Text Analysis for Economics and Finance

Ruben Durante
ICREA-UPF, BGSE, IPEG, CEPR

DIW, October 2020

Sentiment analysis with short text: VADER

- Valence Aware Dictionary and sEntiment Reasoner (VADER): lexicon-based sentiment analysis tool particularly suited for social media content
- A sentiment lexicon is a list of lexical features (e.g., words) labeled according to their semantic orientation as either positive, negative, or neutral
- It has been quite successful dealing with social media texts, newspaper editorials, movie and product reviews.
- It does not only classify text as positive, negative, or neutral, but also provides a composite score that combines all three
- Does not require any training data; constructed from a generalizable human-curated sentiment lexicon. subcategory

Hassan et al. (QJE, 2019): firm-level political risk

- Another example of application of dictionary methods
- But the construction and the use of the dictionary is innovative
- Goal: to construct a measure of political risk faced by individual US firms
- Data: 178,173 transcripts of quarterly earnings conference calls
- Idea: measure the share of the conversations between call participants and firm management centered around risks associated with political matters

Hassan et al. (QJE, 2019): firm-level political risk

- Another example of application of dictionary methods
- But the construction and the use of the dictionary is innovative
- Goal: to construct a measure of political risk faced by individual US firms
- Data: 178,173 transcripts of quarterly earnings conference calls
- Idea: measure the share of the conversations between call participants and firm management centered around risks associated with political matters

Hassan et al. (QJE, 2019): firm-level political risk

- Another example of application of dictionary methods
- But the construction and the use of the dictionary is innovative
- Goal: to construct a measure of political risk faced by individual US firms
- Data: 178,173 transcripts of quarterly earnings conference calls
- Idea: measure the share of the conversations between call participants and firm management centered around risks associated with political matters

Hassan et al. (QJE, 2019): constructing the dictionary

- Training library of political text archetypical of discussions of political topics, **P**
- Training library of non-political text, archetypical of discussion of non-political topics, **N**
- **P**: William T. Bianco and David T. Canon, *American Politics Today*
- **N**: Robert Libby, Patricia A. Libby, and Daniel G. Short's, *Financial Accounting*
- Each library is the set of all adjacent two-word combinations ("bigrams") contained in the text
- $\boldsymbol{P} \backslash \boldsymbol{N}$: terms that are in the political texts but not in the non-political text

Hassan et al. (QJE, 2019): constructing the dictionary

- Training library of political text archetypical of discussions of political topics, **P**
- Training library of non-political text, archetypical of discussion of non-political topics, **N**
- **P**: William T. Bianco and David T. Canon, *American Politics Today*
- **N**: Robert Libby, Patricia A. Libby, and Daniel G. Short's, *Financial Accounting*
- Each library is the set of all adjacent two-word combinations ("bigrams") contained in the text
- $\boldsymbol{P} \setminus \boldsymbol{N}$: terms that are in the political texts but not in the non-political text

Hassan et al. (QJE, 2019): constructing the dictionary

- Training library of political text archetypical of discussions of political topics, **P**
- Training library of non-political text, archetypical of discussion of non-political topics, **N**
- **P**: William T. Bianco and David T. Canon, *American Politics Today*
- **N**: Robert Libby, Patricia A. Libby, and Daniel G. Short's, *Financial Accounting*
- Each library is the set of all adjacent two-word combinations ("bigrams") contained in the text
- $\boldsymbol{P}\backslash\boldsymbol{N}$: terms that are in the political texts but not in the non-political text

Hassan et al. (QJE, 2019): constructing the dictionary

- First key statistics for them is $f_{b,\boldsymbol{P}}/B_{\boldsymbol{P}}$
  - $f_{b,\boldsymbol{P}}$: frequency of bigram b in the political training library
  - $B_{\boldsymbol{P}}$ is the total number of bigrams in the political training library
  - **What is this?**
  - Relative term frequency of b in P. Similar to $tf_{i,j}$

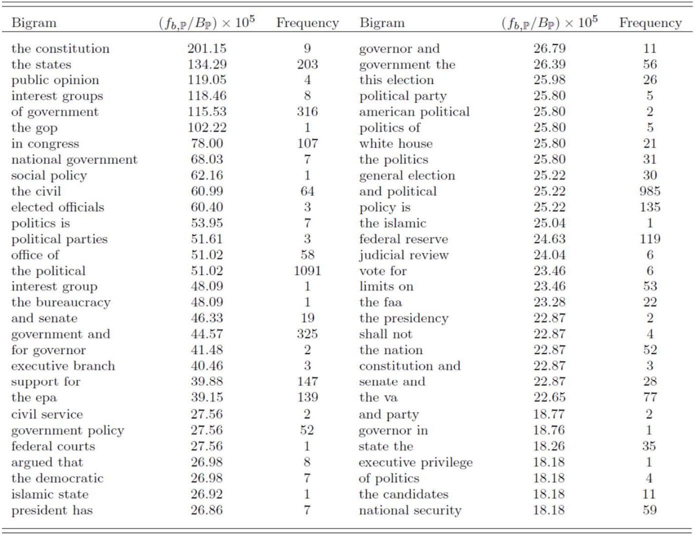
- Second key statistics is $\mathbf{1}[b \in \boldsymbol{P} \backslash \boldsymbol{N}]$
  - Where $\mathbf{1}[\cdot]$ is an indicator function.
  - This is an extreme way of doing an idf_b across libraries. **Why?**
  - idf_b would give more weight to terms that are "special" to library $\boldsymbol{P}$, i.e. not as frequent in $\boldsymbol{N}$.
  - Here the weight is set to 0 for all terms in $\boldsymbol{P}$ that are also in $\boldsymbol{N}$.

Hassan et al. (QJE, 2019): constructing the dictionary

- First key statistics for them is $f_{b,\boldsymbol{P}}/B_{\boldsymbol{P}}$
  - $f_{b,\boldsymbol{P}}$: frequency of bigram b in the political training library
  - $B_{\boldsymbol{P}}$ is the total number of bigrams in the political training library
  - **What is this?**
  - Relative term frequency of b in P. Similar to $tf_{i,j}$
- Second key statistics is $\mathbf{1}[b \in \boldsymbol{P} \backslash \boldsymbol{N}]$
  - Where $\mathbf{1}[\cdot]$ is an indicator function.
  - This is an extreme way of doing an idf_b across libraries. **Why?**
  - idf_b would give more weight to terms that are “special” to library $\boldsymbol{P}$, i.e. not as frequent in $\boldsymbol{N}$.
  - Here the weight is set to 0 for all terms in $\boldsymbol{P}$ that are also in $\boldsymbol{N}$.

Hassan et al. (QJE, 2019): constructing the dictionary

Table 2: Top 120 political bigrams used in construction of $PRisk_{i,t}$



Bigram	$(f_{b,\mathbb{P}}/B_{\mathbb{P}}) \times 10^5$	Frequency	Bigram	$(f_{b,\mathbb{P}}/B_{\mathbb{P}}) \times 10^5$	Frequency
the constitution	201.15	9	governor and	26.79	11
the states	134.29	203	government the	26.39	56
public opinion	119.05	4	this election	25.98	26
interest groups	118.46	8	political party	25.80	5
of government	115.53	316	american political	25.80	2
the gop	102.22	1	politics of	25.80	5
in congress	78.00	107	white house	25.80	21
national government	68.03	7	the politics	25.80	31
social policy	62.16	1	general election	25.22	30
the civil	60.99	64	and political	25.22	985
elected officials	60.40	3	policy is	25.22	135
politics is	53.95	7	the islamic	25.04	1
political parties	51.61	3	federal reserve	24.63	119
office of	51.02	58	judicial review	24.04	6
the political	51.02	1091	vote for	23.46	6
interest group	48.09	1	limits on	23.46	53
the bureaucracy	48.09	1	the faa	23.28	22
and senate	46.33	19	the presidency	22.87	2
government and	44.57	325	shall not	22.87	4
for governor	41.48	2	the nation	22.87	52
executive branch	40.46	3	constitution and	22.87	3
support for	39.88	147	senate and	22.87	28
the epa	39.15	139	the va	22.65	77
civil service	27.56	2	and party	18.77	2
government policy	27.56	52	governor in	18.76	1
federal courts	27.56	1	state the	18.26	35
argued that	26.98	8	executive privilege	18.18	1
the democratic	26.98	7	of politics	18.18	4
islamic state	26.92	1	the candidates	18.18	11
president has	26.86	7	national security	18.18	59

Hassan et al. (QJE, 2019): using the dictionary

- Count the number of instances where political bigrams are used in conjunction with synonyms for "risk"

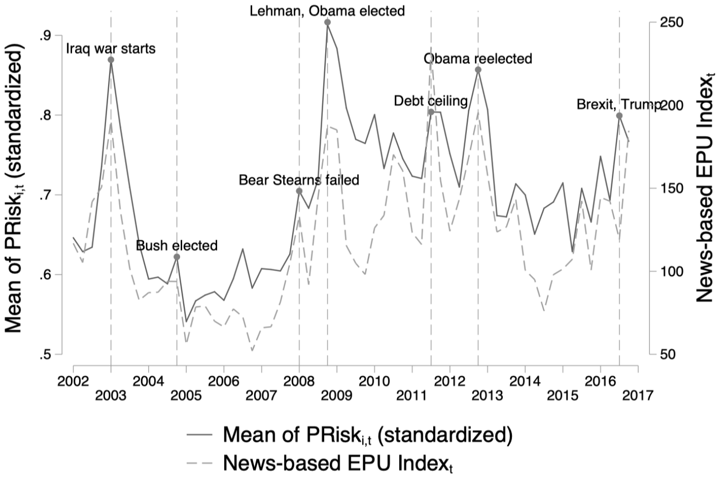
- Conference-call transcript of firm i in quarter t into a list of bigrams contained in the transcript $b = 1, ..., B_{it}$.

$$PRisk_{it} = \frac{\sum_{b=1}^{B_{it}} \left(\mathbf{1}\left[b \in \boldsymbol{P} \backslash \boldsymbol{N}\right] \times \mathbf{1}\left[|b - r| < 10\right] \times \frac{f_{b,\boldsymbol{P}}}{B_{\boldsymbol{P}}} \right)}{B_{it}}$$

- r is the position of the nearest synonim for risk or uncertainty

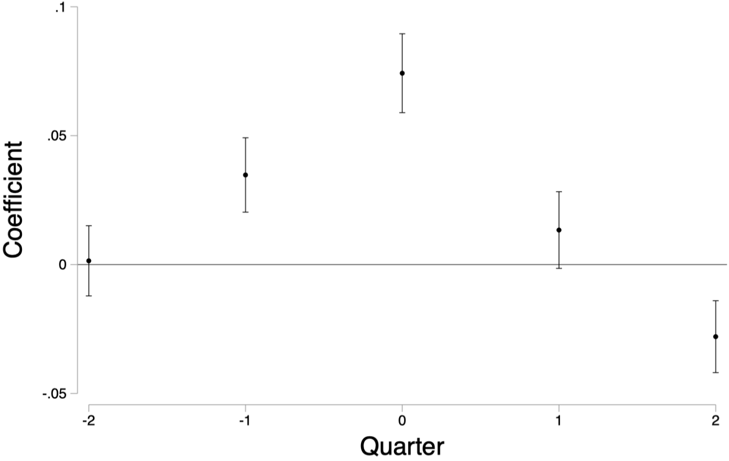
Hassan et al. (QJE, 2019): using the dictionary

Figure 1: Variation in $PRisk_{i,t}$ over time and correlation with EPU

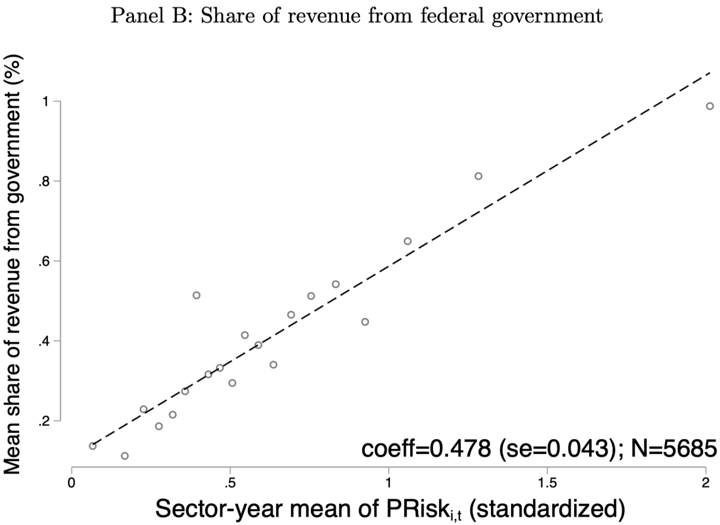


Hassan et al. (QJE, 2019): using the dictionary

Figure 2: Variation in $PRisk_{i,t}$ around federal elections

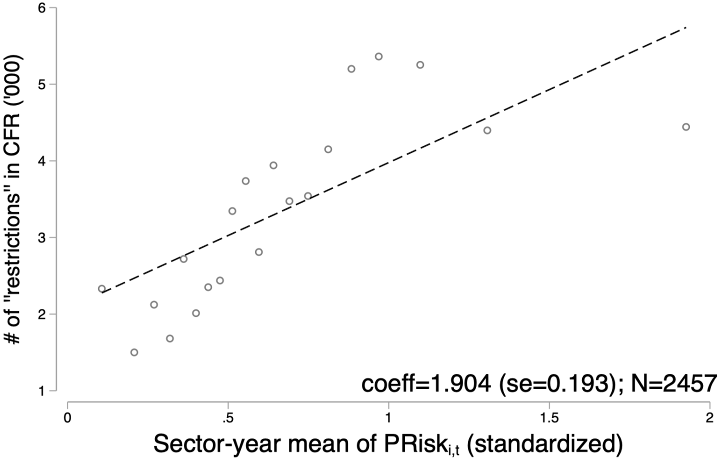


Hassan et al. (QJE, 2019): using the dictionary



Hassan et al. (QJE, 2019): using the dictionary

Panel A: Index of regulatory constraints



Documents as vectors

- In the document-feature-matrix each document is represented by a row-vector
- Each vector contains the (weighted) frequencies of each feature in the document
- Idea: these vectors can be used to measure the similarity/distance between documents

Property of distance measures

- Let A and B be any two documents in a set and $d(A, B)$ be the distance between A and B

1. $d(x, y) \geq 0$: the distance between any two points must be non-negative
2. $d(A, B) = 0$ iff $A = B$: the distance between two documents must be zero if and only if the two objects are identical
3. $d(A, B) = d(B, A)$: distance must be symmetric
4. $d(A, C) \leq d(A, B) + d(B, C)$ must satisfy the triangle inequality

Text analysis of patent innovation

"Measuring technological innovation over the very long run", Kelly et al. (2019)

- Goal:
  - Construct a new measure of novelty and impact of innovations based on similarity and distance between the text of patents
- Data:
  - 9 million patents since 1840, from U.S. Patent Office and Google Scholar Patents.
  - Date, inventor, backward citations
  - Text (abstract, claims, and description)
- Text pre-processing:
  - Drop HTML markup, punctuation, numbers, capitalization, and stopwords
  - Remove terms that appear in less than 20 patents
  - 1.6 million words in vocabulary.

Text analysis of patent innovation

"Measuring technological innovation over the very long run", Kelly et al. (2019)

- Goal:
  - Construct a new measure of novelty and impact of innovations based on similarity and distance between the text of patents
- Data:
  - 9 million patents since 1840, from U.S. Patent Office and Google Scholar Patents.
  - Date, inventor, backward citations
  - Text (abstract, claims, and description)
- Text pre-processing:
  - Drop HTML markup, punctuation, numbers, capitalization, and stopwords
  - Remove terms that appear in less than 20 patents
  - 1.6 million words in vocabulary.

Text analysis of patent innovation

"Measuring technological innovation over the very long run", Kelly et al. (2019)

- Goal:
  - Construct a new measure of novelty and impact of innovations based on similarity and distance between the text of patents
- Data:
  - 9 million patents since 1840, from U.S. Patent Office and Google Scholar Patents.
  - Date, inventor, backward citations
  - Text (abstract, claims, and description)
- Text pre-processing:
  - Drop HTML markup, punctuation, numbers, capitalization, and stopwords
  - Remove terms that appear in less than 20 patents
  - 1.6 million words in vocabulary.

Measuring Innovation

- Backward IDF weighting of word w in patent p:

$$\text{BIDF}(w,p) = \frac{\text{\# of patents prior to } p}{\log\left(1 + \text{\# documents prior to } p \text{ that include } w\right)}$$

  - Down-weights words that appeared frequently before a patent, but up-weights new words

- For each patent:
  - Compute cosine similarity to all future patents, using BIDF of earlier patent

- 9m×9m similarity matrix = 30TB of data
  - Enforce sparsity by setting similarity $< .05$ to zero (93.4% of pairs).

Measuring Innovation

- Backward IDF weighting of word w in patent p:

$$\text{BIDF}(w,p) = \frac{\text{\# of patents prior to } p}{\log\left(1 + \text{\# documents prior to } p \text{ that include } w\right)}$$

  - Down-weights words that appeared frequently before a patent, but up-weights new words

- For each patent:
  - Compute cosine similarity to all future patents, using BIDF of earlier patent

- 9m×9m similarity matrix = 30TB of data
  - Enforce sparsity by setting similarity $< .05$ to zero (93.4% of pairs).

Measuring Innovation

- Backward IDF weighting of word w in patent p:

$$\text{BIDF}(w,p) = \frac{\#\text{ of patents prior to } p}{\log\left(1 + \#\text{ documents prior to } p \text{ that include } w\right)}$$

  - Down-weights words that appeared frequently before a patent, but up-weights new words

- For each patent:
  - Compute cosine similarity to all future patents, using BIDF of earlier patent

- 9m$\times$9m similarity matrix $=$ 30TB of data
  - Enforce sparsity by setting similarity $< .05$ to zero (93.4% of pairs).

Novelty, Impact, and Quality

- “Novelty” is defined by (negative) similarity to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

  where $B(j)$ is the set of previous patents (in, e.g., last 20 years).

- “Impact” is defined as similarity to subsequent patents:

$$\text{Impact}_i = \sum_{i \in F(j)} \rho_{ij}$$

  where $F(j)$ is the set of future patents (in, e.g., next 100 years).

- A patent has high quality if it is novel and impactful:

$$\text{Quality}_i = \frac{\text{Impact}_i}{-\text{Novelty}_i}$$

Novelty, Impact, and Quality

- "Novelty" is defined by (negative) similarity to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

  where $B(j)$ is the set of previous patents (in, e.g., last 20 years).

- "Impact" is defined as similarity to subsequent patents:

$$\text{Impact}_i = \sum_{i \in F(j)} \rho_{ij}$$

  where $F(j)$ is the set of future patents (in, e.g., next 100 years).

- A patent has high quality if it is novel and impactful:

$$\text{Quality}_i = \frac{\text{Impact}_i}{-\text{Novelty}_i}$$

Novelty, Impact, and Quality

- "Novelty" is defined by (negative) similarity to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

  where $B(j)$ is the set of previous patents (in, e.g., last 20 years).

- "Impact" is defined as similarity to subsequent patents:

$$\text{Impact}_i = \sum_{i \in F(j)} \rho_{ij}$$

  where $F(j)$ is the set of future patents (in, e.g., next 100 years).

- A patent has high quality if it is novel and impactful:

$$\text{Quality}_i = \frac{\text{Impact}_i}{-\text{Novelty}_i}$$

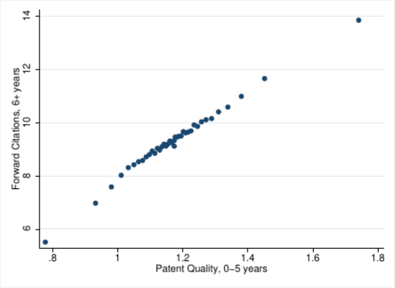
Validation

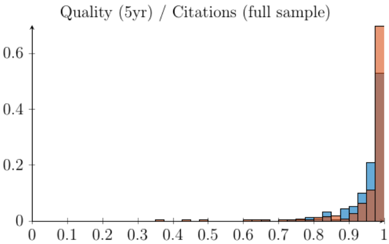
1. For pairs with higher $\rho_{i,j}$, patent j is more likely to cite patent i.

2. Patent office assigns 3-digit technology class code; similarity is significantly higher within class compared to across class.

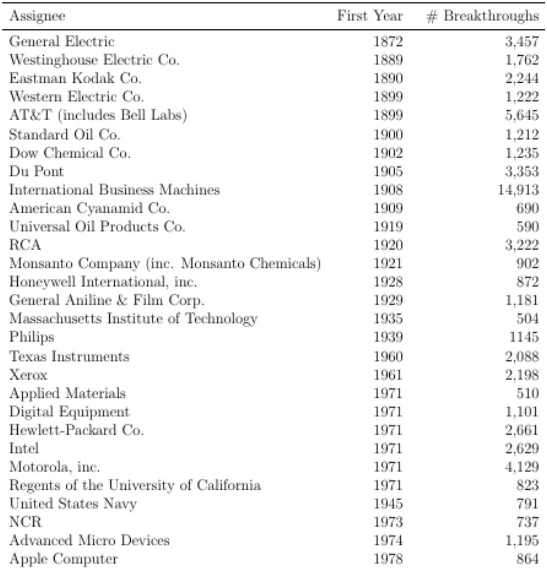
3. Higher quality patents get more cites:

Validation (cont.)



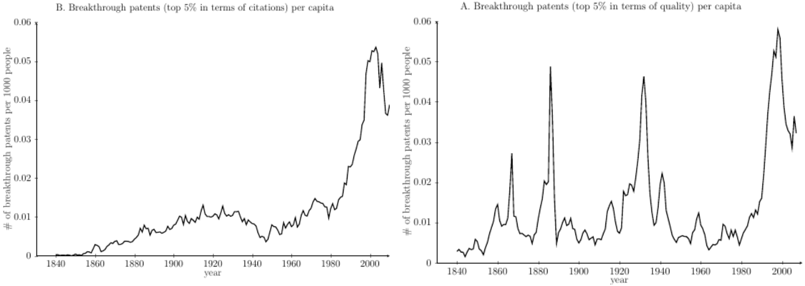


Most Innovative Firms



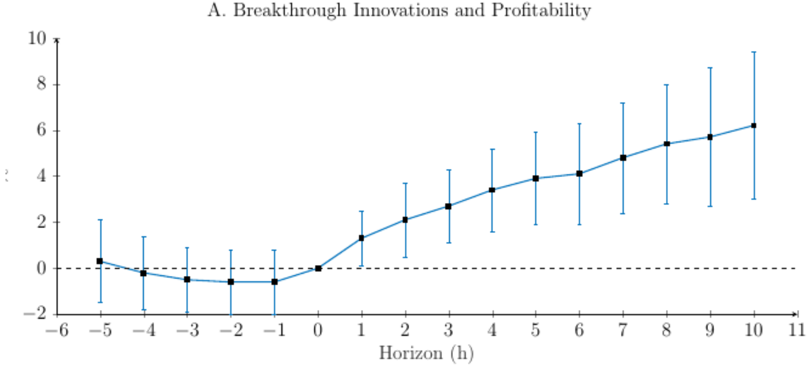
Assignee	First Year	# Breakthroughs
General Electric	1872	3,457
Westinghouse Electric Co.	1889	1,762
Eastman Kodak Co.	1890	2,244
Western Electric Co.	1899	1,222
AT&T (includes Bell Labs)	1899	5,645
Standard Oil Co.	1900	1,212
Dow Chemical Co.	1902	1,235
Du Pont	1905	3,353
International Business Machines	1908	14,913
American Cyanamid Co.	1909	690
Universal Oil Products Co.	1919	590
RCA	1920	3,222
Monsanto Company (inc. Monsanto Chemicals)	1921	902
Honeywell International, inc.	1928	872
General Aniline & Film Corp.	1929	1,181
Massachusetts Institute of Technology	1935	504
Philips	1939	1145
Texas Instruments	1960	2,088
Xerox	1961	2,198
Applied Materials	1971	510
Digital Equipment	1971	1,101
Hewlett-Packard Co.	1971	2,661
Intel	1971	2,629
Motorola, inc.	1971	4,129
Regents of the University of California	1971	823
United States Navy	1945	791
NCR	1973	737
Advanced Micro Devices	1974	1,195
Apple Computer	1978	864

Breakthrough patents per capita


B. Breakthrough patents (top 5% in terms of citations) per capita

A. Breakthrough patents (top 5% in terms of quality) per capita

Breakthrough patents and firm profits



Jegadeesh and Wu (2013)

- Goal: estimate the response of a company's stock returns to information contained in the company's 10-K filings
- Idea: the content and wording of 10-Ks may prompt investor reactions in the following days which, in turn, can impact the firm's stock returns
- Combine dictionary methods and text regressions to map the response of stock returns to the words included in a firm's 10-Ks
- Main finding: stronger and more stable relationship between the tone of 10-Ks and a firm's stock market returns than previously found using simpler word counting techniques

Data

- Use all 10-Ks filed from January 1995 through December 2010 from the SEC's EDGAR database using a web crawling algorithm
- The final sample contains 45,860 filings by 7,606 unique firms
- For the dictionary, use the negative and positive word lists constructed by Loughran and McDonald (2011) (LM).
- The LM list contains **353** positive words and **2,337** negative words. By grouping similar words, the list is reduced to **123** positive words and **718** negative words

Empirical strategy

- The objective is to fit the regression:

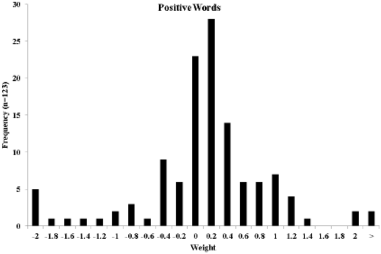
$$v_i = a + b\left(\sum_j w_j \frac{c_{ij}}{\sum_j c_{ij}}\right) + \varepsilon_i$$

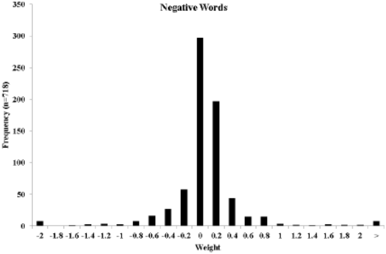
  where c_{ij} is a count of occurrences of word j in 10-K report i, and v_i is the company's stock excess returns in the post-filing days

- The coefficient w_j summarizes the average association between an occurrence of word j and the stock's return

- These coefficients are estimated using a cross-sectional regression, while a subsequent rescaling of all coefficients removes the common influence parameter b.

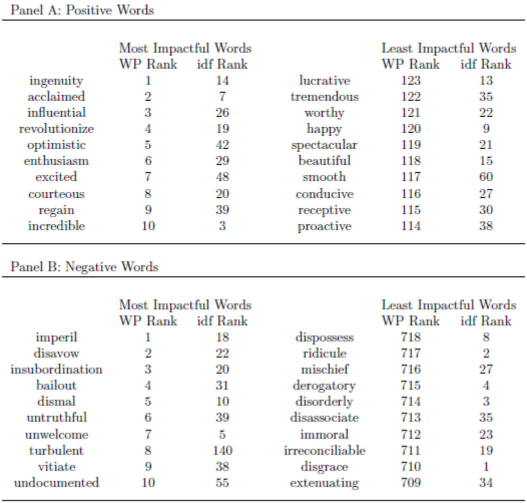
Empirical Strategy (cont.)





- The distribution of the resulting standardized LM word weights for positive and negative words
- They emphasize that these weights are very different from those computed with simple tf-idf

Empirical Strategy (cont.)



Panel A: Positive Words

	Most Impactful Words			Least Impactful Words	
	WP Rank	idf Rank		WP Rank	idf Rank
ingenuity	1	14	lucrative	123	13
acclaimed	2	7	tremendous	122	35
influential	3	26	worthy	121	22
revolutionize	4	19	happy	120	9
optimistic	5	42	spectacular	119	21
enthusiasm	6	29	beautiful	118	15
excited	7	48	smooth	117	60
courteous	8	20	conducive	116	27
regain	9	39	receptive	115	30
incredible	10	3	proactive	114	38

Panel B: Negative Words

	Most Impactful Words			Least Impactful Words	
	WP Rank	idf Rank		WP Rank	idf Rank
imperil	1	18	dispossess	718	8
disavow	2	22	ridicule	717	2
insubordination	3	20	mischief	716	27
bailout	4	31	derogatory	715	4
dismal	5	10	disorderly	714	3
untruthful	6	39	disassociate	713	35
unwelcome	7	5	immoral	712	23
turbulent	8	140	irreconciliable	711	19
vitiate	9	38	disgrace	710	1
undocumented	10	55	extenuating	709	34

- Weights rank substantially different from tf-idf counterparts

Results

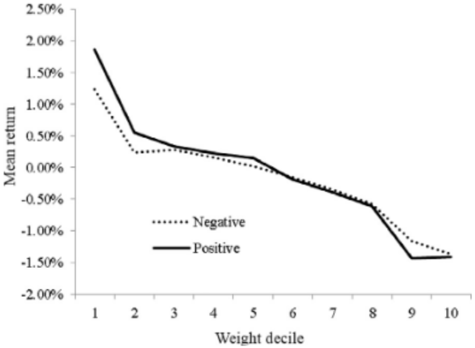


Fig. 3. Mean filing period abnormal return using word power weights. This figure presents the distribution of filing window abnormal returns, defined as a firm's buy-and-hold return minus the CRSP value-weighted index return over the four-day window of [filing date, filing date + 3] across various deciles of filings sorted based on the word power scores of the 10-Ks. We compute the word power weights for each year using Eqs. (6) and (7) over the sample period prior to the filing of 10-Ks. We compute the positive and negative tone for each 10-K using Eq. (4). Decile 1 is comprised of the decile of firms with the most positive (or least negative) document scores and decile 10 is comprised of the decile of firms with the least positive (or most negative) document scores. Mean return is the average filing period abnormal returns for all firms in that decile. The sample covers 45,860 10-Ks over the 1995 to 2010 sample period.

Results (cont.)

- Propose an approach to avoid subjectivity inherent to using lexicons composed of words with positive or negative connotations
- Move away from the idea that all positive or negative words "count" the same, let the data determine importance of each term for outcome
- Robustness checks show that this weighting strategy reliably quantifies tone even when the subjectivity of the lexicon is increased
- Out-of-sample properties are superior to their dictionary-based counterparts. Highlight the limitations of the latter for purpose of prediction

Manela and Moreira (2017): measuring NVIX

- Goal: construct an index of news-implied market volatility by identifying a small subset of words whose frequencies are most useful for predicting distress in financial markets

- Idea: relative coverage of topics by the business press is a good proxy for the evolution of investors' concerns about these topics

- Use supervised machine learning (SVM) to identify relevant features from text and to build predictions of distress

- Main finding: news-implied volatility captures well the disaster concerns of the average investor over a long time span
  - War and government policy are good predictors of risk premia

Manela and Moreira (2017): data

- Extract the title and abstract of all front-page articles of the Wall Street Journal from July 1889 to December 2009
- Transform text data into c_{it}, the relative frequency of over 400,000 n-grams
- Volatility measured by the VIX implied volatility index
- Break the sample into three subsamples:
  - *Train* subsample estimates the relation between news data and implied volatility using data from 1996 to 2009
  - *Test* subsample covers 1986-1995 and is used for out-of-sample tests of model fit
  - *Prediction* subsample includes observations prior to 1986 for which features are not available

Manela and Moreira (2017): empirical strategy

- We need to estimate

$$VIX_t = w_0 + \mathbf{w} \cdot \mathbf{x}_t + v_t$$

but with OLS we will overfit ($T_{train} = 168$, $len(x_t) = 468,091$)

- A solution is to use support vector regressions, which instead minimizes the following objective

$$L(\mathbf{w}, w_0) = \sum_{t \in train} g_\epsilon(v_t - w_0 - \mathbf{w} \cdot \mathbf{x}_t) + c(\mathbf{w} \cdot \mathbf{w})$$

where $g_\varepsilon(e) = \max\{0, |e| - \varepsilon\}$ is an ε-insensitive error

- The minimizing coefficients vector w is a weighted average of regressors

$$\hat{\mathbf{w}}_{SVR} = \sum_{t \in \text{train}} (\hat{\alpha}_t^* - \hat{\alpha}_t)\mathbf{x}_t$$

Manela and Moreira (2017): empirical strategy

- We need to estimate

  $$VIX_t = w_0 + \mathbf{w} \cdot \mathbf{x}_t + v_t$$

  but with OLS we will overfit ($T_{train} = 168$, $len(x_t) = 468,091$)

- A solution is to use support vector regressions, which instead minimizes the following objective

  $$L(\mathbf{w}, w_0) = \sum_{t \in train} g_\epsilon (v_t - w_0 - \mathbf{w} \cdot \mathbf{x}_t) + c(\mathbf{w} \cdot \mathbf{w})$$

  where $g_\varepsilon(e) = \max\{0, |e| - \varepsilon\}$ is an ε-insensitive error

- The minimizing coefficients vector w is a weighted average of regressors

  $$\hat{\mathbf{w}}_{SVR} = \sum_{t \in \text{train}} (\hat{\alpha}_t^* - \hat{\alpha}_t)\mathbf{x}_t$$

Manela and Moreira (2017): empirical strategy

- We need to estimate

$$VIX_t = w_0 + \mathbf{w} \cdot \mathbf{x}_t + v_t$$

but with OLS we will overfit ($T_{train} = 168$, $len(x_t) = 468,091$)

- A solution is to use support vector regressions, which instead minimizes the following objective

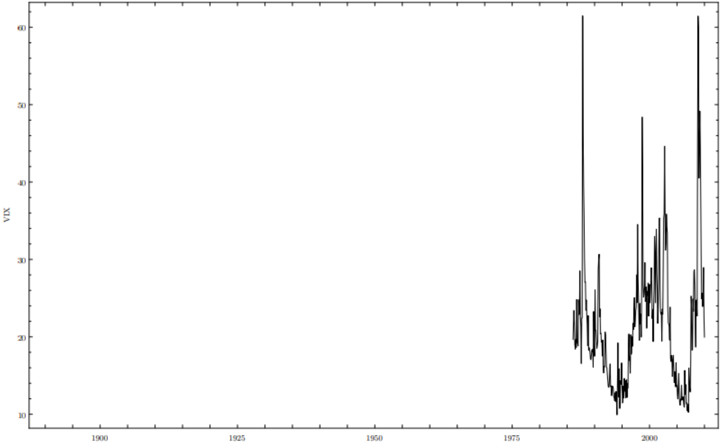
$$L(\mathbf{w}, w_0) = \sum_{t \in train} g_{\epsilon}(v_t - w_0 - \mathbf{w} \cdot \mathbf{x}_t) + c(\mathbf{w} \cdot \mathbf{w})$$

where $g_{\varepsilon}(e) = \max\{0, |e| - \varepsilon\}$ is an ε-insensitive error

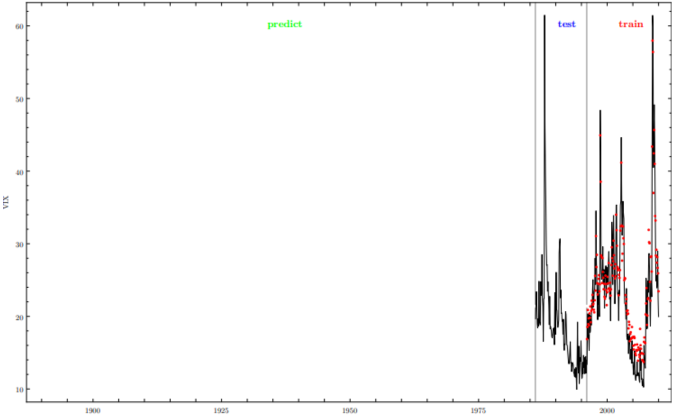
- The minimizing coefficients vector w is a weighted average of regressors

$$\hat{\mathbf{w}}_{SVR} = \sum_{t \in \text{train}} (\hat{\alpha}_t^* - \hat{\alpha}_t)\mathbf{x}_t$$

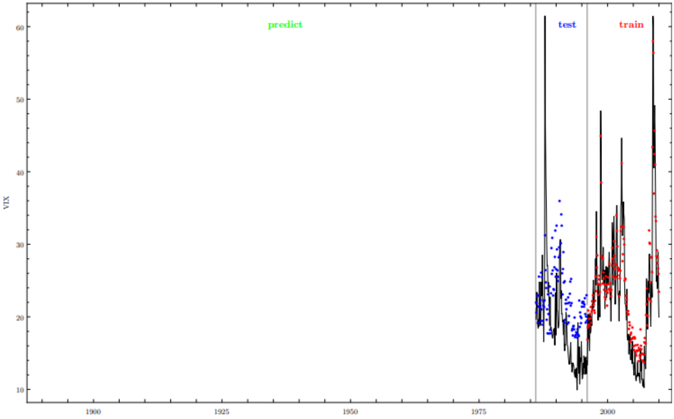
Manela and Moreira (2017): estimation results



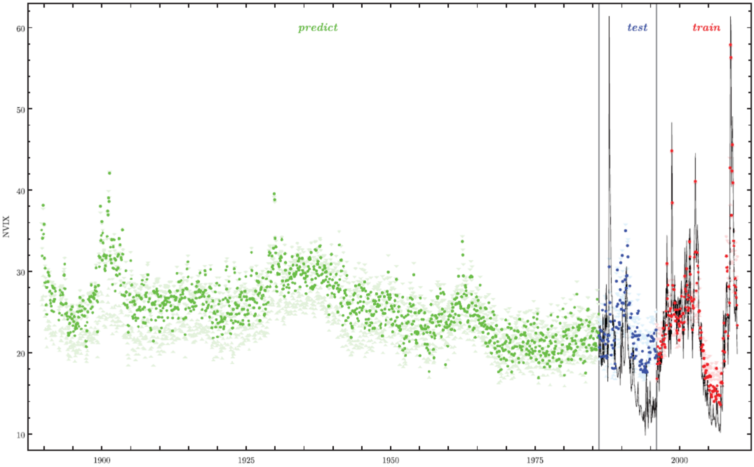
Manela and Moreira (2017): NVIX construction



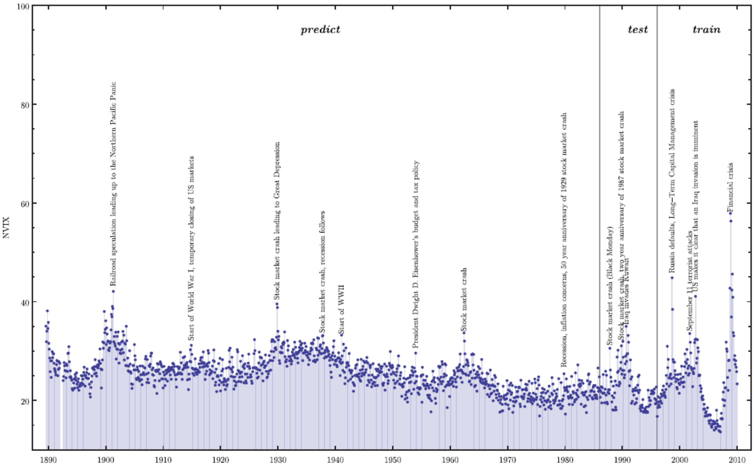
Manela and Moreira (2017): NVIX construction



Manela and Moreira (2017): NVIX construction



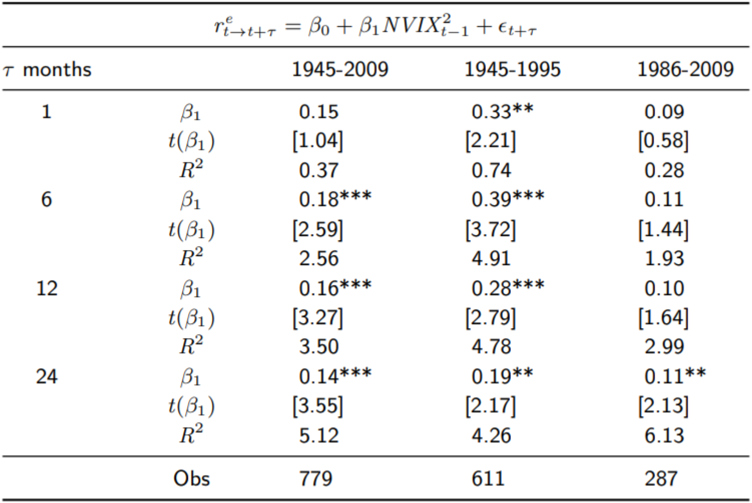
Manela and Moreira (2017): NVIX construction



Manela and Moreira (2017): Empirical Strategy

- Asset pricing models with time-varying risk premia suggest that times when risk is relatively high would be followed by above average stock market returns
  - Time-varying volatility (Merton, 1973)
  - Time-varying disaster risk (Gabaix, 2012)
- Goal: explain future excess return on the market protfolio at various horizons with lagged forward-looking measures of risk as measured by NVIX squared

Manela and Moreira (2017): Estimation Results



$$r^{e}_{t\to t+\tau} = \beta_0 + \beta_1 NVIX^2_{t-1} + \epsilon_{t+\tau}$$

τ months		1945-2009	1945-1995	1986-2009
1	β_1	0.15	0.33**	0.09
	$t(\beta_1)$	[1.04]	[2.21]	[0.58]
	R^2	0.37	0.74	0.28
6	β_1	0.18***	0.39***	0.11
	$t(\beta_1)$	[2.59]	[3.72]	[1.44]
	R^2	2.56	4.91	1.93
12	β_1	0.16***	0.28***	0.10
	$t(\beta_1)$	[3.27]	[2.79]	[1.64]
	R^2	3.50	4.78	2.99
24	β_1	0.14***	0.19**	0.11**
	$t(\beta_1)$	[3.55]	[2.17]	[2.13]
	R^2	5.12	4.26	6.13
	Obs	779	611	287

Supervised machine learning

- Goal: classify documents into pre-existing categories
- What do we need:
    - Hand-coded labeled dataset, to be split into training set and a validation/test set
    - Method to extrapolate from hand-coding to unlabeled, i.e., classifier
- Possible classifiers:
    - Regularized regressions
    - Naive Bayes
    - Support Vector Machines
    - Ensemble classifiers

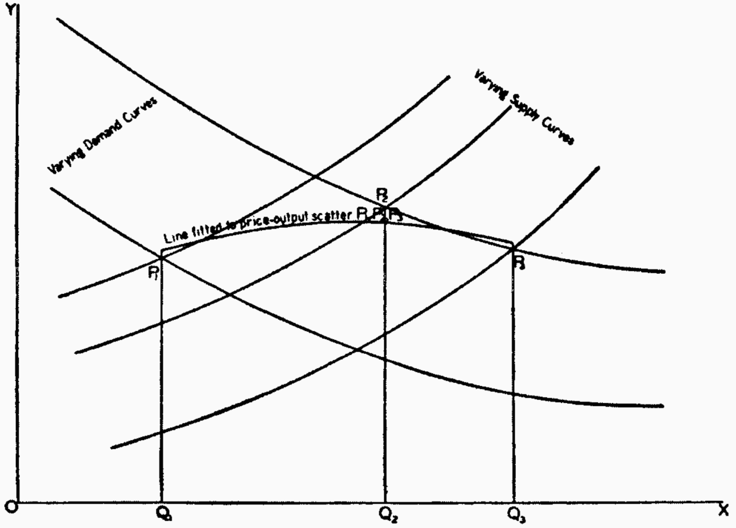
Supervised machine learning

- Goal: classify documents into pre-existing categories
- What do we need:
  - Hand-coded labeled dataset, to be split into training set and a validation/test set
  - Method to extrapolate from hand-coding to unlabeled, i.e., classifier
- Possible classifiers:
  - Regularized regressions
  - Naive Bayes
  - Support Vector Machines
  - Ensemble classifiers

Stock-Trebbi 2003: who invented instrumental variables?

- First derivation of IV estimator in Appendix B of *The Tariff on Animal and Vegetable Oils* by Philip G. Wright (1928)
  - First 285 pages: "a painfully detailed treatise on animal and vegetable oils, their production, uses, markets and tariffs"
  - Appendix B: "Out of the blue... a succinct and insightful of why price and quantity data alone are in general inadequate, two separate and correct derivations of IV, and an empirical application to butter and flaxseed"
- Because Appendix B is so different many people (e.g., Manski 1988) have suggested it might have been written by Philip's son Sewall Wright, a famous genetic statistician

FIGURE 4. PRICE-OUTPUT DATA FAIL TO REVEAL EITHER SUPPLY OR DEMAND CURVE.



Authorship debate

- Case for Sewall:
  - Appendix uses method of "path coefficients" which Sewall had recently invented
  - A more eminent statitician
- Case for Philip:
  - Was an economist while Sewall was not
  - Had written frequently about the identification problem

Sample

- Each “document” is a block of 1000 words
- “Training set”
  - 20 undisputedly by Sewall
  - 25 undisputedly by Philip
- “Prediction set”
  - 6 from Appendix B
  - 1 from Chapter 1

From text to data

- Define columns of C to be:
    - Counts of 70 “function words” taken from Mosteller & Wallace
    - Counts of 18 “grammatical constructions” from Mannion & Dixon (1997)
- Result: $n = 52$, $p = 88$, with V observed for 45 documents

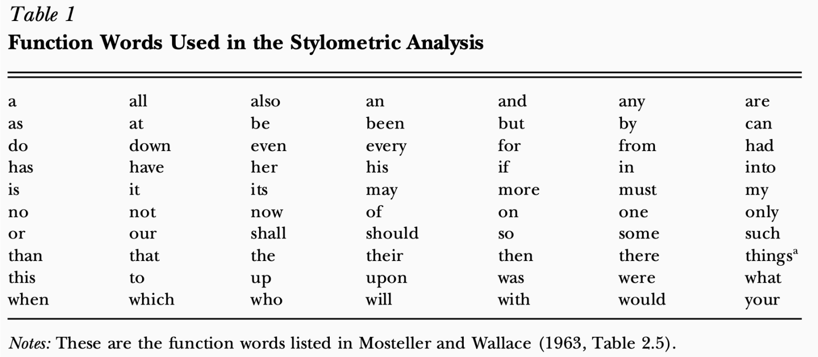


Table 1

Function Words Used in the Stylometric Analysis

a	all	also	an	and	any	are
as	at	be	been	but	by	can
do	down	even	every	for	from	had
has	have	her	his	if	in	into
is	it	its	may	more	must	my
no	not	now	of	on	one	only
or	our	shall	should	so	some	such
than	that	the	their	then	there	things[a]
this	to	up	upon	was	were	what
when	which	who	will	with	would	your

Notes: These are the function words listed in Mosteller and Wallace (1963, Table 2.5).

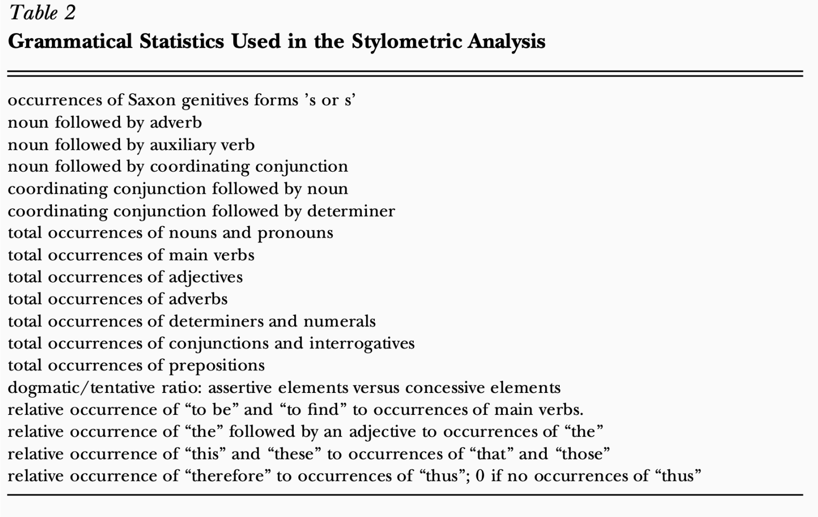


Table 2

Grammatical Statistics Used in the Stylometric Analysis

occurrences of Saxon genitives forms ’s or s’
noun followed by adverb
noun followed by auxiliary verb
noun followed by coordinating conjunction
coordinating conjunction followed by noun
coordinating conjunction followed by determiner
total occurrences of nouns and pronouns
total occurrences of main verbs
total occurrences of adjectives
total occurrences of adverbs
total occurrences of determiners and numerals
total occurrences of conjunctions and interrogatives
total occurrences of prepositions
dogmatic/tentative ratio: assertive elements versus concessive elements
relative occurrence of “to be” and “to find” to occurrences of main verbs.
relative occurrence of “the” followed by an adjective to occurrences of “the”
relative occurrence of “this” and “these” to occurrences of “that” and “those”
relative occurrence of “therefore” to occurrences of “thus”; 0 if no occurrences of “thus”

Method

- We still have $p > n$, so just regressing V on c is unfeasible
- Two measures, each computed for function words and grammar separately, to produce $(\hat{V}_{fw}, \hat{V}_{g})$

Approach #1: PCA

- Compute 4 principal components of C_{fw} and C_g
- Regress V on each separately
- (Unsupervised dimension reduction)

Approach #1: linear discriminant analysis (Fischer, 1936)

- Compute “linear discriminant” of C_{fw} and C_g

$$\hat{V} = \sum_p w_p c_p$$

$$w_p = \frac{\bar{C}_{p:P} - \bar{C}_{p:S}}{s^2_{p:P} + s^2_{p:S}}$$

- This is the optimal Bayes classifier if X’s are normal with equal variance
- Supervised dimension reduction using generative model

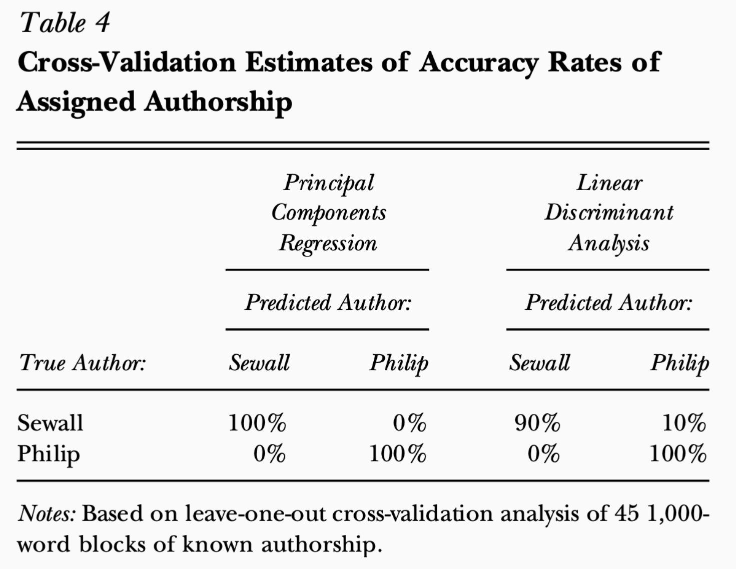


Table 4

Cross-Validation Estimates of Accuracy Rates of Assigned Authorship

	Principal Components Regression		*Linear Discriminant Analysis*	
	Predicted Author:		*Predicted Author:*	
True Author:	*Sewall*	*Philip*	*Sewall*	*Philip*
Sewall	100%	0%	90%	10%
Philip	0%	100%	0%	100%

Notes: Based on leave-one-out cross-validation analysis of 45 1,000-word blocks of known authorship.

Figure 1

Scatterplot of Predicted Values from Regression on First Four Principal Components: Grammatical Statistics versus Function Words

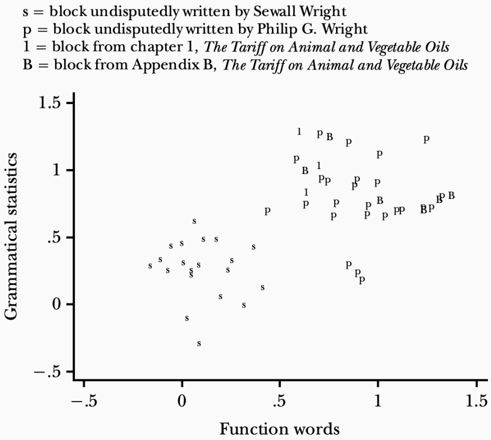
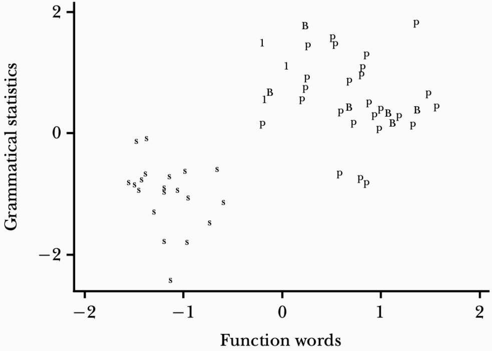


Figure 2

Scatterplot of Linear Discriminant Based on Grammatical Statistics versus Linear Discriminant Based on Function Words

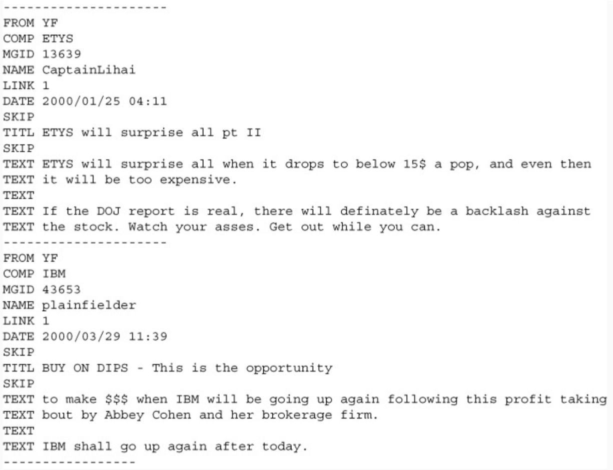


Conclusion

- Philip is undoubtedly the author
- Of course this didn't mean it was his *idea*
- But Stock & Trebbi conclude it probably was

Financial news and asset price: Antweiler-Frank 2004

- Data: Message board contents on Yahoo!Finance & Raging Bull



```
---------------------
FROM YF
COMP ETYS
MGID 13639
NAME CaptainLihai
LINK 1
DATE 2000/01/25 04:11
SKIP
TITL ETYS will surprise all pt II
SKIP
TEXT ETYS will surprise all when it drops to below 15$ a pop, and even then
TEXT it will be too expensive.
TEXT
TEXT If the DOJ report is real, there will definately be a backlash against
TEXT the stock. Watch your asses. Get out while you can.
---------------------
FROM YF
COMP IBM
MGID 43653
NAME plainfielder
LINK 1
DATE 2000/03/29 11:39
SKIP
TITL BUY ON DIPS - This is the opportunity
SKIP
TEXT to make $$$ when IBM will be going up again following this profit taking
TEXT bout by Abbey Cohen and her brokerage firm.
TEXT
TEXT IBM shall go up again after today.
-----------------
```

Antweiler and Frank 2004

- Count words
- Create training set of 1,000 messages hand-coded as: buy, sell, hold
- Compute "naive Bayes classification:" posterior guess assuming words are independent

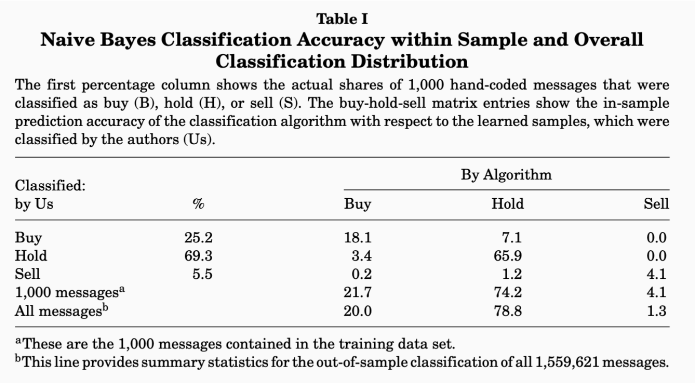


Table I

Naive Bayes Classification Accuracy within Sample and Overall Classification Distribution

The first percentage column shows the actual shares of 1,000 hand-coded messages that were classified as buy (B), hold (H), or sell (S). The buy-hold-sell matrix entries show the in-sample prediction accuracy of the classification algorithm with respect to the learned samples, which were classified by the authors (Us).

Classified:		By Algorithm		
by Us	%	Buy	Hold	Sell
Buy	25.2	18.1	7.1	0.0
Hold	69.3	3.4	65.9	0.0
Sell	5.5	0.2	1.2	4.1
1,000 messages[a]		21.7	74.2	4.1
All messages[b]		20.0	78.8	1.3

[a] These are the 1,000 messages contained in the training data set.

[b] This line provides summary statistics for the out-of-sample classification of all 1,559,621 messages.

Antweiler and Frank 2004

- Small amount of predictability in returns
- Messages predict volatility
- Disagreement (variable recommendations) predicts volume

Research question

- How to measure partisan bias in the media?
- Mass media can slant the news to favor a particular point of view (*partisan bias*)
- Forms of partisan bias:
  - Unbalanced reporting of political events (language, citations, etc.)
  - More newstime devoted to like-minded politicians/experts
  - More emphasis on issues on which a party is perceived as stronger
  - More emphasis on bad performance or scandals of opposing party
- In any case, a very elusive object!
- Important to study other questions, e.g., what drives media bias?

Research question

- How to measure partisan bias in the media?
- Mass media can slant the news to favor a particular point of view (*partisan bias*)
- Forms of partisan bias:
  - Unbalanced reporting of political events (language, citations, etc.)
  - More newstime devoted to like-minded politicians/experts
  - More emphasis on issues on which a party is perceived as stronger
  - More emphasis on bad performance or scandals of opposing party
- In any case, a very elusive object!
- Important to study other questions, e.g., what drives media bias?

Research question

- How to measure partisan bias in the media?
- Mass media can slant the news to favor a particular point of view (*partisan bias*)
- Forms of partisan bias:
  - Unbalanced reporting of political events (language, citations, etc.)
  - More newstime devoted to like-minded politicians/experts
  - More emphasis on issues on which a party is perceived as stronger
  - More emphasis on bad performance or scandals of opposing party
- In any case, a very elusive object!
- Important to study other questions, e.g., what drives media bias?

Research question

- How to measure partisan bias in the media?
- Mass media can slant the news to favor a particular point of view (*partisan bias*)
- Forms of partisan bias:
  - Unbalanced reporting of political events (language, citations, etc.)
  - More newstime devoted to like-minded politicians/experts
  - More emphasis on issues on which a party is perceived as stronger
  - More emphasis on bad performance or scandals of opposing party
- In any case, a very elusive object!
- Important to study other questions, e.g., what drives media bias?

Media bias: an example

- Fox News:

  "In one of the deadliest reported firefights in Iraq since the fall of Saddam Hussein's regime, US forces killed at least 54 Iraqis and captured eight others while fending off simultaneous convoy ambushes Sunday in the northern city of Samarra."

- New York Times:

  "American commanders vowed Monday that the killing of as many as 54 insurgents in this central Iraqi town would serve as a lesson to those fighting the United States, but Iraqis disputed the death toll and said anger against America would only rise."

- Al-Jazeera.net:

  "The US military has vowed to continue aggressive tactics after saying it killed 54 Iraqis following an ambush, but commanders admitted they had no proof to back up their claims. The only corpses at Samarra's hospital were those of civilians, including two elderly Iranian visitors and a child."

A Measure of Media Bias (Groseclose-Milyo, 2005)

- Goal:
  - Estimate relative ideological scores for several major U.S. media outlets
- Empirical Idea:
  - Count the times a media outlet cites various think tanks and compare with the times members of Congress cite the same groups
  - Citations are the language, politicians the reference point
  - Only look at content (no editorials, letters, etc.)
- Findings:
  - Strong liberal bias: all news outlets but two (Fox News, Washington Times) have scores to the left of the average member of Congress
  - Most leftist: CBS Evening News, NYT
  - Most centrist: PBS NewsHour, CNN's Newsnight, ABC's Good Morning America, USA Today

A Measure of Media Bias (Groseclose-Milyo, 2005)

- Goal:
  - Estimate relative ideological scores for several major U.S. media outlets
- Empirical Idea:
  - Count the times a media outlet cites various think tanks and compare with the times members of Congress cite the same groups
  - Citations are the language, politicians the reference point
  - Only look at content (no editorials, letters, etc.)
- Findings:
  - Strong liberal bias: all news outlets but two (Fox News, Washington Times) have scores to the left of the average member of Congress
  - Most leftist: CBS Evening News, NYT
  - Most centrist: PBS NewsHour, CNN's Newsnight, ABC's Good Morning America, USA Today

A Measure of Media Bias (Groseclose-Milyo, 2005)

- Goal:
  - Estimate relative ideological scores for several major U.S. media outlets
- Empirical Idea:
  - Count the times a media outlet cites various think tanks and compare with the times members of Congress cite the same groups
  - Citations are the language, politicians the reference point
  - Only look at content (no editorials, letters, etc.)
- Findings:
  - Strong liberal bias: all news outlets but two (Fox News, Washington Times) have scores to the left of the average member of Congress
  - Most leftist: CBS Evening News, NYT
  - Most centrist: PBS NewsHour, CNN's Newsnight, ABC's Good Morning America, USA Today

Data

- List of 200 of the most prominent US think tanks and policy groups
- Search the *Congressional Record* for instances where a member of Congress cited one of these think tanks (from 1993 to 2002)
- Record the average (adjusted) ADA score of the member who cited the think tank (0-conservative to 100-liberal).
- Include citations in written reports.
- Do not consider mentions of actions, but just views of the TT/PG.
- Omit cases where a politician/media outlet mentions a TT/PG just to criticize it (only 5% and 1% of mentions respectively)
- Omit cases where a politician/media gave an ideological label to a TT/PG (only 2% and 5% respectively
- "The idea is that we only wanted cases where the legislator/journalist cited the think tank as if it were a disinterested expert on the topic"

Data

- List of 200 of the most prominent US think tanks and policy groups
- Search the *Congressional Record* for instances where a member of Congress cited one of these think tanks (from 1993 to 2002)
- Record the average (adjusted) ADA score of the member who cited the think tank (0-conservative to 100-liberal).
- Include citations in written reports.
- Do not consider mentions of actions, but just views of the TT/PG.
- Omit cases where a politician/media outlet mentions a TT/PG just to criticize it (only 5% and 1% of mentions respectively)
- Omit cases where a politician/media gave an ideological label to a TT/PG (only 2% and 5% respectively
- "The idea is that we only wanted cases where the legislator/journalist cited the think tank as if it were a disinterested expert on the topic"

Data

- List of 200 of the most prominent US think tanks and policy groups
- Search the *Congressional Record* for instances where a member of Congress cited one of these think tanks (from 1993 to 2002)
- Record the average (adjusted) ADA score of the member who cited the think tank (0-conservative to 100-liberal).
- Include citations in written reports.
- Do not consider mentions of actions, but just views of the TT/PG.
- Omit cases where a politician/media outlet mentions a TT/PG just to criticize it (only 5% and 1% of mentions respectively)
- Omit cases where a politician/media gave an ideological label to a TT/PG (only 2% and 5% respectively
- "The idea is that we only wanted cases where the legislator/journalist cited the think tank as if it were a disinterested expert on the topic"

Think Tanks

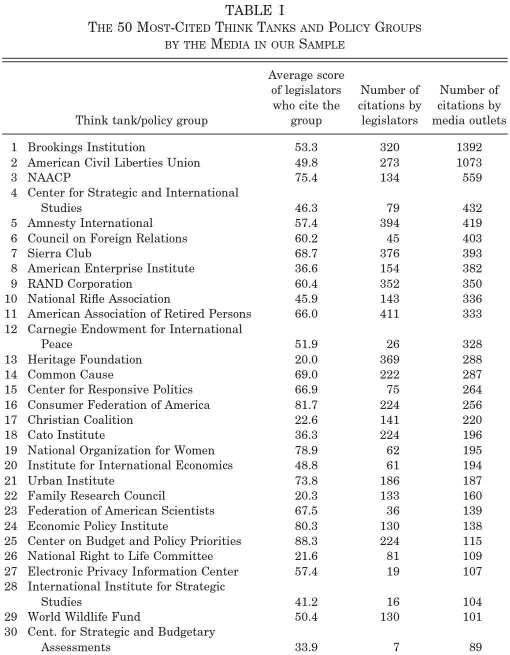


TABLE I
The 50 Most-Cited Think Tanks and Policy Groups
by the Media in our Sample

	Think tank/policy group	Average score of legislators who cite the group	Number of citations by legislators	Number of citations by media outlets
1	Brookings Institution	53.3	320	1392
2	American Civil Liberties Union	49.8	273	1073
3	NAACP	75.4	134	559
4	Center for Strategic and International Studies	46.3	79	432
5	Amnesty International	57.4	394	419
6	Council on Foreign Relations	60.2	45	403
7	Sierra Club	68.7	376	393
8	American Enterprise Institute	36.6	154	382
9	RAND Corporation	60.4	352	350
10	National Rifle Association	45.9	143	336
11	American Association of Retired Persons	66.0	411	333
12	Carnegie Endowment for International Peace	51.9	26	328
13	Heritage Foundation	20.0	369	288
14	Common Cause	69.0	222	287
15	Center for Responsive Politics	66.9	75	264
16	Consumer Federation of America	81.7	224	256
17	Christian Coalition	22.6	141	220
18	Cato Institute	36.3	224	196
19	National Organization for Women	78.9	62	195
20	Institute for International Economics	48.8	61	194
21	Urban Institute	73.8	186	187
22	Family Research Council	20.3	133	160
23	Federation of American Scientists	67.5	36	139
24	Economic Policy Institute	80.3	130	138
25	Center on Budget and Policy Priorities	88.3	224	115
26	National Right to Life Committee	21.6	81	109
27	Electronic Privacy Information Center	57.4	19	107
28	International Institute for Strategic Studies	41.2	16	104
29	World Wildlife Fund	50.4	130	101
30	Cent. for Strategic and Budgetary Assessments	33.9	7	89

Politicians

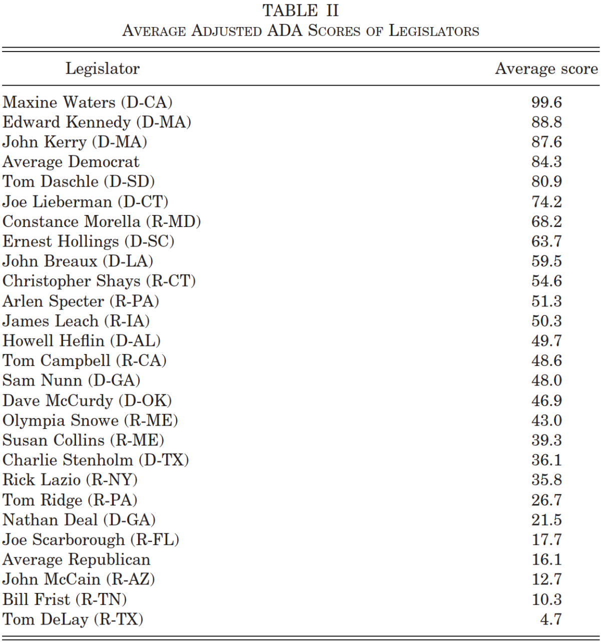


TABLE II

AVERAGE ADJUSTED ADA SCORES OF LEGISLATORS

Legislator	Average score
Maxine Waters (D-CA)	99.6
Edward Kennedy (D-MA)	88.8
John Kerry (D-MA)	87.6
Average Democrat	84.3
Tom Daschle (D-SD)	80.9
Joe Lieberman (D-CT)	74.2
Constance Morella (R-MD)	68.2
Ernest Hollings (D-SC)	63.7
John Breaux (D-LA)	59.5
Christopher Shays (R-CT)	54.6
Arlen Specter (R-PA)	51.3
James Leach (R-IA)	50.3
Howell Heflin (D-AL)	49.7
Tom Campbell (R-CA)	48.6
Sam Nunn (D-GA)	48.0
Dave McCurdy (D-OK)	46.9
Olympia Snowe (R-ME)	43.0
Susan Collins (R-ME)	39.3
Charlie Stenholm (D-TX)	36.1
Rick Lazio (R-NY)	35.8
Tom Ridge (R-PA)	26.7
Nathan Deal (D-GA)	21.5
Joe Scarborough (R-FL)	17.7
Average Republican	16.1
John McCain (R-AZ)	12.7
Bill Frist (R-TN)	10.3
Tom DeLay (R-TX)	4.7

Definition of Bias

- Nothing to do with honesty or accuracy; more like a preference
- The centrist US voter in the late 1990s had a left-right ideology approximately equal to Specter (R-PA; 51.3) or Nunn (D-GA; 48.0)
- The average NYT article has an ideology approximately equal to Joe Lieberman (D-CT; 74.2)
- Since Liberman is more liberal then Specter and Nunn, hence the NYT has a liberal bias
- Bias here reflects omission, selective choice of points of view.

Results

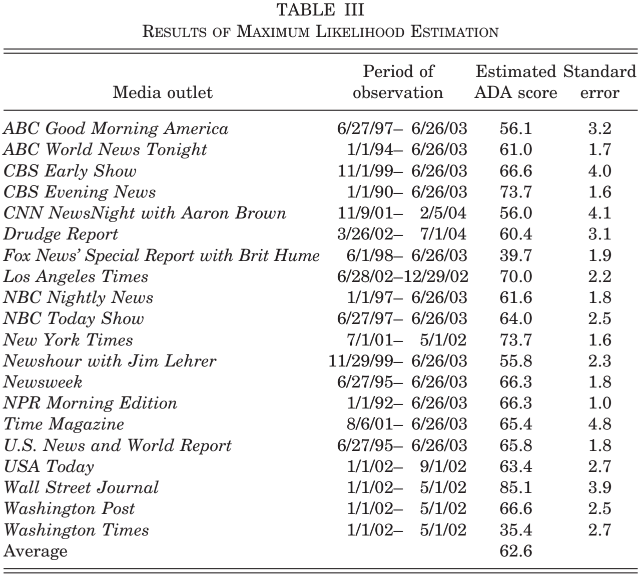
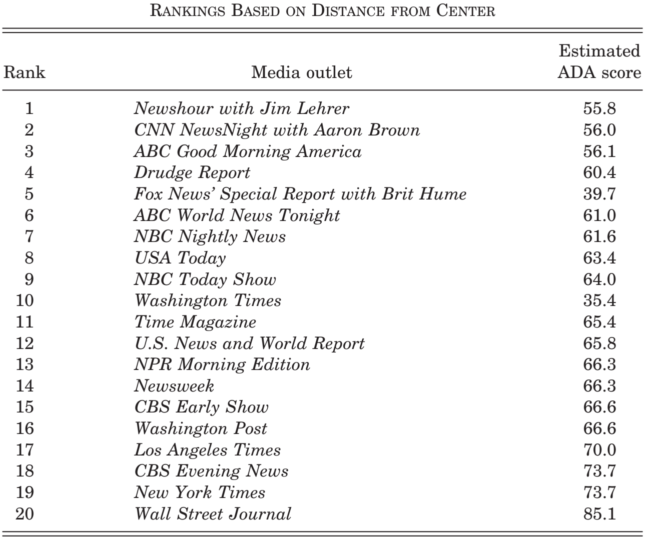


TABLE III
Results of Maximum Likelihood Estimation

Media outlet	Period of observation	Estimated ADA score	Standard error
ABC Good Morning America	6/27/97– 6/26/03	56.1	3.2
ABC World News Tonight	1/1/94– 6/26/03	61.0	1.7
CBS Early Show	11/1/99– 6/26/03	66.6	4.0
CBS Evening News	1/1/90– 6/26/03	73.7	1.6
CNN NewsNight with Aaron Brown	11/9/01– 2/5/04	56.0	4.1
Drudge Report	3/26/02– 7/1/04	60.4	3.1
Fox News' Special Report with Brit Hume	6/1/98– 6/26/03	39.7	1.9
Los Angeles Times	6/28/02–12/29/02	70.0	2.2
NBC Nightly News	1/1/97– 6/26/03	61.6	1.8
NBC Today Show	6/27/97– 6/26/03	64.0	2.5
New York Times	7/1/01– 5/1/02	73.7	1.6
Newshour with Jim Lehrer	11/29/99– 6/26/03	55.8	2.3
Newsweek	6/27/95– 6/26/03	66.3	1.8
NPR Morning Edition	1/1/92– 6/26/03	66.3	1.0
Time Magazine	8/6/01– 6/26/03	65.4	4.8
U.S. News and World Report	6/27/95– 6/26/03	65.8	1.8
USA Today	1/1/02– 9/1/02	63.4	2.7
Wall Street Journal	1/1/02– 5/1/02	85.1	3.9
Washington Post	1/1/02– 5/1/02	66.6	2.5
Washington Times	1/1/02– 5/1/02	35.4	2.7
Average		62.6	

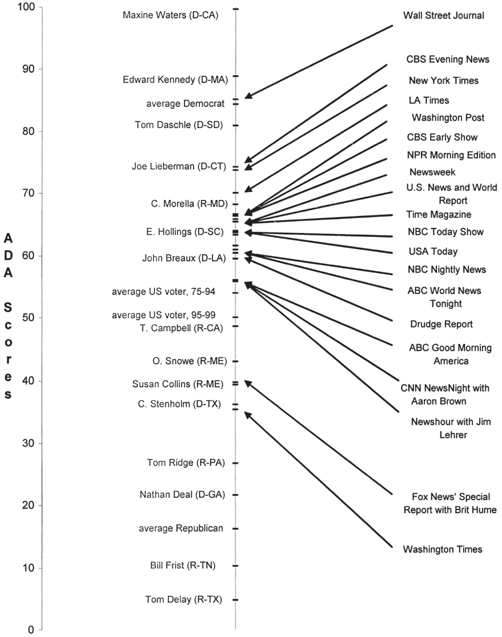
Results (cont.)



Rankings Based on Distance from Center

Rank	Media outlet	Estimated ADA score
1	*Newshour with Jim Lehrer*	55.8
2	*CNN NewsNight with Aaron Brown*	56.0
3	*ABC Good Morning America*	56.1
4	*Drudge Report*	60.4
5	*Fox News' Special Report with Brit Hume*	39.7
6	*ABC World News Tonight*	61.0
7	*NBC Nightly News*	61.6
8	*USA Today*	63.4
9	*NBC Today Show*	64.0
10	*Washington Times*	35.4
11	*Time Magazine*	65.4
12	*U.S. News and World Report*	65.8
13	*NPR Morning Edition*	66.3
14	*Newsweek*	66.3
15	*CBS Early Show*	66.6
16	*Washington Post*	66.6
17	*Los Angeles Times*	70.0
18	*CBS Evening News*	73.7
19	*New York Times*	73.7
20	*Wall Street Journal*	85.1

Results (cont.)



What Drives Media Slant (Gentzkow-Shapiro, 2010)

- Goal:
  - Construct a measure of media slant based on the similarity of the language used by newspapers and politicians
- Methodology:
  - Consider official speeches by US congressmen
  - Identify the two- and three-word expressions most representative of *Republicans* and *Democrats*
  - Compute the frequency of “Democratic” and “Republican” expressions in the articles published in over 400 newspapers
  - Define the slant of each newspaper with respect to politicians
  - Test whether slant is driven by consumers’ vs. owners’ preferences
- Findings:
  - Slant highly correlated with political leaning of potential readers
  - Identity of media owner does not explain much

What Drives Media Slant (Gentzkow-Shapiro, 2010)

- Goal:
  - Construct a measure of media slant based on the similarity of the language used by newspapers and politicians
- Methodology:
  - Consider official speeches by US congressmen
  - Identify the two- and three-word expressions most representative of *Republicans* and *Democrats*
  - Compute the frequency of "Democratic" and "Republican" expressions in the articles published in over 400 newspapers
  - Define the slant of each newspaper with respect to politicians
  - Test whether slant is driven by consumers' vs. owners' preferences
- Findings:
  - Slant highly correlated with political leaning of potential readers
  - Identity of media owner does not explain much

What Drives Media Slant (Gentzkow-Shapiro, 2010)

- Goal:
  - Construct a measure of media slant based on the similarity of the language used by newspapers and politicians
- Methodology:
  - Consider official speeches by US congressmen
  - Identify the two- and three-word expressions most representative of *Republicans* and *Democrats*
  - Compute the frequency of "Democratic" and "Republican" expressions in the articles published in over 400 newspapers
  - Define the slant of each newspaper with respect to politicians
  - Test whether slant is driven by consumers' vs. owners' preferences
- Findings:
  - Slant highly correlated with political leaning of potential readers
  - Identity of media owner does not explain much

Data

- Politicians:
  - All speeches from 2005 *Congressional Record*
  - Ideological score: vote share to Bush in 2004 presidential election in congressperson's constituency
- News content:
  - Headlines and text of all articles published on 433 US daily newspapers in 2005
  - Source: Newslibrary, ProQuest
  - Only news articles, no editorials
- Other:
  - Newspaper HQ location and relevant market (PMSA)
  - Vote shares for Republicans and Democrats in relevant media market
  - Identity of newspaper's owner

Data

- Politicians:
  - All speeches from 2005 *Congressional Record*
  - Ideological score: vote share to Bush in 2004 presidential election in congressperson's constituency
- News content:
  - Headlines and text of all articles published on 433 US daily newspapers in 2005
  - Source: Newslibrary, ProQuest
  - Only news articles, no editorials
- Other:
  - Newspaper HQ location and relevant market (PMSA)
  - Vote shares for Republicans and Democrats in relevant media market
  - Identity of newspaper's owner

Data

- Politicians:
  - All speeches from 2005 *Congressional Record*
  - Ideological score: vote share to Bush in 2004 presidential election in congressperson's constituency
- News content:
  - Headlines and text of all articles published on 433 US daily newspapers in 2005
  - Source: Newslibrary, ProQuest
  - Only news articles, no editorials
- Other:
  - Newspaper HQ location and relevant market (PMSA)
  - Vote shares for Republicans and Democrats in relevant media market
  - Identity of newspaper's owner

Methodology: step #1

- Pre-process the *Congressional Record* corpus: remove stop words, stemming (Porter)
- Consider all 2-word and 3-word phrases appearing in the corpus
- For each phrase p of length l, compute the total number of times it is used by Democrats and Republicans (f_{pld}, f_{plr})
- For each phrase compute the Pearson's χ^2 statistic:

$$\chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}$$

- χ^2: test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans.
- Simple to compute, only requires f_{pld} and f_{plr}. Preferable to other naive statistics such as the ratio of uses by D or R

Methodology: step #1

- Pre-process the *Congressional Record* corpus: remove stop words, stemming (Porter)
- Consider all 2-word and 3-word phrases appearing in the corpus
- For each phrase p of length l, compute the total number of times it is used by Democrats and Republicans (f_{pld}, f_{plr})
- For each phrase compute the Pearson's χ^2 statistic:

$$\chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}$$

- χ^2: test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans.
- Simple to compute, only requires f_{pld} and f_{plr}. Preferable to other naive statistics such as the ratio of uses by D or R

Methodology: step #1

- Pre-process the *Congressional Record* corpus: remove stop words, stemming (Porter)

- Consider all 2-word and 3-word phrases appearing in the corpus

- For each phrase p of length l, compute the total number of times it is used by Democrats and Republicans (f_{pld}, f_{plr})

- For each phrase compute the Pearson's χ^2 statistic:

$$\chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}$$

- χ^2: test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans.

- Simple to compute, only requires f_{pld} and f_{plr}. Preferable to other naive statistics such as the ratio of uses by D or R

Methodology: step #1

- Pre-process the *Congressional Record* corpus: remove stop words, stemming (Porter)
- Consider all 2-word and 3-word phrases appearing in the corpus
- For each phrase p of length l, compute the total number of times it is used by Democrats and Republicans (f_{pld}, f_{plr})
- For each phrase compute the Pearson's χ^2 statistic:

$$\chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}$$

- χ^2: test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans.
- Simple to compute, only requires f_{pld} and f_{plr}. Preferable to other naive statistics such as the ratio of uses by D or R

Methodology: step #1

- Pre-process the *Congressional Record* corpus: remove stop words, stemming (Porter)
- Consider all 2-word and 3-word phrases appearing in the corpus
- For each phrase p of length l, compute the total number of times it is used by Democrats and Republicans (f_{pld}, f_{plr})
- For each phrase compute the Pearson's χ^2 statistic:

$$\chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}$$

- χ^2: test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans.
- Simple to compute, only requires f_{pld} and f_{plr}. Preferable to other naive statistics such as the ratio of uses by D or R

Methodology: step #1

- Pre-process the *Congressional Record* corpus: remove stop words, stemming (Porter)

- Consider all 2-word and 3-word phrases appearing in the corpus

- For each phrase p of length l, compute the total number of times it is used by Democrats and Republicans (f_{pld}, f_{plr})

- For each phrase compute the Pearson's χ^2 statistic:

$$\chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}$$

- χ^2: test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans.

- Simple to compute, only requires f_{pld} and f_{plr}. Preferable to other naive statistics such as the ratio of uses by D or R

Methodology: step #2

- Eliminate phrases that are not likely to be useful for diagnosing newspaper partisanship.

- 2-word phrases appearing less than 200 times in newspaper headlines (e.g., procedural phrases not used by the media)

- 2-word phrases that appeared more than 15,000 times in headlines (e.g., "third quarter" or "exchange rate")

- 3-word phrases that appeared less than 5 times in headlines

- 3-word phrases that appeared more than 1,000 times in headlines

- Any phrase that appeared in the full text of more than 400,000 documents

- Among the remaining ones, select the 500 phrases of each length with the highest values of χ^2_{pl}

Methodology: step #2

- Eliminate phrases that are not likely to be useful for diagnosing newspaper partisanship.
- 2-word phrases appearing less than 200 times in newspaper headlines (e.g., procedural phrases not used by the media)
- 2-word phrases that appeared more than 15,000 times in headlines (e.g., “third quarter” or “exchange rate”)
- 3-word phrases that appeared less than 5 times in headlines
- 3-word phrases that appeared more than 1,000 times in headlines
- Any phrase that appeared in the full text of more than 400,000 documents
- Among the remaining ones, select the 500 phrases of each length with the highest values of χ^2_{pl}

Methodology: step #2

- Eliminate phrases that are not likely to be useful for diagnosing newspaper partisanship.
- 2-word phrases appearing less than 200 times in newspaper headlines (e.g., procedural phrases not used by the media)
- 2-word phrases that appeared more than 15,000 times in headlines (e.g., "third quarter" or "exchange rate")
- 3-word phrases that appeared less than 5 times in headlines
- 3-word phrases that appeared more than 1,000 times in headlines
- Any phrase that appeared in the full text of more than 400,000 documents
- Among the remaining ones, select the 500 phrases of each length with the highest values of χ^2_{pl}

Most representative Democratic phrases

Panel A: Phrases Used More Often by Democrats

Two-Word Phrases

private accounts
trade agreement
American people
tax breaks
trade deficit
oil companies
credit card
nuclear option
war in Iraq
middle class
Rosa Parks
President budget
Republican party
change the rules
minimum wage
budget deficit
Republican senators
privatization plan
wildlife refuge
card companies
workers rights
poor people
Republican leader
Arctic refuge
cut funding
American workers
living in poverty
Senate Republicans
fuel efficiency
national wildlife

Three-Word Phrases

veterans health care
congressional black caucus
VA health care
billion in tax cuts
credit card companies
security trust fund
social security trust
privatize social security
American free trade
central American free
corporation for public broadcasting
additional tax cuts
pay for tax cuts
tax cuts for people
oil and gas companies
prescription drug bill
caliber sniper rifles
increase in the minimum wage
system of checks and balances
middle class families
cut health care
civil rights movement
cuts to child support
drilling in the Arctic National
victims of gun violence
solvency of social security
Voting Rights Act
war in Iraq and Afghanistan
civil rights protections
credit card debt

Most representative Republican phrases

Panel B: Phrases Used More Often by Republicans

Two-Word Phrases

- stem cell
- natural gas
- death tax
- illegal aliens
- class action
- war on terror
- embryonic stem
- tax relief
- illegal immigration
- date the time
- personal accounts
- Saddam Hussein
- pass the bill
- private property
- border security
- President announces
- human life
- Chief Justice
- human embryos
- increase taxes
- retirement accounts
- government spending
- national forest
- minority leader
- urge support
- cell lines
- cord blood
- action lawsuits
- economic growth
- food program

Three-Word Phrases

- embryonic stem cell
- hate crimes legislation
- adult stem cells
- oil for food program
- personal retirement accounts
- energy and natural resources
- global war on terror
- hate crimes law
- change hearts and minds
- global war on terrorism
- Circuit Court of Appeals
- death tax repeal
- housing and urban affairs
- million jobs created
- national flood insurance
- oil for food scandal
- private property rights
- temporary worker program
- class action reform
- Chief Justice Rehnquist
- Tongass national forest
- pluripotent stem cells
- Supreme Court of Texas
- Justice Priscilla Owen
- Justice Janice Rogers
- American Bar Association
- growth and job creation
- natural gas natural
- Grand Ole Opry
- reform social security

Methodology: step #3

- Use politicians' language and ideology to map phrases to ideology
- Re-index the phrases in the sample by $p \in [1, ..., 1000]$
- For each congressperson $c \in C$ we observe ideology y_c and phrase frequencies $\{f_{pc}\}_{p=1}^{1000}$
- Compute relative frequencies as: $\tilde{f}_{pc} = f_{pc} / \sum_{p=1}^{P} f_{pc}$
- For each phrase p, we regress $\tilde{f}_{pc}$ on y_c for the sample of congresspeople, obtaining an intercept and a slope parameter, a_p and b_p
- Hence we estimate 1,000 separate regressions, each with a sample the size of C
- Intuition: the larger b_p the more the use of a phrase is correlated with ideology (y_c)

Methodology: step #3

- Use politicians' language and ideology to map phrases to ideology
- Re-index the phrases in the sample by $p \in [1, ..., 1000]$
- For each congressperson $c \in C$ we observe ideology y_c and phrase frequencies $\{f_{pc}\}_{p=1}^{1000}$
- Compute relative frequencies as: $\tilde{f}_{pc} = f_{pc} / \sum_{p=1}^{P} f_{pc}$
- For each phrase p, we regress $\tilde{f}_{pc}$ on y_c for the sample of congresspeople, obtaining an intercept and a slope parameter, a_p and b_p
- Hence we estimate 1,000 separate regressions, each with a sample the size of C
- Intuition: the larger b_p the more the use of a phrase is correlated with ideology (y_c)

Methodology: step #3

- Use politicians' language and ideology to map phrases to ideology
- Re-index the phrases in the sample by $p \in [1, ..., 1000]$
- For each congressperson $c \in C$ we observe ideology y_c and phrase frequencies $\{f_{pc}\}_{p=1}^{1000}$
- Compute relative frequencies as: $\tilde{f}_{pc} = f_{pc} / \sum_{p=1}^{P} f_{pc}$
- For each phrase p, we regress $\tilde{f}_{pc}$ on y_c for the sample of congresspeople, obtaining an intercept and a slope parameter, a_p and b_p
- Hence we estimate 1,000 separate regressions, each with a sample the size of C
- Intuition: the larger b_p the more the use of a phrase is correlated with ideology (y_c)

Methodology: step #3

- Use politicians' language and ideology to map phrases to ideology

- Re-index the phrases in the sample by $p \in [1, ..., 1000]$

- For each congressperson $c \in C$ we observe ideology y_c and phrase frequencies $\{f_{pc}\}_{p=1}^{1000}$

- Compute relative frequencies as: $\tilde{f}_{pc} = f_{pc} / \sum_{p=1}^{P} f_{pc}$

- For each phrase p, we regress $\tilde{f}_{pc}$ on y_c for the sample of congresspeople, obtaining an intercept and a slope parameter, a_p and b_p

- Hence we estimate 1,000 separate regressions, each with a sample the size of C

- Intuition: the larger b_p the more the use of a phrase is correlated with ideology (y_c)

Methodology: step #3

- Use politicians' language and ideology to map phrases to ideology
- Re-index the phrases in the sample by $p \in [1, ..., 1000]$
- For each congressperson $c \in C$ we observe ideology y_c and phrase frequencies $\{f_{pc}\}_{p=1}^{1000}$
- Compute relative frequencies as: $\tilde{f}_{pc} = f_{pc} / \sum_{p=1}^{P} f_{pc}$
- For each phrase p, we regress $\tilde{f}_{pc}$ on y_c for the sample of congresspeople, obtaining an intercept and a slope parameter, a_p and b_p
- Hence we estimate 1,000 separate regressions, each with a sample the size of C
- Intuition: the larger b_p the more the use of a phrase is correlated with ideology (y_c)

Methodology: step #3

- Use politicians' language and ideology to map phrases to ideology
- Re-index the phrases in the sample by $p \in [1, ..., 1000]$
- For each congressperson $c \in C$ we observe ideology y_c and phrase frequencies $\{f_{pc}\}_{p=1}^{1000}$
- Compute relative frequencies as: $\tilde{f}_{pc} = f_{pc} / \sum_{p=1}^{P} f_{pc}$
- For each phrase p, we regress $\tilde{f}_{pc}$ on y_c for the sample of congresspeople, obtaining an intercept and a slope parameter, a_p and b_p
- Hence we estimate 1,000 separate regressions, each with a sample the size of C
- Intuition: the larger b_p the more the use of a phrase is correlated with ideology (y_c)

Methodology: step #4

- We want to assign each paper a measure of slant based on the frequency of phrases it uses and their ideological valence

- For each paper $n \in N$, we observe the relative frequency for each phrase $\tilde{f}_{pc}$, but not the ideology y_n, which we want to estimate

- For each newspaper n we regress $(\tilde{f}_{pn} - a_p)$ on b_p for the sample of phrases, obtaining the slope estimate:

$$\hat{y}_n = \frac{\sum_{p=1}^{1000} b_p(\tilde{f}_{pn} - a_p)}{\sum_{p=1}^{1000} b_p^2}$$

- We estimate N separate regressions, each with a sample of 1,000

- Intuition: the higher the frequency $(\tilde{f}_{pn})$ of more ideological phrases (b_p), the higher the measure of slant $(\hat{y}_n)$

Methodology: step #4

- We want to assign each paper a measure of slant based on the frequency of phrases it uses and their ideological valence

- For each paper $n \in N$, we observe the relative frequency for each phrase $\tilde{f}_{pc}$, but not the ideology y_n, which we want to estimate

- For each newspaper n we regress $(\tilde{f}_{pn} - a_p)$ on b_p for the sample of phrases, obtaining the slope estimate:

$$\hat{y}_n = \frac{\sum_{p=1}^{1000} b_p(\tilde{f}_{pn} - a_p)}{\sum_{p=1}^{1000} b_p^2}$$

- We estimate N separate regressions, each with a sample of 1,000

- Intuition: the higher the frequency $(\tilde{f}_{pn})$ of more ideological phrases (b_p), the higher the measure of slant $(\hat{y}_n)$

Methodology: step #4

- We want to assign each paper a measure of slant based on the frequency of phrases it uses and their ideological valence

- For each paper $n \in N$, we observe the relative frequency for each phrase $\tilde{f}_{pc}$, but not the ideology y_n, which we want to estimate

- For each newspaper n we regress $(\tilde{f}_{pn} - a_p)$ on b_p for the sample of phrases, obtaining the slope estimate:

$$\hat{y}_n = \frac{\sum_{p=1}^{1000} b_p(\tilde{f}_{pn} - a_p)}{\sum_{p=1}^{1000} b_p^2}$$

- We estimate N separate regressions, each with a sample of 1,000

- Intuition: the higher the frequency ($\tilde{f}_{pn}$) of more ideological phrases (b_p), the higher the measure of slant ($\hat{y}_n$)

Methodology: step #4

- We want to assign each paper a measure of slant based on the frequency of phrases it uses and their ideological valence
- For each paper $n \in N$, we observe the relative frequency for each phrase $\tilde{f}_{pc}$, but not the ideology y_n, which we want to estimate
- For each newspaper n we regress $(\tilde{f}_{pn} - a_p)$ on b_p for the sample of phrases, obtaining the slope estimate:

$$\hat{y}_n = \frac{\sum_{p=1}^{1000} b_p(\tilde{f}_{pn} - a_p)}{\sum_{p=1}^{1000} b_p^2}$$

- We estimate N separate regressions, each with a sample of 1,000
- Intuition: the higher the frequency ($\tilde{f}_{pn}$) of more ideological phrases (b_p), the higher the measure of slant ($\hat{y}_n$)

Methodology: step #4

- We want to assign each paper a measure of slant based on the frequency of phrases it uses and their ideological valence

- For each paper $n \in N$, we observe the relative frequency for each phrase $\tilde{f}_{pc}$, but not the ideology y_n, which we want to estimate

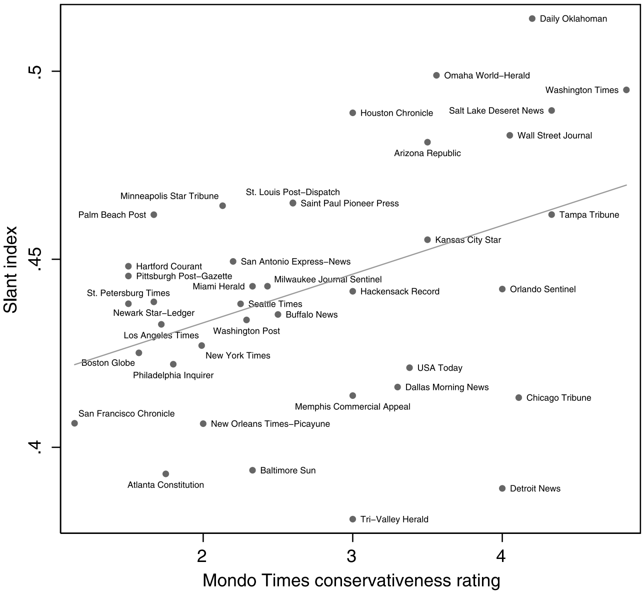
- For each newspaper n we regress $(\tilde{f}_{pn} - a_p)$ on b_p for the sample of phrases, obtaining the slope estimate:

$$\hat{y}_n = \frac{\sum_{p=1}^{1000} b_p(\tilde{f}_{pn} - a_p)}{\sum_{p=1}^{1000} b_p^2}$$

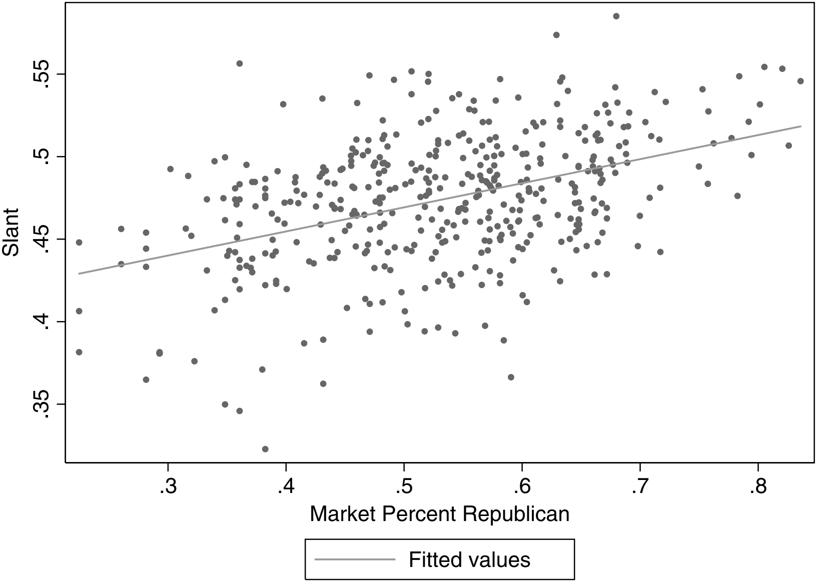
- We estimate N separate regressions, each with a sample of 1,000

- Intuition: the higher the frequency $(\tilde{f}_{pn})$ of more ideological phrases (b_p), the higher the measure of slant $(\hat{y}_n)$

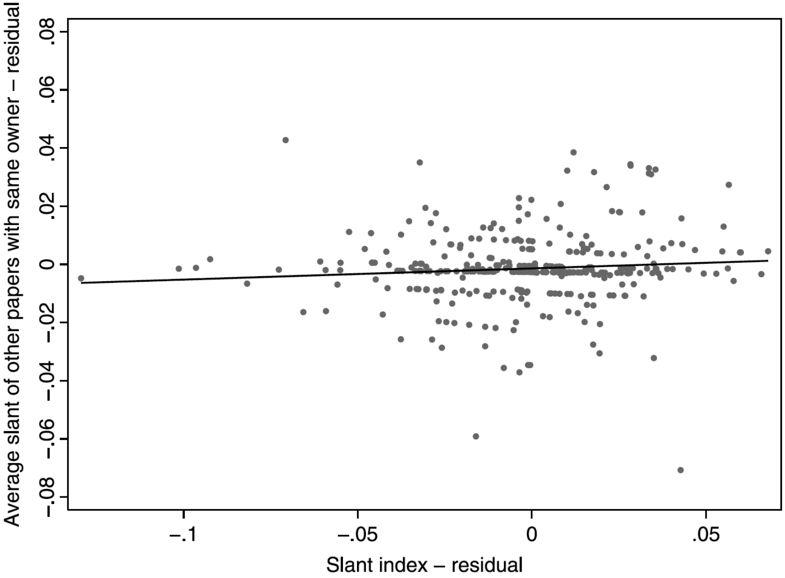
Validating the measure of slant



Using the measure: slant and readers' ideology



Using the measure: slant and owners' ideology



Using the measure: slant and owners' ideology

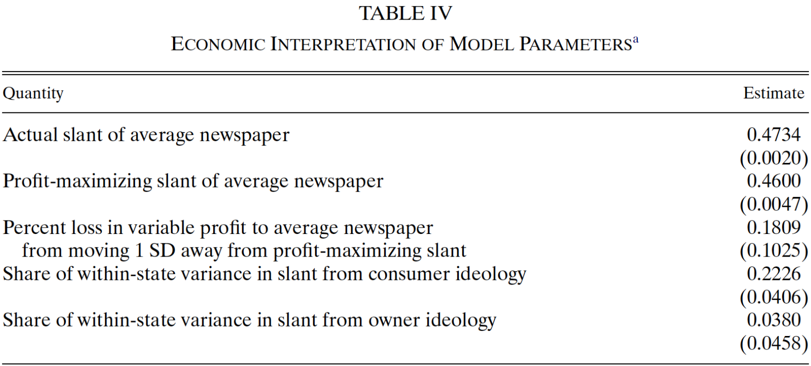


TABLE IV

ECONOMIC INTERPRETATION OF MODEL PARAMETERS[a]

Quantity	Estimate
Actual slant of average newspaper	0.4734
	(0.0020)
Profit-maximizing slant of average newspaper	0.4600
	(0.0047)
Percent loss in variable profit to average newspaper from moving 1 SD away from profit-maximizing slant	0.1809
	(0.1025)
Share of within-state variance in slant from consumer ideology	0.2226
	(0.0406)
Share of within-state variance in slant from owner ideology	0.0380
	(0.0458)

Production of Information in an Online World (Cage et al., 19)

- Goal:
  - Study how much online media outlets copy content from each other in the news production process
- Methodology:
  - Consider all online news content produced by French media in 2013
  - Identify 25K news events with an event-detection algorithm
  - Identify first news item that breaks news about an event
  - Measure how much copying is used by subsequent news stories related to event with plagiarism detection algorithm
  - Measure effects of copying on readership/audience
- Findings:
  - Online copying in news production is widespread: 61.8% of content presents some form of copying
  - Producing original content is rewarded with larger viewership shares

Production of Information in an Online World (Cage et al., 19)

- Goal:
  - Study how much online media outlets copy content from each other in the news production process
- Methodology:
  - Consider all online news content produced by French media in 2013
  - Identify 25K news events with an event-detection algorithm
  - Identify first news item that breaks news about an event
  - Measure how much copying is used by subsequent news stories related to event with plagiarism detection algorithm
  - Measure effects of copying on readership/audience
- Findings:
  - Online copying in news production is widespread: 61.8% of content presents some form of copying
  - Producing original content is rewarded with larger viewership shares

Production of Information in an Online World (Cage et al., 19)

- Goal:
  - Study how much online media outlets copy content from each other in the news production process
- Methodology:
  - Consider all online news content produced by French media in 2013
  - Identify 25K news events with an event-detection algorithm
  - Identify first news item that breaks news about an event
  - Measure how much copying is used by subsequent news stories related to event with plagiarism detection algorithm
  - Measure effects of copying on readership/audience
- Findings:
  - Online copying in news production is widespread: 61.8% of content presents some form of copying
  - Producing original content is rewarded with larger viewership shares

Data

- Online News Content:
  - More than 2.5 million French news articles published online in 2013 (7K/day)
  - Transmedia approach: content from 86 media outlets including 1 news agency (AFP), 59 newspapers, 10 online-only media outlets, 7 radio stations, 9 TV channels
  - Source: French National Audiovisual Institute (public company)
- Viewership:
  - Daily audience measures for 58 out of the 86 outlets (AFP and some local newspapers not covered)
  - Number of shares of the article on Facebook and Twitter. Proxy for the number of views of an article

Methodology: step #1 event detection algorithm

- Consider headline and text of each article and compute its TF-IDF vector representation
- Compute the cosine similarity of each article-pair
- Iteratively aggregate articles into event-clusters if the cosine similarity is above a certain threshold (determined manually)
- "Close" an event if no article is aggregated to it within a 24-hour window
- Drop events that contain less than 2 distinct media outlets and less than 10 articles.
- Results in 25,215 news events, each lasting about 41 hours
- 33.4% of of the 2.5M articles are classified into an event
- Very large clusters are mostly "garbage clusters"

Methodology: step #2 plagiarism detection algorithm

- Consider all the articles within an event-cluster and order them by time. The first one is the news-breaking article.
- Define the reaction time of an article as the time elapsed since publication of the news-breaking article
- Compare the text of each article to that of all the preceding articles in the event-cluster
- If a portion of text of at least 100 characters in the article is identical to any previous portion that is already published, then that portion is defined as a copy
- For each article, calculate the originality rate, defined as:

$$\text{originality rate } = \frac{\#\text{ original characters}}{\#\text{ total characters}}$$

Summary statistics

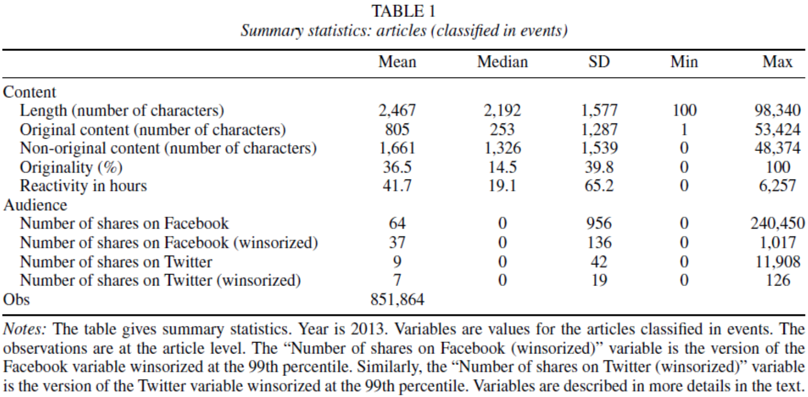
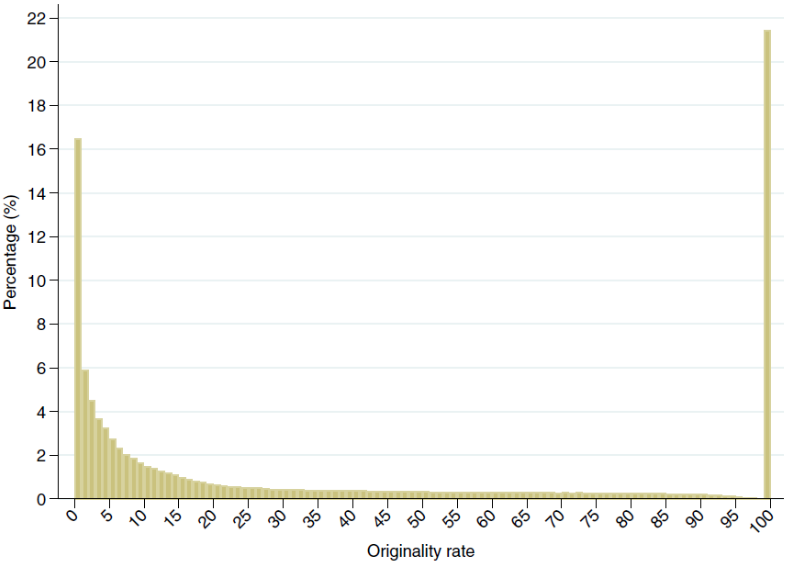


TABLE 1
Summary statistics: articles (classified in events)

	Mean	Median	SD	Min	Max
Content					
Length (number of characters)	2,467	2,192	1,577	100	98,340
Original content (number of characters)	805	253	1,287	1	53,424
Non-original content (number of characters)	1,661	1,326	1,539	0	48,374
Originality (%)	36.5	14.5	39.8	0	100
Reactivity in hours	41.7	19.1	65.2	0	6,257
Audience					
Number of shares on Facebook	64	0	956	0	240,450
Number of shares on Facebook (winsorized)	37	0	136	0	1,017
Number of shares on Twitter	9	0	42	0	11,908
Number of shares on Twitter (winsorized)	7	0	19	0	126
Obs	851,864				

Notes: The table gives summary statistics. Year is 2013. Variables are values for the articles classified in events. The observations are at the article level. The "Number of shares on Facebook (winsorized)" variable is the version of the Facebook variable winsorized at the 99th percentile. Similarly, the "Number of shares on Twitter (winsorized)" variable is the version of the Twitter variable winsorized at the 99th percentile. Variables are described in more details in the text.

Originality rate distribution



Positive relationship between originality and reaction time

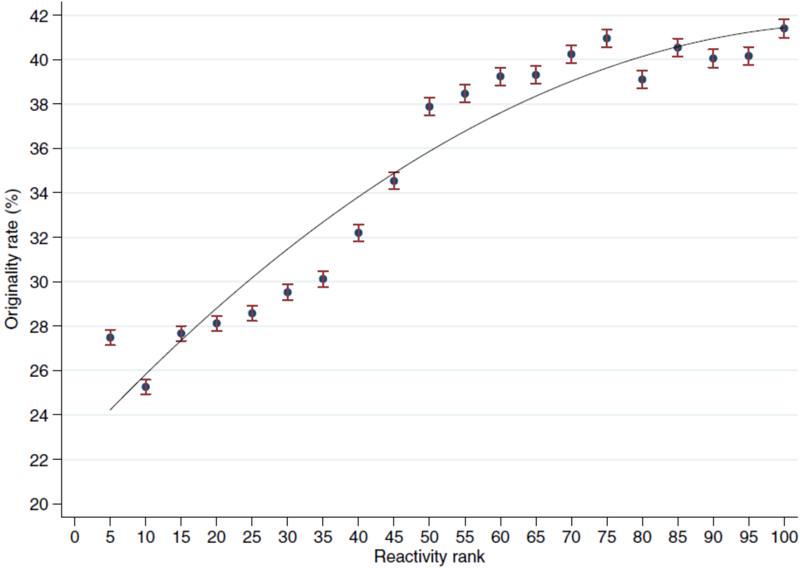


FIGURE 3

Correlation between originality and reaction time: average originality rate depending on the reactivity rank

Methodology: step #3 strategies to proxy article views

- "Naive" approach:
  - On a given day and for a given outlet, assume all articles are equally popular
  - Number of views is naively defined as number of page views on website divided by number of articles
- Linear approach:
  - Number of article views is proportional to the relative number of shares of the article, within its outlet-day pair
- Social media approach:
  - Collect data from leading French newspaper *Le Monde* on article views from April to August 2017
  - Link the URL of the online article to Facebook (resp. Twitter) data in order to identify the number of shares on social media
  - Examine correlation between the number of views on *Le Monde* website and the number of social media shares
  - Use relationship to infer the number of views for other outlets

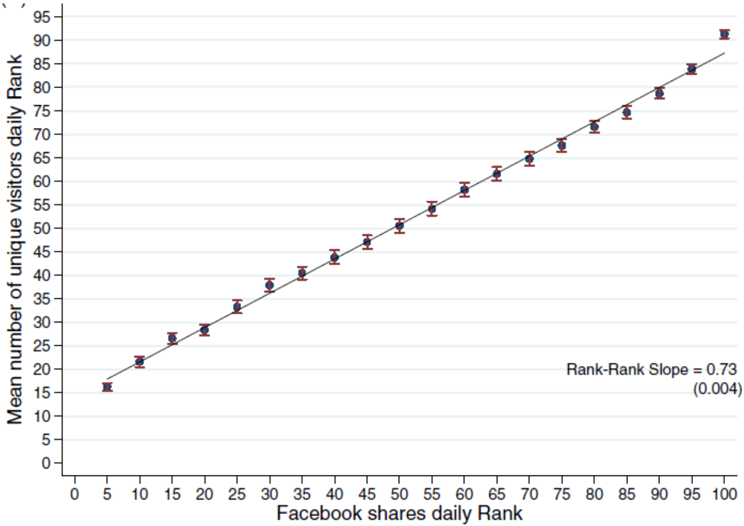
Methodology: step #3 strategies to proxy article views

- "Naive" approach:
  - On a given day and for a given outlet, assume all articles are equally popular
  - Number of views is naively defined as number of page views on website divided by number of articles
- Linear approach:
  - Number of article views is proportional to the relative number of shares of the article, within its outlet-day pair
- Social media approach:
  - Collect data from leading French newspaper *Le Monde* on article views from April to August 2017
  - Link the URL of the online article to Facebook (resp. Twitter) data in order to identify the number of shares on social media
  - Examine correlation between the number of views on *Le Monde* website and the number of social media shares
  - Use relationship to infer the number of views for other outlets

Methodology: step #3 strategies to proxy article views

- “Naive” approach:
  - On a given day and for a given outlet, assume all articles are equally popular
  - Number of views is naively defined as number of page views on website divided by number of articles
- Linear approach:
  - Number of article views is proportional to the relative number of shares of the article, within its outlet-day pair
- Social media approach:
  - Collect data from leading French newspaper *Le Monde* on article views from April to August 2017
  - Link the URL of the online article to Facebook (resp. Twitter) data in order to identify the number of shares on social media
  - Examine correlation between the number of views on *Le Monde* website and the number of social media shares
  - Use relationship to infer the number of views for other outlets

Social media shares and actual online views



Association between number of Unique visitors' and Facebook shares' Percentile Ranks

Positive relationship between popularity and originality

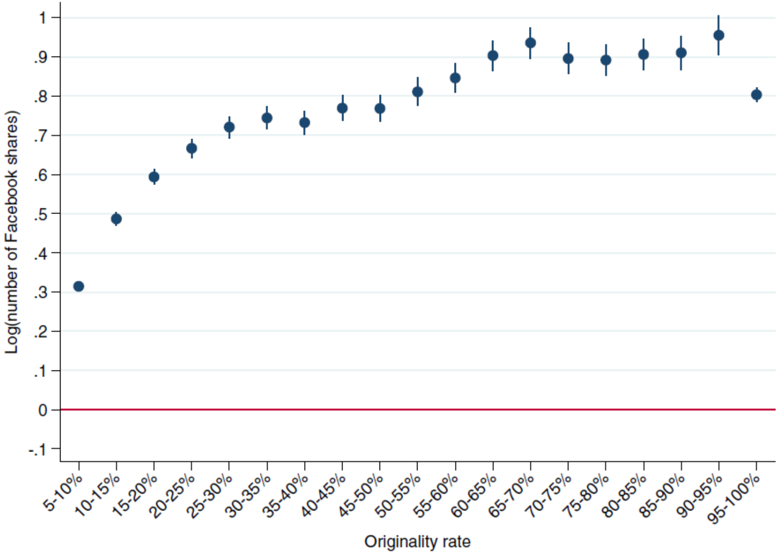


FIGURE 5

Facebook shares and originality rate

Share of original content in articles data

$$\frac{\sum_a \text{original content}_a \cdot \text{number of views}_a}{\sum_a \text{original content}_a \cdot \text{number of views}_a + \sum_a \text{non-original content}_a \cdot \text{number of views}_a}$$

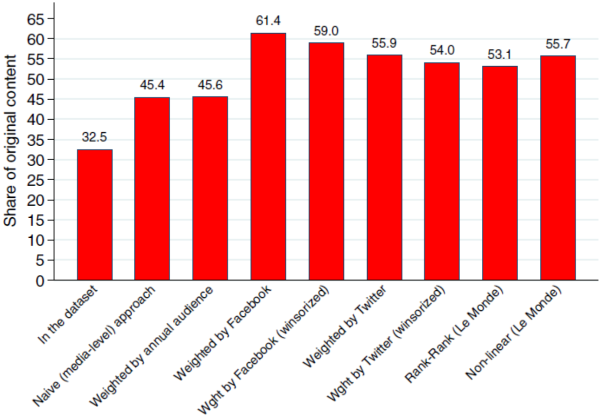


FIGURE 6

The audience-weighted share of original content