DSME 6635: Artificial Intelligence for Business Research

Traditional NLP: Downstream Tasks, N-Gram, Naïve Bayes, and Language Model Evaluation

Renyu (Philip) Zhang

1

Agenda

- NLP Downstream Tasks
- N-Gram and Naïve Bayes
- Traditional NLP Applications in Business/Econ Research

2

2

Text as Data

Journal of Economic Literature 2019, 57(3), 535–574
https://doi.org/10.1257/jel.20181020

Text as Data†

Matthew Gentzkow, Bryan Kelly, and Matt Taddy*

An ever-increasing share of human interaction, communication, and culture is recorded as digital text. We provide an introduction to the use of text as an input to economic research. We discuss the features that make text different from other forms of data, offer a practical overview of relevant statistical methods, and survey a variety of applications. (JEL C38, C55, L82, Z13)

0. Pre-processing;

1. Represent raw text $\mathcal{D}$ as a numerical array $\mathbf{C}$;
2. Map $\mathbf{C}$ to predicted values $\hat{\mathbf{V}}$ of unknown outcomes $\mathbf{V}$; and
3. Use $\hat{\mathbf{V}}$ in subsequent descriptive or causal analysis.

3

3

Downstream Task: Sentiment Classification

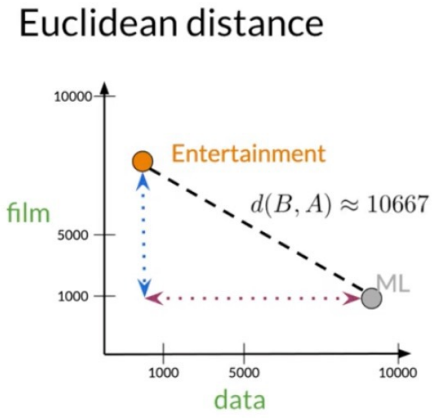
- Now your data is pre-processed from: Sentence_i --> Label_i = positive or negative
- To: (X1_i, X2_i, X3_i) --> y_i = 1 or 0, where X is the word/sentence/document-representation vector.
- We can use different ML methods to find the mapping between X and y. For example,
  - Linear probability model;
  - Logistic regression or other generalized linear models;
  - Deep learning.

4

4

Downstream Task: Semantic Similarities

- How do we measure similarities? Distance or angle.



Corpus A: (500,7000)

Corpus B: (9320,1000)

$$d(B, A) = \sqrt{(B_1 - A_1)^2 + (B_2 - A_2)^2}$$

$$c^2 = a^2 + b^2$$

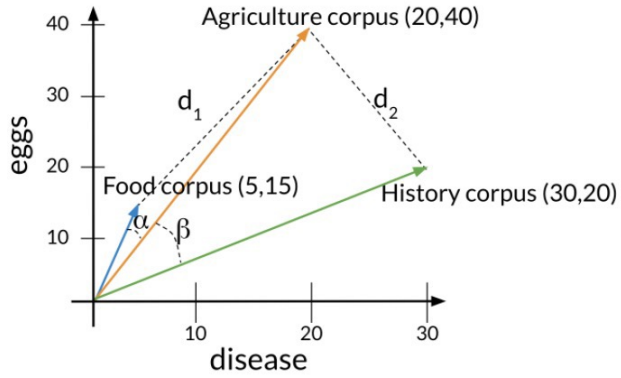
$$d(B, A) = \sqrt{(8820)^2 + (-6000)^2}$$

5

5

Downstream Task: Semantic Similarities

- How do we measure similarities? Distance or angle.



Euclidean distance: $d_2 < d_1$

Angles comparison: $\beta > \alpha$

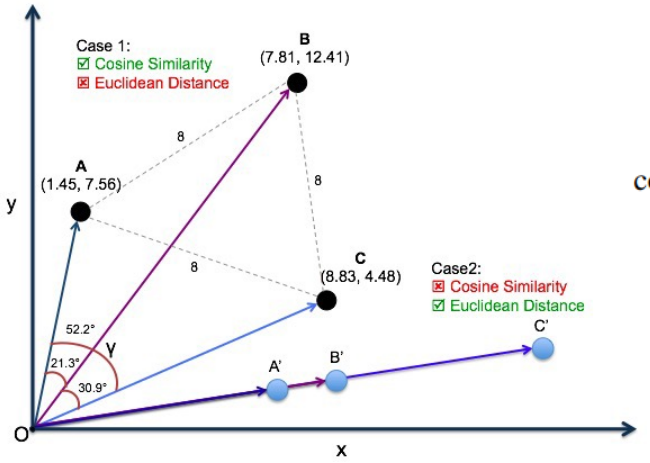
The cosine of the angle between the vectors

6

6

Downstream Task: Semantic Similarities

- Cosine Similarity: distance(OA, OC) > distance (OA, OB)
- Euclidean Similarity: distance(OA, OC) < distance(OA, OB)

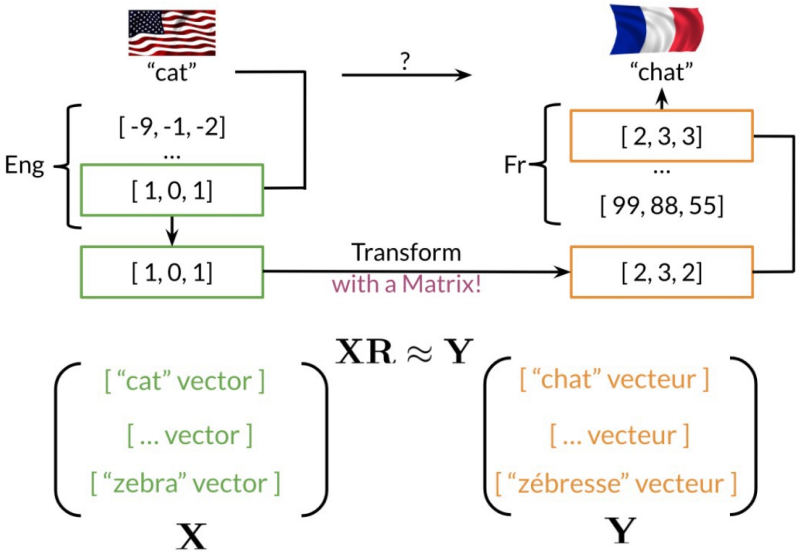


$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

7

7

Downstream Task: Machine Translation



8

8

Putting Things Together

1. **Pre-processing**: Take a dataset and do text normalization.
2. **Word-representation**: Use your favorite way to do word representation.
3. **Sentence/Document-representation**: Use your favorite way to do sentence/document representation.
4. **Downstream-task**: Use your favorite model to do downstream tasks.

9

9

Agenda

- NLP Downstream Tasks
- N-Gram and Naïve Bayes
- Traditional NLP Applications in Business/Econ Research

10

10

Bayesian Perspective: N-Gram

- Reference: https://web.stanford.edu/~jurafsky/slp3/, Chapter 2.
- Language Model: To find the probability of a word given the entire history of the sentence so far.

$$Pr(W_n|W_1,W_2,W_3,\ldots,W_{n-1})$$

- For example, suppose the sentence is "its water is so transparent that the…"

$$P(\text{the}\,|\,\text{its water is so transparent that}) = \frac{Count(\text{its water is so transparent that the})}{Count(\text{its water is so transparent that})}$$

- This model is too complex and suffers from the curse of dimensionality.
  - The number of combinations of word history grow exponentially with text length, and it requires prohibitively large datasets to compute.

11

11

Bayesian Perspective: N-Gram

- N-gram model refines the idea by limiting the dependencies on history.
- The N-gram model leverages the chain rule of probability:

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \,|\, w_1 w_2 \ldots w_{i-1})$$

P("its water is so transparent") =
P(its) × P(water|its) × P(is|its water)
× P(so|its water is) × P(transparent|its water is so)

12

12

Bayesian Perspective: N-gram

- N-gram model refines the idea by limiting the dependencies on history, a.k.a. **Markov Chain**.
- Unigram: $$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i)$$
- Bi-gram: $$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-1})$$
- We can extend to trigrams, 4-grams, 5-grams, etc., but there is a trade-off:
  - A larger N implies a longer-distance dependency.
  - A larger N also implies much more data to estimate the conditional probabilities.

13

13

Estimating Bigram

Maximum Likelihood Estimation (MLE) → $$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

An Example →

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

$P(\texttt{I} \mid \texttt{<s>}) = \frac{2}{3} = .67$ $P(\texttt{Sam} \mid \texttt{<s>}) = \frac{1}{3} = .33$ $P(\texttt{am} \mid \texttt{I}) = \frac{2}{3} = .67$

$P(\texttt{</s>} \mid \texttt{Sam}) = \frac{1}{2} = 0.5$ $P(\texttt{Sam} \mid \texttt{am}) = \frac{1}{2} = .5$ $P(\texttt{do} \mid \texttt{I}) = \frac{1}{3} = .33$

14

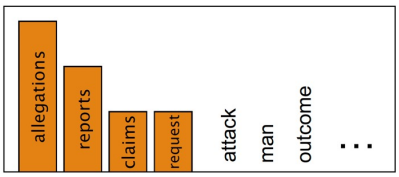
14

Smoothing

- What do we do if a word appears in the test set but not in the training set?
- The current n-gram model will fail (because the probability is 0).
- To keep a language model from assigning zero probability to these unseen events, we'll have to shave off a bit of probability mass from some more frequent events and give it to the events we've never seen. This is called smoothing.

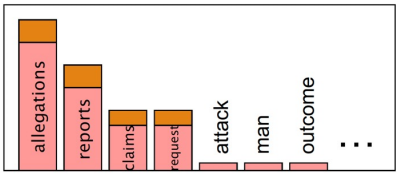
When we have sparse statistics:

P(w | denied the)
3 allegations
2 reports
1 claims
1 request
7 total



Steal probability mass to generalize better

P(w | denied the)
2.5 allegations
1.5 reports
0.5 claims
0.5 request
2 other
7 total



15

15

Laplace/Add-One Smoothing

Pretend that we saw each word one more time than we did.

$$P_{MLE}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

$$P_{Add-1}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i) + 1}{c(w_{i-1}) + V}$$

See (w_{i-1},w_i) one more time.

See each of (w_{i-1},w_k) one more time for all w_k in the vocabulary.

16

16

More about N-gram

- If we have 10,000 unique words, we will have 10,000*10,000=10e8 possible combinations of bigram.
- How about we use tri-gram or even 4-gram?
- Some toolkits for n-gram models:
  - SRILM: http://www.speech.sri.com/projects/srilm/
  - KenLM: https://kheafield.com/code/kenlm/
- Google n-gram viewer: https://books.google.com/ngrams/
  - Dataset: https://storage.googleapis.com/books/ngrams/books/datasetsv2.html

What drives media slant? Evidence from US daily newspapers
M Gentzkow, JM Shapiro
Econometrica, 2010 · Wiley Online Library

We construct a new index of media slant that measures the similarity of a news outlet's language to that of a congressional Republican or Democrat. We estimate a model of newspaper demand that incorporates slant explicitly, estimate the slant that would be chosen if newspapers independently maximized their own profits, and compare these profit-maximizing points with firms' actual choices. We find that readers have an economically significant preference for like-minded news. Firms respond strongly to

SHOW MORE

☆ Save 99 Cite Cited by 2330 Related articles All 33 versions Web of Science: 710 »

Measuring group differences in high-dimensional choices: method and application to congressional speech
M Gentzkow, JM Shapiro, M Taddy
Econometrica, 2019 · Wiley Online Library

We study the problem of measuring group differences in choices when the dimensionality of the choice set is large. We show that standard approaches suffer from a severe finite-sample bias, and we propose an estimator that applies recent advances in machine learning to address this bias. We apply this method to measure trends in the partisanship of congressional speech from 1873 to 2016, defining partisanship to be the ease with which an observer could infer a congressperson's party from a single utterance. Our

SHOW MORE

☆ Save 99 Cite Cited by 372 Related articles All 14 versions Web of Science: 95 »

17

17

Bayesian Perspective: Naïve Bayes

- Reference: https://web.stanford.edu/~jurafsky/slp3/, Chapter 3.
- Text classification is a fundamental application in NLP.
  - Sentiment analysis
  - Spam detection
  - Authorship identification
  - Language identification
  - Assigning subject categories, topics, or genres
  - Many more……
- Input:
  - A document d
  - A fixed set of classes $C=\{c_1,c_2,\ldots,c_J\}$
- Output:
  - A predicted class of document d in C.
- For a document d and a class c:

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Dropping the denominator

18

18

Bayesian Perspective: Naïve Bayes

- Reference: https://web.stanford.edu/~jurafsky/slp3/, Chapter 3.

"Likelihood" "Prior"

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c)P(c)$$

Document d represented as features x1..xn

O(|X|^n•|C|) parameters

How often does this class occur?

Could only be estimated if a very, very large number of training examples was available.

We can just count the relative frequencies in a corpus

19

19

Bayesian Perspective: Naïve Bayes

- Reference: https://web.stanford.edu/~jurafsky/slp3/, Chapter 3.
- Bag of words assumption: Position does not matter; all words are of similar importance (you may apply TF-IDF instead).
- Conditional independence: The feature probabilities Pr(x_i|c) are independent conditioned on class c.

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

positions ← all word positions in test document

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c)P(c)$$

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

20

20

Bag of Words Assumption

- Reference: https://web.stanford.edu/~jurafsky/slp3/, Chapter 3.

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

21

21

Naïve Bayes Estimation

- Reference: https://web.stanford.edu/~jurafsky/slp3/, Chapter 3.
- Maximum likelihood estimates:

The fraction of documents in class c_j

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

The fraction of times word w_i appears among all words in documents of class c_j

$$c_{MAP} = \text{argmax}_c \, \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

Most likely class of the document

22

22

Naïve Bayes Example

- Reference: https://web.stanford.edu/~jurafsky/slp3/, Chapter 3.

Positive tweets

I am happy because I am learning NLP

I am happy, not sad.

Negative tweets

I am sad, I am not learning NLP

I am sad, not happy

word	Pos	Neg
I	3	3
am	3	3
happy	2	1
because	1	0
learning	1	1
NLP	1	1
sad	1	2
not	1	2
N_{class}	13	12

word	Pos	Neg
I	0.24	0.25
am	0.24	0.25
happy	0.15	0.08
because	0.08	0
learning	0.08	0.08
NLP	0.08	0.08
sad	0.08	0.17
not	0.08	0.17

Let's classify the following tweets as positive or negative:

1. I am not sad.
2. I am learning NLP.

23

23

Naïve Bayes (Laplace) Smoothing

- Reference: https://web.stanford.edu/~jurafsky/slp3/, Chapter 3.
- Like n-gram, we need to normalize the estimated probabilities and bound them away from 0.

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) + 1 \right)}$$

$$= \frac{count(w_i, c) + 1}{\left(\sum_{w \in V} count(w, c) \right) + |V|}$$

See the word w_i in class c one more time.

See each word w of the vocabulary in class c one more time.

24

24

Putting Things Together for Naïve Bayes

1. **Labeling**: Annotate a dataset with class labels (positive, negative, etc.).
2. **Pre-processing**: Pre-process the text to words.
3. **Frequency computation**: Compute the Freq(word, class).
4. **Probability computation**: Compute P(word|class) and P(document|class).
5. **Classification**.

25

25

Remarks on Pre-processing

- All these preprocessing techniques are **standardized cross models**.
- Regardless whether you are using deep learning or other methods, the preprocessing techniques are similar.
- The **fundamental ways of thinking about words (probabilistic linguistic model)** is also similar across models.
- These techniques are much more useful **when you use your own methods** (to do tasks that are not accomplished in CS before).

26

26

Agenda

- NLP Downstream Tasks
- N-Gram and Naïve Bayes
- Traditional NLP Applications in Business/Econ Research

27

27

Application: Authorship Identification

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 302 JUNE, 1963 *Volume 58*

INFERENCE IN AN AUTHORSHIP PROBLEM[1,2]

A comparative study of discrimination methods applied to the authorship of the disputed *Federalist* papers

FREDERICK MOSTELLER
Harvard University
and
Center for Advanced Study in the Behavioral Sciences
AND
DAVID L. WALLACE
University of Chicago

This study has four purposes: to provide a comparison of discrimination methods; to explore the problems presented by techniques based strongly on Bayes' theorem when they are used in a data analysis of large scale; to solve the authorship question of *The Federalist* papers; and to propose routine methods for solving other authorship problems.

Word counts are the variables used for discrimination. Since the topic written about heavily influences the rate with which a word is used, care in selection of words is necessary. The filler words of the language such as *an*, *of*, and *upon*, and, more generally, articles, prepositions, and conjunctions provide fairly stable rates, whereas more meaningful words like *war*, *executive*, and *legislature* do not.

After an investigation of the distribution of these counts, the authors execute an analysis employing the usual discriminant function and an analysis based on Bayesian methods. The conclusions about the authorship problem are that Madison rather than Hamilton wrote all 12 of the disputed papers.

The findings about methods are presented in the closing section on conclusions.

This report, summarizing and abbreviating a forthcoming monograph [8], gives some of the results but very little of their empirical and theoretical foundation. It treats two of the four main studies presented in the monograph, and none of the side studies.

- Who wrote the *Federalist Papers*, Alexander Hamilton or James Madison?
- Applying Naïve Bayes, where the class c is either Hamilton or Madison, Mosteller and Wallace (1963) find overwhelming evidence that the disputed papers were authored by Madison.
- A similar method was applied to identify who invented instrumental variables estimator.

[PDF] **Retrospectives**: Who **invented instrumental variable** regression?
JH Stock, F Trebbi - Journal of Economic Perspectives, 2003 - pubs.aeaweb.org
… derivations of the **instrumental variables** estimators of the … B was showing that **instrumental variables** regression can be … do, which makes **instrumental variables** regression a central …
☆ Save 99 Cite Cited by 242 Related articles All 13 versions Web of Science: 82 »

28

28

Application: Sentiment and Stock Price

THE JOURNAL OF FINANCE • VOL. LXII, NO. 3 • JUNE 2007

Giving Content to Investor Sentiment: The Role of Media in the Stock Market

PAUL C. TETLOCK*

ABSTRACT

I quantitatively measure the interactions between the media and the stock market using daily content from a popular *Wall Street Journal* column. I find that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals, and unusually high or low pessimism predicts high market trading volume. These and similar results are consistent with theoretical models of noise and liquidity traders, and are inconsistent with theories of media content as a proxy for new information about fundamental asset values, as a proxy for market volatility, or as a sideshow with no relationship to asset markets.

- Based on the dictionary approach, convert the word counts in WSJ's "Abstract of the Market" into sentiment scores and condense them into a single principal component, called "pessimism factor".
- This pessimism score is used to forecast stock market activity.

When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks
T Loughran, B McDonald - The Journal of finance, 2011 - Wiley Online Library
… this paper **is** whether a … **is** inherently imprecise, we provide evidence based on 50,115 firm-year 10-Ks between 1994 and 2008 that the H4N list substantially misclassifies words **when** …
☆ Save 99 Cite Cited by 5231 Related articles All 8 versions Web of Science: 1875 »

Giving content to **investor sentiment**: The role of media in the stock market
PC Tetlock - The Journal of finance, 2007 - Wiley Online Library
… **investor sentiment** or … **investor sentiment**, resulting in downward pressure on prices. It is unclear whether media pessimism forecasts **investor sentiment** or reflects past **investor sentiment**…
☆ Save 99 Cite Cited by 5098 Related articles All 18 versions Web of Science: 1858 »

29

29

Application: Measuring Policy Uncertainty

THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. 131 November 2016 Issue 4

MEASURING ECONOMIC POLICY UNCERTAINTY*

SCOTT R. BAKER
NICHOLAS BLOOM
STEVEN J. DAVIS

We develop a new index of economic policy uncertainty (EPU) based on newspaper coverage frequency. Several types of evidence—including human readings of 12,000 newspaper articles—indicate that our index proxies for movements in policy-related economic uncertainty. Our U.S. index spikes near tight presidential elections, Gulf Wars I and II, the 9/11 attacks, the failure of Lehman Brothers, the 2011 debt ceiling dispute, and other major battles over fiscal policy. Using firm-level data, we find that policy uncertainty is associated with greater stock price volatility and reduced investment and employment in policy-sensitive sectors like defense, health care, finance, and infrastructure construction. At the macro level, innovations in policy uncertainty foreshadow declines in investment, output, and employment in the United States and, in a panel vector autoregressive setting, for 12 major economies. Extending our U.S. index back to 1900, EPU rose dramatically in the 1930s (from late 1931) and has drifted upward since the 1960s. *JEL Codes:* D80, E22, E66, G18, L50.

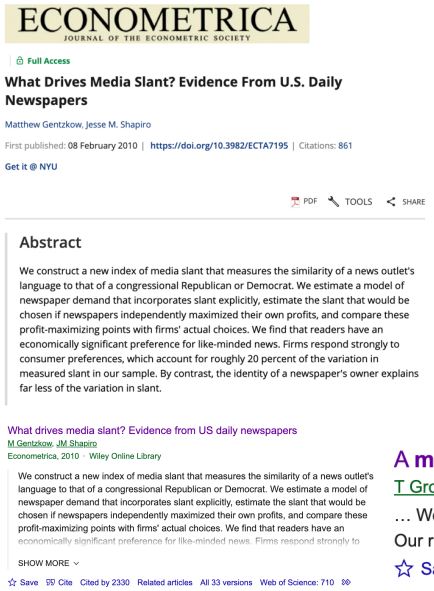
- Based on the dictionary approach, count the number of news articles containing at least one key word from the tree categories: economy, policy, and uncertainty. Use these counts to predict the level of economic policy uncertainty, defined as the simple average of the counts across different newspapers
- The created index is validated by a human audit, i.e., it is highly correlated with the human-coded index.

Measuring economic policy uncertainty
SR Baker, N Bloom, SJ Davis - … quarterly journal of **economics**, 2016 - academic.oup.com
… battles over fiscal **policy**. Using firm-level data, we find that **policy uncertainty** is associated with greater stock price volatility and reduced investment and employment in **policy**-sensitive …
☆ Save 99 Cite Cited by 10872 Related articles All 53 versions Web of Science: 4167 »

30

30

Application: Media Slant



ECONOMETRICA
JOURNAL OF THE ECONOMETRIC SOCIETY

Full Access

What Drives Media Slant? Evidence From U.S. Daily Newspapers

Matthew Gentzkow, Jesse M. Shapiro

First published: 08 February 2010 | https://doi.org/10.3982/ECTA7195 | Citations: 861

Get it @ NYU

PDF TOOLS SHARE

Abstract

We construct a new index of media slant that measures the similarity of a news outlet's language to that of a congressional Republican or Democrat. We estimate a model of newspaper demand that incorporates slant explicitly, estimate the slant that would be chosen if newspapers independently maximized their own profits, and compare these profit-maximizing points with firms' actual choices. We find that readers have an economically significant preference for like-minded news. Firms respond strongly to consumer preferences, which account for roughly 20 percent of the variation in measured slant in our sample. By contrast, the identity of a newspaper's owner explains far less of the variation in slant.

- Produce the 2-grams and 3-grams by speaker and select the top one thousand phrases through a chi-square test to identify the frequently and asymmetrically used phrases by Democrats and Republicans.

- Predict newspaper slant from the counts of the selected phrases by a two-stage supervised generative method.

What drives media slant? Evidence from US daily newspapers

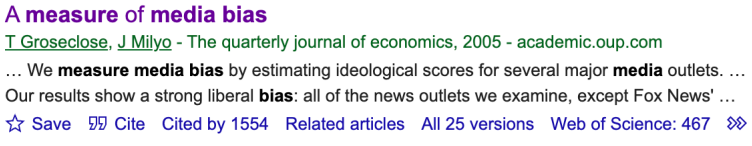
M Gentzkow, JM Shapiro

Econometrica, 2010 · Wiley Online Library

We construct a new index of media slant that measures the similarity of a news outlet's language to that of a congressional Republican or Democrat. We estimate a model of newspaper demand that incorporates slant explicitly, estimate the slant that would be chosen if newspapers independently maximized their own profits, and compare these profit-maximizing points with firms' actual choices. We find that readers have an economically significant preference for like-minded news. Firms respond strongly to

SHOW MORE

☆ Save 99 Cite Cited by 2330 Related articles All 33 versions Web of Science: 710 »



A measure of media bias

T Groseclose, J Milyo - The quarterly journal of economics, 2005 - academic.oup.com

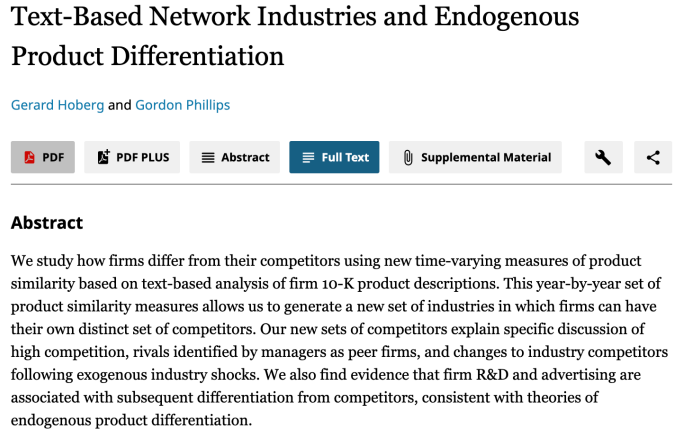
… We **measure media bias** by estimating ideological scores for several major **media** outlets. … Our results show a strong liberal **bias**: all of the news outlets we examine, except Fox News' …

☆ Save 99 Cite Cited by 1554 Related articles All 25 versions Web of Science: 467 »

31

31

Application: Industry Segmentation



Text-Based Network Industries and Endogenous Product Differentiation

Gerard Hoberg and Gordon Phillips

PDF PDF PLUS Abstract Full Text Supplemental Material

Abstract

We study how firms differ from their competitors using new time-varying measures of product similarity based on text-based analysis of firm 10-K product descriptions. This year-by-year set of product similarity measures allows us to generate a new set of industries in which firms can have their own distinct set of competitors. Our new sets of competitors explain specific discussion of high competition, rivals identified by managers as peer firms, and changes to industry competitors following exogenous industry shocks. We also find evidence that firm R&D and advertising are associated with subsequent differentiation from competitors, consistent with theories of endogenous product differentiation.

- Classify industries based on product descriptions of company disclosure text, the 10-K report.

- The cosine-similarities between the token counts of different product offerings are computed.

- Industries are defined by clustering firms according to their cosine similarities.

Text-based network industries and endogenous product differentiation

G Hoberg, G Phillips - Journal of Political Economy, 2016 - journals.uchicago.edu

… **industries** as time-varying intransitive networks. We name these new **industries text-based network industry** … Relative to existing **industry** classifications, our **text-based** classifications …

☆ Save 99 Cite Cited by 1884 Related articles All 20 versions Web of Science: 597 »

32

32