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

Matthew Gentzkow and Jesse M. Shapiro, 2010

SOCI 40133 1 Computational Content Analysis

Presented by Luxin Tian

luxintian@uchicago.edu

January 30, 2020

Overview

1. A Glimpse of the Research Design
   - Research question
   - Research Design
   - Data and Measurement

2. Measurement based on Computational Text Analysis
   - Methodology
   - Selecting Phrases for Analysis
   - Mapping Phrases to Ideology

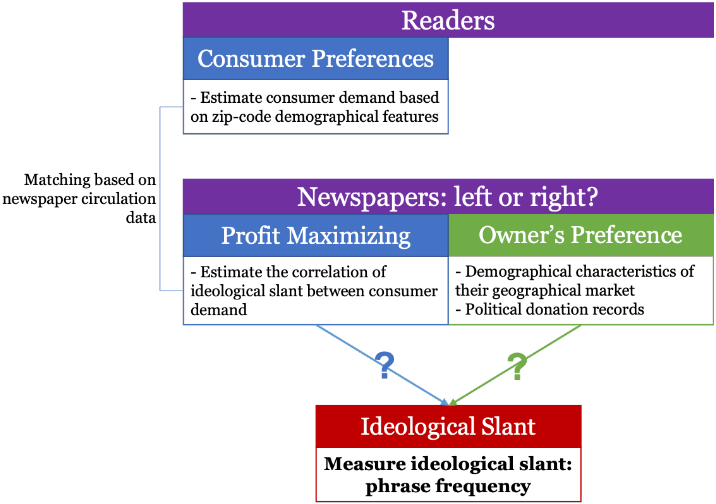
3. Conclusions

Research Question

Research question:
What are the incentives that shape ideological content on media?

- Why is it an important question?
  - “”’. . . news content has a powerful impact on politics, with ideologically diverse content producing socially desirable outcomes.”
  - “. . . unregulated markets will tend to produce too little ideological diversity.”
- This paper presents evidence on the incentives that shape ideological content and on the role of media ownership.

Research Design



Readers

Consumer Preferences

- Estimate consumer demand based on zip-code demographical features

Matching based on newspaper circulation data

Newspapers: left or right?

Profit Maximizing

- Estimate the correlation of ideological slant between consumer demand

Owner's Preference

- Demographical characteristics of their geographical market
- Political donation records

Ideological Slant

Measure ideological slant: phrase frequency

Data and Measurement

Measuring ideological slant: labelled data for supervised learning

- Congressional Record (text)
  - automated speech scripts
- Congressperson Party Identity (categorical label)
  - vote share for George W. Bush (Republican) in one's congressional district or state
  - robustness-check using the commonly-used roll-call measures

Measuring ideological slant: unlabelled data for classification

- Newspaper text
  - a total sample of 433 newspapers from ProQuest and Newslibrary

Data and Measurement

Measuring ownership and market chracteristics

- 2001 Editor and Publisher International Yearbook CD-ROM: owner's identity
- 2000 US Census: ideological identity of the regions where each newspaper's headquarters locate
  - demographic data; vote share for Bush; IV - religion
- political donation records

Measuring consumer chracteristics

- Newspaper GeoCirc: newspapers circulation data, matching with corresponding markets
- similarly, zip-code level demographic and ideological data

Measuring Slant

Approach: comparing phrase frequencies

1. Data for supervised learning:
   - Text: count the frequencies of certain phrases in the congress speech scripts
   - Categorical label: congressmen' party identities
2. Data that needs classification:
   - count the frequencies of the selected subset of phrases in the newspaper's language
3. "...compare phrase frequencies in the newspaper with phrase frequencies in the 2005 Congressional Record to identify whether the newspaper's language is more similar to that of a congressional Republican or a congressional Democrat."

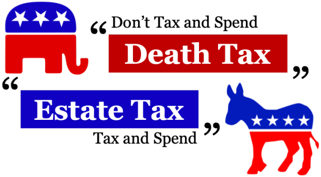
Measuring Slant

Approach: comparing phrase frequencies

1. Data for supervised learning:
   - Text: count the frequencies of certain phrases in the congress speech scripts
   - Categorical label: congressmen' party identities
2. Data that needs classification:
   - count the frequencies of the selected subset of phrases in the newspaper's language
3. "...compare phrase frequencies in the newspaper with phrase frequencies in the 2005 Congressional Record to identify whether the newspaper's language is more similar to that of a congressional Republican or a congressional Democrat."

Measuring Slant: Why is this meaningful?

There can be significant difference between the two party regarding the choice of phrases to use.



“Republicans [. . .] had to work hard to get members to act in unison, including training members to say ’death tax’. Estate tax sounds like it only hits the wealthy but ’death tax’ sounds like it hits everyone.” (Graetz and Shapiro, 2005)

Build A Criteria to Select Phrases for Analysis

- Count phrases

Definition (Phrase frequency)

f_{pld}, f_{plr}: the total number of times that phrase p of length l (two or three words) is used by Democrats and Republicans.
$f_{\sim pld}$, $f_{\sim plr}$: the total occurrences of length-l phrases that are not phrase p spoken by Democrats and Republicans.

Build A Criteria to Select Phrases for Analysis

- Build a χ^2-test statistic to select the phrases that the propensity of using them is equal for Democrats and Republicans

Definition (χ^2-test statistic)

If the counts f_{pld} and f_{plr} are drawn from (possibly different) multinomial distributions, χ^2_{pl} is a test statistic for the null hypothesis that the propensity to use phrase p of length l is equal for Democrats and Republicans.

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

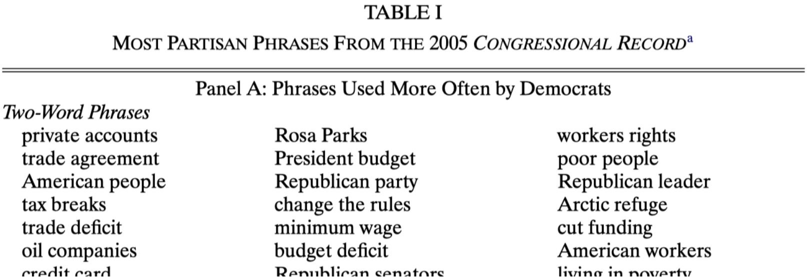


TABLE I

MOST PARTISAN PHRASES FROM THE 2005 *CONGRESSIONAL RECORD*[a]

Panel A: Phrases Used More Often by Democrats

Two-Word Phrases

private accounts	Rosa Parks	workers rights
trade agreement	President budget	poor people
American people	Republican party	Republican leader
tax breaks	change the rules	Arctic refuge
trade deficit	minimum wage	cut funding
oil companies	budget deficit	American workers
credit card	Republican senators	living in poverty

- Using this criteria, the authors restrict attention to those two-word and three-word phrases that are frequently used but not common English words by setting proper cutoffs.
- Finally, they select the 500 phrases of each length l with the greatest values of χ^2_{pl}, for a total of 1000 phrases.

Regression Models: Mapping Phrases to Ideology

The general idea is,

1. Train a regression model by using the labelled congress speech text data to estimate the relationship between the use of a phrase p and the ideology of the speaker.
2. Use the trained model to infer the ideology of newspapers by asking whether a newspaper tends to use phrases favored by more Republicans.

Regression Models: Mapping Phrases to Ideology

The regression models are pretty simple:

1. For each phrase p, regress its frequency of usage $\tilde{f}_{pc}$ on ideology y_c of the congresspeople c for the samples of congresspeople, obtaining the parameters a_p and b_p.

$$\tilde{f}_{pc} = a_p + b_p y_c + \epsilon_p$$

2. For each newspaper n, regress the same model for the sample of phrases but taking the estimated parameters a_p and b_p as given, obtaining the estimated y_c, which is, exactly, the vector of the predicted ideological identity.

$$(\tilde{f}_{pn} - a_p) = y_n b_p + e_{pn}$$

Regression Models: Mapping Phrases to Ideology

The regression models are pretty simple:

1. For each phrase p, regress its frequency of usage $\tilde{f}_{pc}$ on ideology y_c of the congresspeople c for the samples of congresspeople, obtaining the parameters a_p and b_p.

$$\tilde{f}_{pc} = a_p + b_p y_c + \epsilon_p$$

2. For each newspaper n, regress the same model for the sample of phrases but taking the estimated parameters a_p and b_p as given, obtaining the estimated y_c, which is, exactly, the vector of the predicted ideological identity.

$$(\tilde{f}_{pn} - a_p) = y_n b_p + e_{pn}$$

Validating the prediction

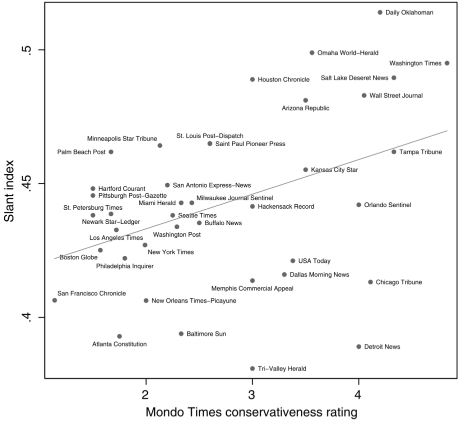
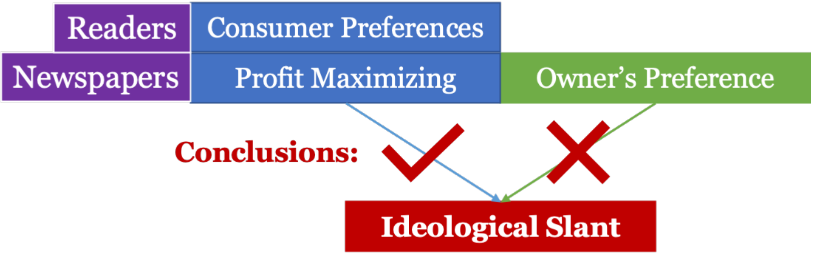


FIGURE 1.—Language-based and reader-submitted ratings of slant. The slant index (y axis) is shown against the average Mondo Times user rating of newspaper conservativeness (x axis), which ranges from 1 (liberal) to 5 (conservative). Included are all papers rated by at least two users on Mondo Times, with at least 25,000 mentions of our 1000 phrases in 2005. The line is predicted slant from an OLS regression of slant on Mondo Times rating. The correlation coefficient is 0.40 ($p = 0.0114$).

Validating the prediction using some labelled data from other sources (incomplete data):

- It performs well :)

Conclusions



- Readers have a significant preferences for like-minded news.
- Firms respond strongly to consumer preferences.
- By contrast, the identity of a newspaper's owner explains far less of the variation in slant.

Takeaways

- Counting words and phrases can help investigate rich characteristics underlying the text.
- It is important to select related words or phrases for analysis and drop common words that may be noisy or are irrelevant to the research.
    - Choose appropriate statistics that is testable (χ^2-test is used in this paper.) However, it may require substantial foundation in statistics.
    - Criteria for subsetting? Arbitrary cutoffs can be useful, but robustness-check is encouraged.
- Content analysis can be a power tool for complementing social science research. Here substantial methods in Economics and Econometrics were used.

Takeaways

- Counting words and phrases can help investigate rich characteristics underlying the text.
- It is important to select related words or phrases for analysis and drop common words that may be noisy or are irrelevant to the research.
  - Choose appropriate statistics that is testable (χ^2-test is used in this paper.) However, it may require substantial foundation in statistics.
  - Criteria for subsetting? Arbitrary cutoffs can be useful, but robustness-check is encouraged.
- Content analysis can be a power tool for complementing social science research. Here substantial methods in Economics and Econometrics were used.

DEAR EU, IT'S TIME TO GO
The Sun
DOVER & OUT
The Daily Telegraph
Unite behind Brexit, says May
the guardian
Today Britain steps into the unknown
After our Legs-it storm SARAH VINE tells her critics: Oh, DO get a life
Daily Mail
CENSORED BY THE LEFT
FREEDOM!
THE TIMES
The eyes of history are watching
May triggers Britain's EU departure
THE KEY QUESTIONS FACING BRITAIN AND THE EU
Brexit begins
DAILY EXPRESS
OUTRAGEOUS! WARTIME LAND
Getty Images

Image source: The Economists. Original author is unknown.

Gentzkow, Matthew, and Jesse M. Shapiro. "What drives media slant? Evidence from US daily newspapers." Econometrica 78, no. 1 (2010): 35-71.