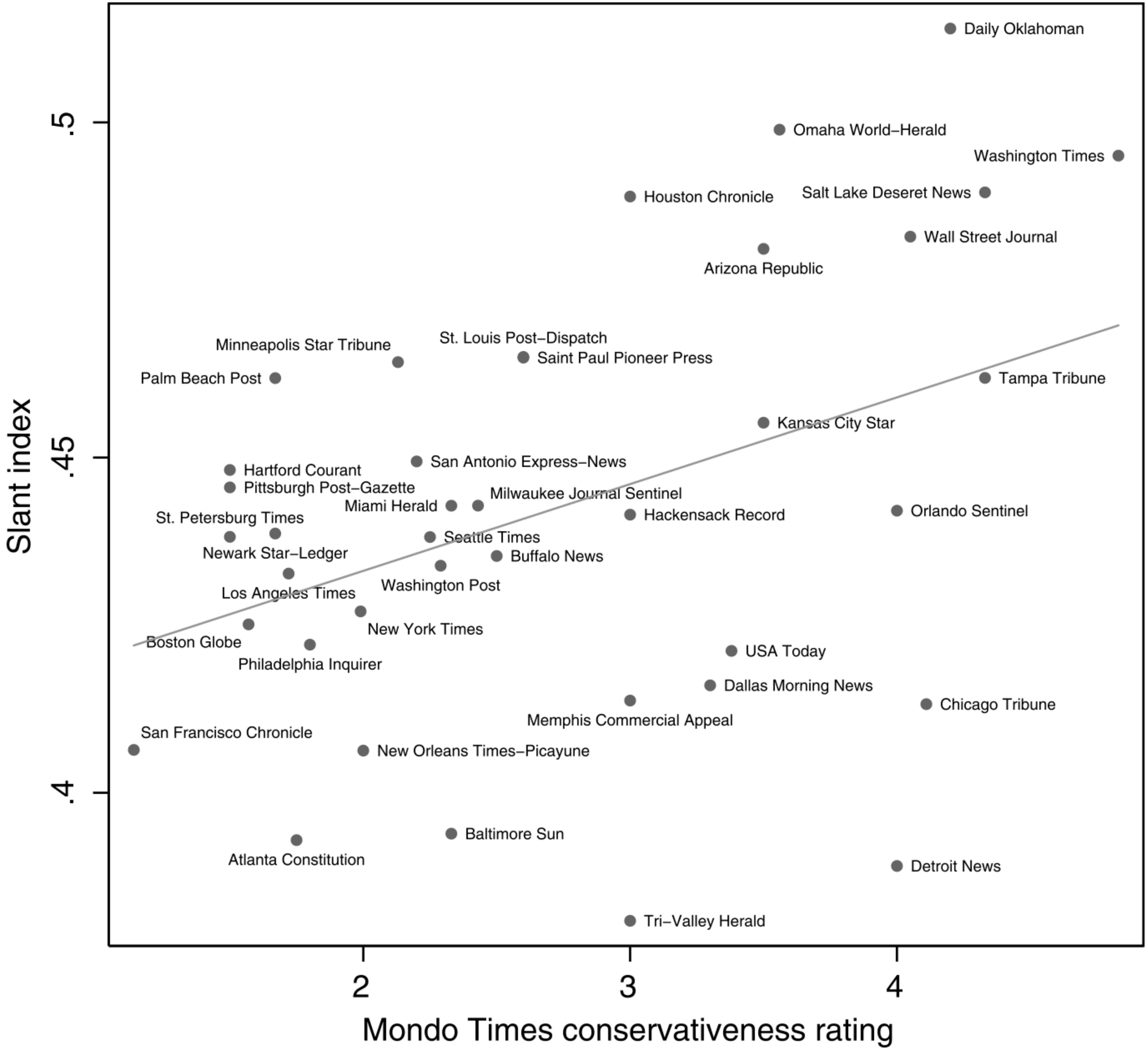
Measuring bias in Indian newspapers

Arun Varghese

Inspiration

- Gentzkow & Shapiro 2010
- First economics paper to use text data in an innovative way
- ‘What drives media slant?’: supply side or demand side?
- Intermediate step: need to quantify ‘slant’ or bias
- Done by comparing text from congressional record (ideology labeled using constituency vote share) against newspaper text

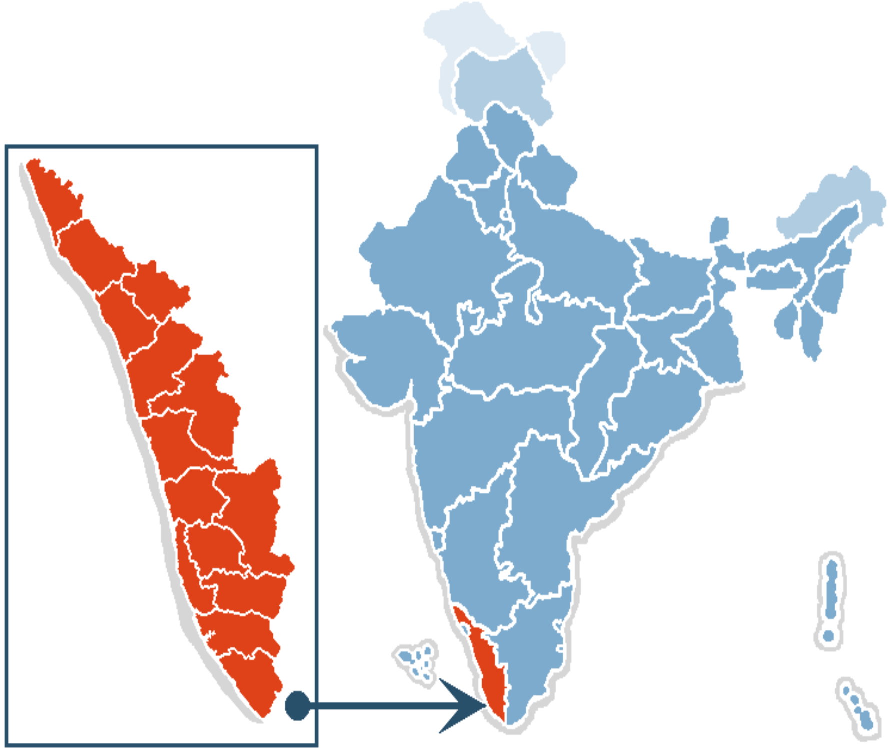

Daily Oklahoman
Omaha World–Herald
Washington Times
Houston Chronicle
Salt Lake Deseret News
Wall Street Journal
Arizona Republic
Minneapolis Star Tribune
St. Louis Post–Dispatch
Saint Paul Pioneer Press
Palm Beach Post
Tampa Tribune
Kansas City Star
San Antonio Express–News
Hartford Courant
Pittsburgh Post–Gazette
Milwaukee Journal Sentinel
Miami Herald
St. Petersburg Times
Hackensack Record
Orlando Sentinel
Seattle Times
Newark Star–Ledger
Buffalo News
Washington Post
Los Angeles Times
New York Times
Boston Globe
Philadelphia Inquirer
USA Today
Dallas Morning News
Chicago Tribune
Memphis Commercial Appeal
San Francisco Chronicle
New Orleans Times–Picayune
Baltimore Sun
Atlanta Constitution
Detroit News
Tri–Valley Herald
.5
.45
.4
Slant index
2
3
4
Mondo Times conservativeness rating

My project

- Similar exercise with Malayalam-language newspapers in the Indian state of Kerala
- Want to get better at language, learn something about politics
- Challenges: dealing with non-standard text, not many pre-existing NLP resources, agglutinative language

Kerala

- South-western tip of India
- First state to democratically elect communist government
- Now, multiple parties but dominated by two coalitions: Left Democratic Front (LDF) and congress-led United Democratic Front (UDF) who have traded majority power in legislature since Kerala founding



More background

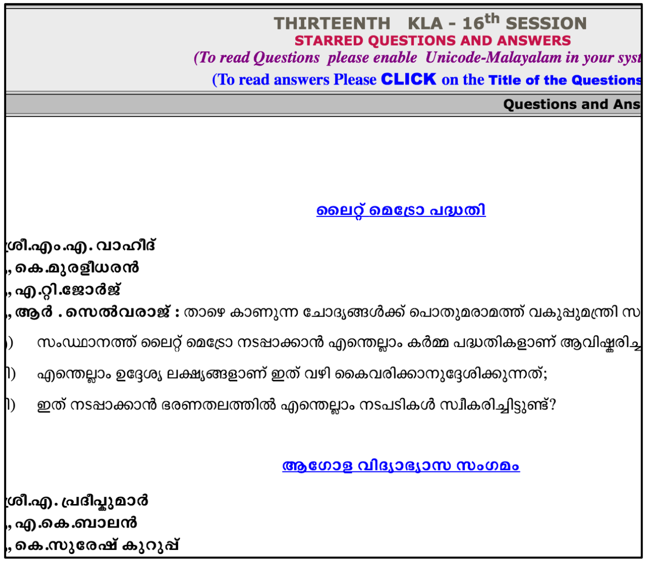
- Legislative Assembly: Niyamasabha (140 elected members)
- Last election in 2011
- Print media widely consumed
- Bias is widely thought to exist – some papers are owned by certain parties

Method

- Get text from legislative hearings
- Label by political slant/ideology from vote share data
- Generate text features and build regression model of slant
- Get newspaper article text and fit to model to get slant score for each newspaper

Data

- Text of legislative meetings available online
  - But mostly as PDFs
  - But ‘Question and Answer’ sessions available on html pages (bad sample?)
  - Scraped a month’s worth of these (labeled with speaker/s)
- Vote share data from 2011 election available online
- Text from newspaper articles available online



THIRTEENTH KLA - 16th SESSION

STARRED QUESTIONS AND ANSWERS

(To read Questions please enable Unicode-Malayalam in your syst

(To read answers Please CLICK on the Title of the Questions

Questions and Ans

ലൈറ്റ് മെട്രോ പദ്ധതി

ശ്രീ.എം.എ. വാഹിദ്
,, കെ.മുരളീധരന്‍
,, എ.റ്റി.ജോര്‍ജ്
,, ആര്‍ . സെല്‍വരാജ് : താഴെ കാണുന്ന ചോദ്യങ്ങള്‍ക്ക് പൊതുമരാമത്ത് വകുപ്പുമന്ത്രി സ

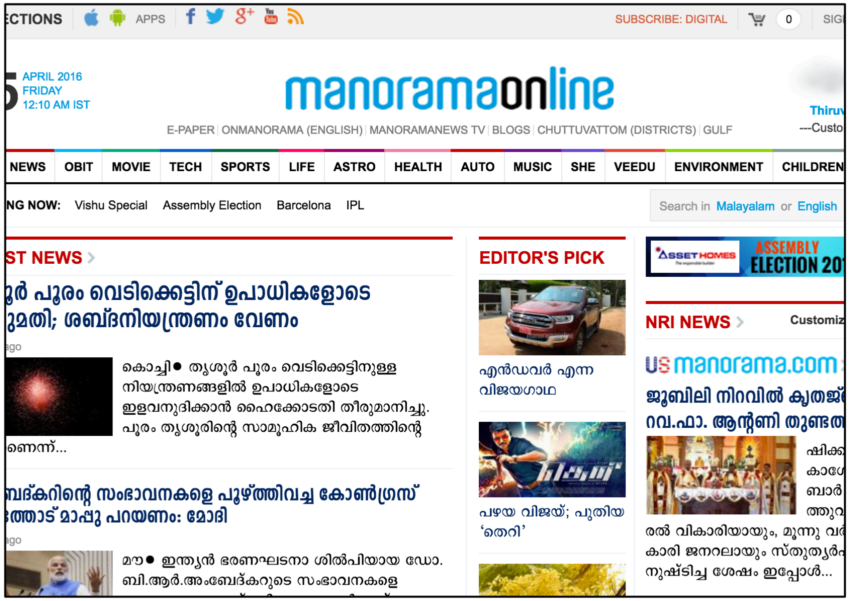
) സംസ്ഥാനത്ത് ലൈറ്റ് മെട്രോ നടപ്പാക്കാന്‍ എന്തെല്ലാം കര്‍മ്മ പദ്ധതികളാണ് ആവിഷ്ക്കരിച്ച

l) എന്തെല്ലാം ഉദ്ദേശ്യ ലക്ഷ്യങ്ങളാണ് ഇത് വഴി കൈവരിക്കാനുദ്ദേശിക്കുന്നത്;

l) ഇത് നടപ്പാക്കാന്‍ ഭരണതലത്തില്‍ എന്തെല്ലാം നടപടികള്‍ സ്വീകരിച്ചിട്ടുണ്ട്?

ആഗോള വിദ്യാഭ്യാസ സംഗമം

ശ്രീ.എ. പ്രദീപ്കുമാര്‍
,, എ.കെ.ബാലന്‍
,, കെ.സുരേഷ് കുറുപ്പ്



ൂര്‍ പൂരം വെടിക്കെട്ടിന് ഉപാധികളോടെ ുമതി; ശബ്ദനിയന്ത്രണം വേണം

ago

കൊച്ചി● തൃശൂര്‍ പൂരം വെടിക്കെട്ടിനുള്ള നിയന്ത്രണങ്ങളില്‍ ഉപാധികളോടെ ഇളവനുദിക്കാന്‍ ഹൈക്കോടതി തീരുമാനിച്ചു. പൂരം തൃശൂരിന്റെ സാമൂഹിക ജീവിതത്തിന്റെ ണെന്ന്...

ബദ്കറിന്റെ സംഭാവനകളെ പൂഴ്ത്തിവച്ച കോണ്‍ഗ്രസ് ത്തോട് മാപ്പു പറയണം: മോദി

ago

മൗ● ഇന്ത്യന്‍ ഭരണഘടനാ ശില്‍പിയായ ഡോ. ബി.ആര്‍.അംബേദ്കറുടെ സംഭാവനകളെ

എന്‍ഡവര്‍ എന്ന വിജയഗാഥ

പഴയ വിജയ്; പുതിയ ‘തെറി’

ജൂബിലി നിറവില്‍ കൃതജ്ഞ റവ.ഫാ. ആന്റണി തുണ്ടത്

ഷിക്ക കാഗേ ബാര്‍ ത്തുറ രല്‍ വികാരിയായും, മൂന്നു വര്‍ കാരി ജനറലായും സ്തുത്യര്‍ നുഷ്ടിച്ച ശേഷം ഇപ്പോള്‍...

Feature engineering and selection

- Represent legislative text as bag of words
- Drop ‘stop words’, rare words (have to define own stop words)
- Each document has ideology label
- Use chi-squared test to find words/phrases most representative of a particular party

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- Intuition: If we assumed party and words used were randomly assigned, which co-occur more/less often than we would expect?

Gentzkow & Shapiro

Democrats:

- ‘Workers rights’
- ‘Poor people’
- ‘Estate tax’

Republicans:

- ‘Stem cell’
- ‘Saddam Hussein’
- ‘Death tax’

Gentzkow & Shapiro

Democrats:

- ‘Workers rights’
- ‘Poor people’
- ‘Estate tax’

Republicans:

- ‘Stem cell’
- ‘Saddam Hussein’
- ‘Death tax’

Me

Lots of procedural words, question words, etc

LDF:

- അഗ്രികുൽറ്റു (agriculture)
- ആദിവാസി (adivasi)
- More negative words [അവബോധ, അനധികൃത]

UDF:

- Many procedural/question words
- അഞ്ചുവര്‍ഷങ്ങളിലായി വകയിരുത്തി (For five years, we’ve been here)
- And variations

Ongoing work

- Tinkering with model – not getting great explanatory power in-sample
- Use article text to help generate features
- Fit article text and get slant of newspapers!

Thanks!