# Measuring Political Media Slant Using Text Data

Lucas Shen
Feb 2022

# Measuring political media slant with text data

▶ Media affects electoral outcomes (e.g., Adena et al. 2015; Enikolopov et al. 2011)

▶ **Context:** Singapore

▶ **Question:** Is *The Straits Times* (flagship print media) less *accurate when quoting parliamentary speeches of the opposition?* **Yes**

▶ **Prior work**
  • Measures of *coverage intensity* (e.g., Groseclose and Milyo 2005; Gentzkow and Shapiro 2010)

▶ **This paper:**
  • Develops a notion of coverage (quotation) accuracy
  • OLS $\nleftarrow$ ML $+$ NLP
  • Text as data to deal with competing explanations (methodological)

## Preview: Results

- ▶ Unit of analysis: Direct quotations of parliament speeches in news articles

- ▶ Dependent variable: Quotation accuracy

- ▶ Treatment variable: Opposition status of politicians

- ▶ Quotation accuracy for opposition is lower

# Context: Politics & Media in Singapore

- ▶ Parliamentary democracy

- ▶ Ruling party—dominant party since 1959

- ▶ Opposition party—broke through in recent years

- ▶ Since 2006–present: Opposition accounted for ∼5–10 of 100 seats

People's Action Party (Ruling):



Worker's Party (Opposition):

# Context: Politics & Media in Singapore

- *The Straits Times*

- Flagship + newspaper of record

- (Perceived) Ties to the Ruling party via media-related regulations:
  - Newspaper and Printing Presses Act (NPPA)
  - Internal Security Act, Official Secrets Act, Sedition Act, Protection from Online Falsehoods and Manipulation Act*

*Print:*
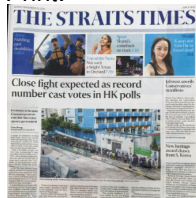


*Digital:*
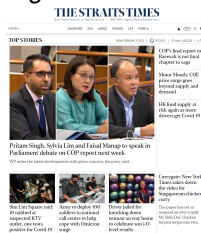
# Text as Data

▶ **Text data on (direct) quotes**:

- *The Straits Times*, 2005–16 from Factiva
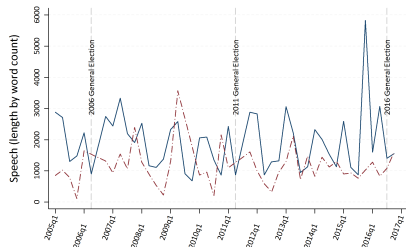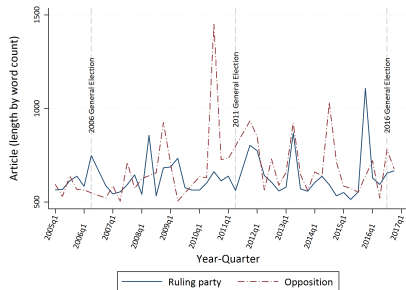
▶ **Text data on parliament speeches**:

- Transcripts from Official Online Repository (Hansard)

▶ Obtaining the data:

- Small scale: Hand-code news articles
- Large scale: Supervised ML

▶ Final panel (12 years from 2005–16):

- 3,425 news articles (from an initial 62,132)
- 5,130 parliament speeches



5

# Scoring quote accuracy example 1

**Parliament speech:** . . . . . . The Workfare Bonus marks a significant step because it provides low-wage workers with a cash bonus. If you work, you get the cash bonus at the end of the year, twice - this year based on 2005 work, next year based on 2006 work. It is a once-off scheme because this is the first time we are doing it. Payments will be for two years. We should **experiment and gain experience** with the scheme first, before considering whether we need a more permanent work-based assistance scheme like this, and if so, what form it should take. . . . . . .

**Quote from news article:**
The Government would 'experiment and gain experience' with the scheme first, he said.

**Substring accuracy**:
100/100

# Scoring quote accuracy example 2

**Parliament speech**: . . . . . . So, let us not take our harmonious social fabric for granted because **the alternative** **may** **be too painful to endure**. This is one pillar of success that we must continue to invest in, especially as we embark on our next lap of development. . . . . . .

**Quote from news article:**
He said Singaporeans should not take racial harmony for granted because 'the alternative will be too painful to endure'

**Substring accuracy**: 84/100

More examples

# Text as Data (1/3): Scoring quote accuracy

**Substring accuracy**: $\mathcal{F}(q, s) = \max_i \left\{ \underbrace{f(q, \overbrace{s^{(i:i+m-1)}}^{\substack{\text{speech} \\ \text{substring} \\ \text{of length } m}})}_{\substack{\text{edit distance b/w} \\ \text{quote \& substring}}} \;\middle|\; \underbrace{i \in \{1, \ldots, n-m+1\}}_{\substack{\text{all possible starting positions} \\ \text{for an m-length substring}}} \right\}$

- q = quote of length m
- s = speech
- i = substring start position
- $f$ is the base edit distance scorer (e.g. Levenshtein distance).
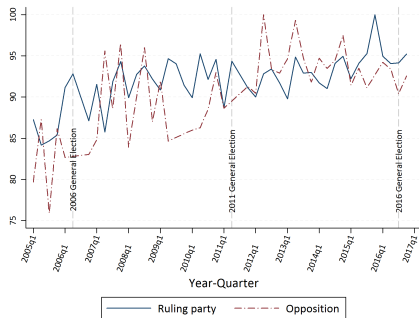
# Empirical approach: OLS

- ▶ Unit of analysis: Speech-quote pairs

- ▶ Dependent variable: Quotation accuracy

- ▶ Treatment variable: Opposition status of politicians

- ▶ Quotation accuracy:
  - Out of 100 points
  - Opposition politicians get quoted 1.5~2.4 points (1.6% to 2.5% less accurate)

- ▶ No difference in coverage intensity

Three alternative interpretations of OLS estimates.

1. **Political topics of speeches**
   Topic distributions of speeches & news articles (Unsupervised ML, LDA)

   $<$*cpf, retirement, minimum_sum, saving, cpf_saving*$>$

   $<$*police, home_team, officer, crime, inquiry*$>$

   $<$*premium, medishield_life, medishield, insurance, insurer*$>$

   $<$*student, school, learn, education, teacher*$>$

   $<$*fare, bus, public_transport, commuter, operator*$>$



Intertopic Distance Map (via multidimensional scaling)

10

## Text as Data (3/3)
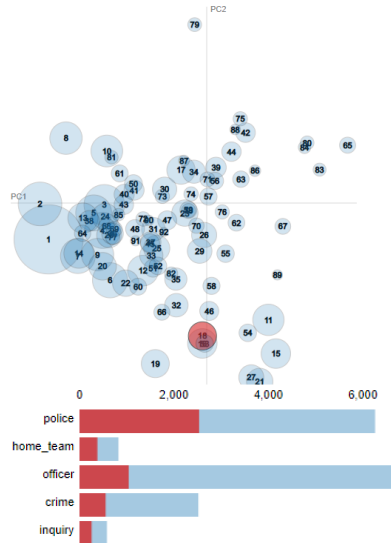
2. **Usage of trivial words (stopwords, e.g. "how", "you", "it", "a", "now")**
   - Alternative accuracy measures: Remove stopwords in pre-processing

3. **Government & opposition speeches are different**
   - Language grade/*readability*

   - Language sophistication/*lexical richness*
     (`https://github.com/lsys/lexicalrichness`, Shen 2022)

   - Speech objectivity & polarity

## Empirical specification

$$\text{quote accuracy}_{irst} = \underbrace{\sum_{k=2006}^{2016} \alpha_k \, year_{kt}}_{\text{year fixed effects}} + \underbrace{\sum_{\ell=11}^{13} \alpha_\ell \, parl_{\ell t}}_{\substack{\text{parliament} \\ \text{term fixed} \\ \text{effects}}} + \underbrace{\boxed{\beta \, opposition_i}}_{\substack{= 1 \text{ if politician} \\ i \text{ is from an} \\ \text{opposition party}}} + \gamma \boldsymbol{X}_{irst} + \varepsilon_{irst} \qquad (1)$$

▶ $i$ = politicians

▶ $r$ = article

▶ $s$ = speech

▶ $t$ = article-speech date

▶ $\boldsymbol{X}$ is the vector of controls

▶ Standard errors adjusted for clusters by news article

List of Controls

12

| | Log of quote length by word count | | Substring accuracy measure | | Bag-of-words accuracy measure | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Opposition | -0.138*** | -0.145*** | -1.455** | -1.485** | -2.434*** | -2.271*** |
| | (0.047) | (0.055) | (0.701) | (0.753) | (0.707) | (0.675) |
| ***Controls*** | | | | | | |
| Time fixed-effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Individual controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Article controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Topic controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Ministerial controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Electoral controls | — | Yes | — | Yes | — | Yes |
| ***F-statistics*** | | | | | | |
| $F$-stat, time fixed-effects | 3.090*** | 2.050** | 5.229*** | 3.189*** | 5.125*** | 2.501*** |
| $F$-stat, individual controls | 1.104 | 1.379 | 2.027** | 0.902 | 2.055** | 0.742 |
| $F$-stat, topic controls | 2.412*** | 2.290*** | 2.106*** | 2.078*** | 1.608*** | 1.459*** |
| $F$-stat, ministerial controls | 3.901*** | 3.997*** | 3.904*** | 3.964*** | 2.865*** | 2.180*** |
| $F$-stat, electoral controls | | 1.096 | | 0.990 | | 1.264 |
| $R^2$ | 0.056 | 0.068 | 0.171 | 0.197 | 0.113 | 0.124 |
| Observations | 14,887 | 10,900 | 14,887 | 10,900 | 14,887 | 10,900 |

Robust standard errors adjusted for clusters by news article in parentheses. *** Significant at the 1 per cent level. ** Significant at the 5 per cent level. * Significant at the 10 per cent level.

## Results:

► Opposition quote is 1.5 points less accurate (relative to ruling party politicians)

► Alternative measure (via BoW) leads to similar conclusions
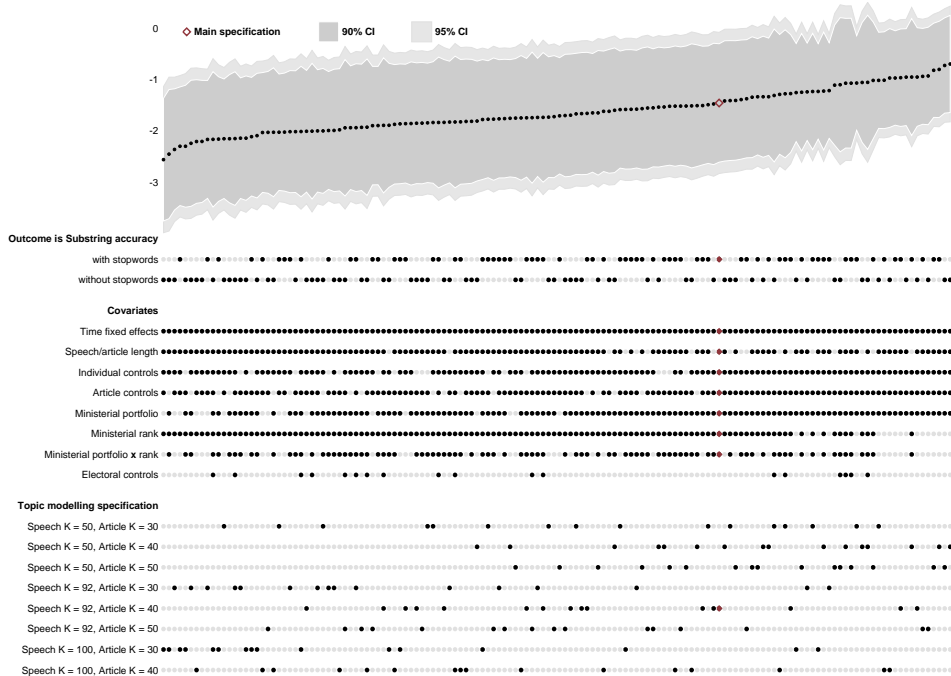
► No difference in coverage intensity

Article-speech level results

13

▶ **Bounding arguments**:
- Selection on observables and unobservables (Altonji et al. 2005; Oster 2017)
  - Data suggest degree of selection on unobservables relative to observables must be very high (58 or 16 times more) for differences to be zero ($\beta = 0$)
- Even if opposition/party status is random, a compositional bias (Angrist and Pischke 2009) suggests that OLS estimates are lower bounds in magnitude
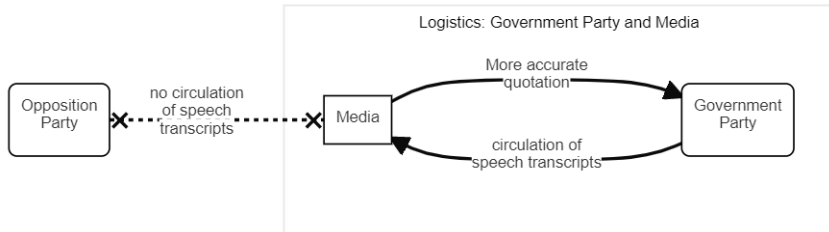
▶ **Robustness:**
- Author & beat FEs
- Dropping translated quotes
- Alternative clustering by speech and author
- Alternative topic distributions of speech and news articles
- Exclusion of stopwords (e.g. "a", "and", "how")
- Controls for language competency and speech tone
  - Language grade/*readability*
  - Language sophistication/*lexical richness*
  - Speech objectivity and polarity

# Mechanism: Media engagement strategy



Logistics: Government Party and Media

- More accurate quotation
- circulation of speech transcripts

Opposition Party — no circulation of speech transcripts — Media — Government Party

- ▶ Interview stakeholders:
  - Senior journalist from political desk
  - Member of parliament(s)

- ▶ Rational choice in a separating equilibrium:
  - Choice: Advance circulation of transcript
  - Cost: Spin
  - Benefit: Higher quote accuracy

- ▶ Separating equilibrium implies private beliefs about slant

**Logistics**

- ▶ Ruling party circulate speech transcripts in advance to the media
- ▶ Opposition do not
- ▶ Explains differences in accuracy

## Recap & Closing

▶ Context: Singapore where slant is subdued

▶ Text data reveals subtle political media slant

(E.g. Groseclose and Milyo 2005; Bernhardt et al. 2008; Ho and Quinn 2008; Sutter 2012)

▶ Consequences of quotation inaccuracy?

▶ **Methodological contributions**:
  - Using both supervised & unsupervised learning
    - Supervised ML expedites data parsing
    - Unsupervised learning annotates data free from bias
  - NLP methods: quantify quotation accuracy
    - Accuracy vs. intensity
    - Accuracy is also less subject to OVB

Thank you!

# References

Adena, M., R. Enikolopov, M. Petrova, V. Santarosa, and E. Zhuravskaya (2015). Radio and the Rise of The Nazis in Prewar Germany. *The Quarterly Journal of Economics 130*(4), 1885–1939.

Altonji, J., T. Elder, and C. Taber (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy 113*(1), 151–184.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton : Princeton University Press, c2009.

Bernhardt, D., S. Krasa, and M. Polborn (2008). Political Polarization and the Electoral Effects of Media Bias. *Journal of Public Economics 92*(5), 1092–1104.

Enikolopov, R., M. Petrova, and E. Zhuravskaya (2011). Media and Political Persuasion: Evidence from Russia. *American Economic Review 101*(7), 3253–3285.

Gentzkow, M. and J. M. Shapiro (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica 78*(1), 35–71.

Groseclose, T. and J. Milyo (2005). A Measure of Media Bias. *Quarterly Journal of Economics 120*(4), 1191–1237.

Ho, D. E. and K. M. Quinn (2008). Measuring Explicit Political Positions of Media. *Quarterly Journal of Political Science 3*(4), 353–377.

Oster, E. (2017). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics 37*(2), 1–18.

Shen, L. (2022). LexicalRichness: A small module to compute textual lexical richness.

Sutter, D. (2012). Is the Media Liberal? An Indirect Test Using News Magazine Circulation. *Applied Economics 44*(27), 3521–3532.

## Data: Scoring quote accuracy example 3

**Parliament speech**: Mr Speaker, Sir, in the course of this debate, I may have misheard certain things and if I misquoted the Members, let me apologise. Mr Low asked whether the GIC money is derived from CPF money. The relationship is not so simple. Let me give an example. You put money in the bank, and you agree that you put it there and you get 2%. The bank publishes a report and says that of all its earnings, it earned 8%. You go to the bank and say you want 8%, it does not work. MOF has **taken on our liabilities**. What MOF does with its money is MOF's consideration but . . . . . .

**Quote from news article**:
. . . . . . Dr Ng said that, in fact, the Ministry of Finance (MOF) has 'taken our liabilities', meaning that MOF bears the risk of losing the money if investments fail.

**Substring accuracy**: 86
**Bag-of-words accuracy**: 100

# Data: Scoring quote accuracy example 4

**Parliament speech**: . . . . . . In my view, this amendment is a **half-baked** Bill which seemingly can solve the employment issue before people are allowed to take out their CPF savings, yet in reality, it cannot guarantee that Singaporeans will continue to be employed until the age of 65, only by then can they take out their CPF savings. . . . . . .

**Quote from news article**: Speaking during the debate on the Retirement and Re-Employment Act yesterday, Mr Low criticised the new law as 'half-cooked'.

**Substring accuracy**: 73
**Bag-of-words accuracy**: 53

## Data: Quote Accuracy Measure 2

The 2nd accuracy measure (using **common words**) is:

$$\mathcal{F}_2(q, s) = \max\Big\{ f(\widetilde{q}, \widetilde{s}), f(\widetilde{q}, \widetilde{C}), f(\widetilde{s}, \widetilde{C}) \Big\},$$

where

- ▶ $\widetilde{C}$ is the sorted set of common words appearing in both quote and speech.
- ▶ $\widetilde{q}$ and $\widetilde{s}$ are the sorted strings of quote and speech.
- ▶ $f$ is the base edit distance scorer (e.g. Levenshtein distance).

**A. Individual controls**

| | |
|---|---|
| Age | Quadratic for age of politician |
| Tenure | Quadratic for political tenure of politician |
| Gender | $\in$ {Male, Female} |
| Race | $\in$ {Chinese, Indian, Malay, Eurasian/others} |

**B. Ministerial controls**

| | |
|---|---|
| Politician type | $\in$ {PM, DPM, Ministers, SPS, Parl Sec, Mayor, Speaker, SMS, MOS, MP, NCMP, NMP} |
| Ministry portfolio | $\in$ {Comms, CCY, Defence, Education, Finance, Foreign, Health, Home, Law, Manpower, National Development, EWR, Trade & Industry, Transport, PMO} |

**C. Electoral controls**

| | |
|---|---|
| Constituency size | $\in$ {1, 4, 5, 6} |
| Voters | Eligible voters in the constituency |
| Votes | Number of votes won in the constituency |
| Votes (%) | Percentage of votes won in the constituency |
| Winner's margin | Number of winner's vote - number of loser's vote |
| Winner's margin (%) | Ratio of winner's margin to voters |

**D. Topic controls** (from *Latent Dirichlet Allocation* (LDA))

| | |
|---|---|
| Speech | vector of $K = 92$ probabilisitc topic association |
| Quote | vector of $K = 92$ probabilisitc topic associations |
| Newspaper Article | vector of $K = 40$ probabilistic topic associations |

**E. Article controls**

| | |
|---|---|
| Day of week | $\in$ {Mon, Tue, Wed, Thu, Fri, Sat, Sun} |
| Section | $\in$ {*Singapore, Prime News, Top of the News, Home, ST, Insight, News, Money, Think, Review- Insight, Sports, Opinion, World, Others*} |
| Translation | $= 1$ if speech $s$ was translated from a vernacular to English |

|  | Log of quote length by word count | | Count of quote fragments | | Bag-of-words accuracy measure | |
|---|---|---|---|---|---|---|
|  | (1) | (2) | (3) | (4) | (5) | (6) |
| Opposition | 0.052 | 0.038 | 0.328*** | 0.318*** | -2.499*** | -2.554*** |
|  | (0.064) | (0.074) | (0.089) | (0.109) | (0.785) | (0.832) |
| ***Controls*** | | | | | | |
| Time fixed-effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Individual controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Article controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Topic controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Ministerial controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Electoral controls | — | Yes | — | Yes | — | Yes |
| ***F-statistics*** | | | | | | |
| $F$-stat, time fixed-effects | 2.682*** | 1.613* | 1.753** | 2.127** | 4.589*** | 2.033** |
| $F$-stat, individual controls | 0.397 | 1.153 | 1.322 | 2.775*** | 2.490** | 1.855* |
| $F$-stat, article controls | 3.290*** | 2.538*** | 5.370*** | 4.487*** | 3.189*** | 2.248*** |
| $F$-stat, topic controls | 2.762*** | 2.413*** | 2.314*** | 2.164*** | 1.563*** | 1.440*** |
| $F$-stat, ministerial controls | 2.694*** | 2.722*** | 5.106*** | 4.987*** | 2.360*** | 1.826*** |
| $F$-stat, electoral controls | | 0.382 | | 1.660 | | 1.476 |
| $R^2$ | 0.181 | 0.199 | 0.241 | 0.256 | 0.118 | 0.139 |
| Observations | 7,087 | 5,143 | 7,087 | 5,143 | 7,087 | 5,143 |

**Article-speech level results:**

► Unit of analysis: All (combined) quotes in an article-speech pair

► No difference in coverage intensity

| Quote fragment | Originating speech | Quote Accuracy |
|---|---|---|
| a fundamental relook | …is taking **a more fundamental relook** at this regulation framework and see how best we can support this strategy … | 95 |
| there will be a need for us to make sure we have regular fare increases of the right quantum | …So **there will be a need for us to make sure that we have regular fare increases of the right quantum**… | 95 |
| likely to go back down the slippery slope | …considering that some of them may be drop-outs or expelled from school, they are **likely to go back down the slippery road** … | 90 |
| core Singapore values | …Rather it is an acknowledgement that the **core Singaporean values** of multi-racialism and meritocracy can and should co-exist with each other… | 90 |
| resilience in response to ground reaction | …commended the PAP's **resilience in response to the ground's reaction** after the election, and I said this augurs well for Singapore… | 85 |
| too employer-focused | …and increase in absentee payroll are **rather employer-focused**… | 85 |
| That is the purpose of these amendments | …That is in fact **one of the purposes of this amendment** which we are bringing to Parliament… | 80 |
| deviates from the concept of free market | …this Bill **deviates from the idea of the concept of a free market**, where supply of services by companies is set by market demand … | 80 |
| the high end | …we noted that the rates were at **a higher end**, but we had the rates that were charged … | 75 |
| They (said) there are two purposes | …They gave the reasons that they wanted the two budget hotels **to serve two purposes**… | 75 |
| look them in the eye | …I want to be able to **look these men and others in the eyes** and say to them … | 70 |
| to help ensure all Singaporeans can afford their education in a public university | …To **ensure that all Singaporeans can afford to attend Government-funded universities** … | 70 |
| temporary or permanent solutions for this important issue | …to propose certain solutions, whether **temporary or permanent, to help resolve this, to me, a very important issue**… | 65 |
| completely unwarranted, alarmist, and show fundamental lack of understanding about the law | …I would venture to suggest that such statements are **alarmist and reveal a fundamental misunderstanding as to what this Bill and the law** is all about … | 65 |

*Notes*—The speech column shows the portion of the speech which contains the best partial substring match. The text **in bold** is an approximation of the best partial substring match. Quote accuracy is as defined by the substring quote accuracy measure.