Quantitative Text Analysis: Applications to Political Science

Kenneth Benoit
London School of Economics and Political Science

ELECDEEM 21 March 2011
The information in text

Huge potential

A way to observe UNOBSERVABLES

Largely untapped
How to use Text?

Read and interpret
How to use Text?

Read and interpret
To be manageable, useful:

Text must be reduced to summary, quantitative information
Text as data

Inherit properties of statistics

Precise characterizations of uncertainty

Concern: Reliability

Concern: Validity
Past - Present - Future
Past - Present - Future

Hand-coded content analysis
Sometimes computer-assisted
Extensively applied to manifestos
Past - Present - Future

“Text as data” scaling approaches
Classification approaches
and still: manifesto hand-coding
Past - Present - Future

Better uncertainty models
Advances in estimation methods
Advances in parametric scaling
The Big Picture

A
Stochastic text generation process

B
Stochastic text coding process

Underlying position to be communicated by text author; fundamentally unobservable by others

Observed text used as indicator author’s position

Data characterizing unobservable underlying position of author in terms of coded observed text
PAST
The Comparative Manifesto Project

3,000+ party programmes
1948-2000
650+ parties
52 countries
three books
hundreds of articles use it
The CMP Coding Process

Human coders unitize the text

CMP text units are called quasi-sentences

Then each unit is assigned a category

Category percentages of the total text are used to measure policy
The CMP Coding Scheme

7 domains of policy
  External relations
  Freedom and Democracy
  Political System
  Economy
  Welfare and Quality of Life
  Fabric of Society
  Social Groups

56 policy categories
  (or “uncoded”)

Human coders are trained to classify text units into each category
**Agriculture**

Lobby for changes in the World Trade Organisation to protect domestic agriculture from being undercut by imports that are not subject to the same quality, health and environmental standards.

- Promote the clean green image of Ireland abroad creating a ‘green Ireland’ brand for food products and ensure Ireland becomes a GM-free zone and ban farming of cloned animals.
- Set a target for 5% of national acreage to be organically converted by 2012.

**501 Environmental Protection: Positive**

Score: $\frac{3}{4} \times 100 = 75\%$
Agriculture

Lobby for changes in the World Trade Organisation to protect domestic agriculture from being undercut by imports that are not subject to the same quality, health and environmental standards.

Promote the clean green image of Ireland abroad creating a ‘green Ireland’ brand for food products and ensure Ireland becomes a GM-free zone and ban farming of cloned animals.

Set a target for 5% of national acreage to be organically converted by 2012.

Or maybe...

Score: 3/6*100 = 50%
Problems with the CMP

Each text coded once, and only once, by a single human coder

Labor intensive, slow, and costly

Concerns exist with

- stochastic nature of text
- unreliability of human unitization
- unreliability of human coding
- scaling (percentage of all text)
Problems with the CMP

Each text coded once, and only once, by a single human coder

Labor intensive, slow, and costly

Concerns exist with

- stochastic nature of text
- unreliability of human unitization
- unreliability of human coding
- scaling (percentage of all text)

Benoit, Laver, and Mikhaylov 2009 (AJPS)
Simulates stochastic manifesto generation by bootstrapping sentences and reconstructing all CMP scores

Assume category counts are generated by a multinomial distribution

Very robust and flexible

No new data needed
Example: British Conservative party: The CMP reported “Rile” value is 25.7, but 95% confidence interval is [20.7, 31.4]
Problems with the CMP

Each text coded once, and only once, by a single human coder

Labor intensive, slow, and costly

Concerns exist with

- stochastic nature of text
- unreliability of human unitization
- unreliability of human coding
- scaling (percentage of all text)

Mikhaylov, Laver, and Benoit 2010
MLB 2010

Experiment: Re-code (one of) two manifestos
(UK Liberal/SDP Alliance 1983
(New Zealand National Party 1972)

Both texts were pre-unitized

Used measures of agreement and misclassification to assess reliability

Recruited experienced and trained coders

Discarded the bottom worst third of the results
MLB 2009 Findings

Coders significantly disagreed with one another
Poor Agreement with the “gold standard” master coding
MLB 2009 Findings

Significant misclassification in individual categories
Problems with the CMP

Each text coded once, and only once, by a single human coder

Labor intensive, slow, and costly

Concerns exist with

- stochastic nature of text
- unreliability of human unitization
- unreliability of human coding
- scaling (percentage of all text)

Lowe, Benoit, Mikhaylov, and Laver 2010
CMP scales policy positions as \( \frac{R - L}{N} \):

**Problem**
- clusters low-frequency categories near zero
- sensitive to irrelevant content

**Alternative** is to use relative \( \frac{R - L}{R + L} \)
- clusters values around extremes
- super-sensitive in the middle

**Solution:** Use logit scale
- \( \log\frac{R + .5}{L + .5} \)
  (and make all categories confrontational)
Comparing scales

Protectionism distributions

Density plot of Saliency Score

Density plot of Relative Proportional Difference Score

Density plot of Logit Score

LBML 2010
(protectionism)
LBML 2010
Comparison with independent expert survey score
(social morality)
Comparing scales
Environmental distributions

\[ L_{\text{Env}} = N_{401} + (\text{Env' Protection: +}) \]
\[ N_{416} \ (\text{Anti-Growth: +}) \]

\[ L_{\text{Env}} = N_{410} \ (\text{Productivity: Positive}) \]
LBML 2010

Full dataset is available

Comes with bootstrapped uncertainty measures from BLM 2009

Contains 21 new left-right scales

Half never before used!
Problems with the CMP

Each text coded once, and only once, by a single human coder

Labor intensive, slow, and costly

Concerns exist with

- stochastic nature of text
- unreliability of human unitization
- unreliability of human coding
- scaling (percentage of all text)
Unitization variance problems

Analysis of coder training data conducted by Andrea Volkens indicate a variance in unitizing texts into “quasi-sentences” of approximately +/- 10%

This introduces additional (non-systematic) noise into the data, but ...

because of the unique nature of the CMP data (where each category is a percentage of all text units), differences in length will bias category estimates

No known solution to this problem except to move to unambiguous text units, such as natural sentences
A test:
Count the Quasi-Sentences

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. The fight against increases in the cost of living is the most important single issue in economic management. People without jobs represent waste of productive effort: National supports a policy of full employment and the dignity of labour. We do not accept unemployment as a balancing factor in economic management. Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.
How many of you said... SEVEN?

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. The fight against increases in the cost of living is the most important single issue in economic management. People without jobs represent waste of productive effort: National supports a policy of full employment and the dignity of labour. We do not accept unemployment as a balancing factor in economic management. Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.
Unitization variance from CMP Training Document
(VERY) Preliminary Results from Recoding Study
Example of what we are recoding:
Sinn Fein 2001 Manifesto

It has delivered the Good Friday Agreement and offered us a route map out of conflict and into a new Ireland based on equality and justice. The Good Friday Agreement is an all-Ireland agreement. It transcends partition and it offers a new opportunity for people across the island.

This potential, this opportunity, has still to be delivered. Yet despite all of the obstacles, Sinn Féin is committed and is determined to see the potential of the Good Friday Agreement realised.

Throughout this entire process, Sinn Féin has been the engine for change. We have given real leadership.

We have been both flexible and imaginative but all the time wedded to our belief that the changes which are clearly necessary and indeed promised under the Agreement must be delivered.

Our peace strategy and the Peace Process which it delivered can create change. With increased political strength this process and the changes it will deliver will become irreversible.

All-Ireland Politics

Sinn Féin is the only all-Ireland party. The only party with a strategy and policies for achieving Irish unity and independence.

The Good Friday Agreement is an all-Ireland agreement.

Over the last 12 months — through the new political institutions created under the Good Friday Agreement — the beginning of all-Ireland cooperation in the crucial fields of education, health,
CMP problems are structural

Too many categories!

Need a more stable text unit

Need a fully confrontational scale

Resists change

Advantage: Good description of actual content
THE PRESENT

“Text as Data” Approaches
"From Text to Policy Positions"

$\theta \xrightarrow{\text{Stochastic Process}} \text{Text} \xrightarrow{\text{Estimation Inference}} \hat{\theta}$

also: From Policy Positions to text
Problem 1: Interpret this data!

<table>
<thead>
<tr>
<th>party</th>
<th>spend_~l</th>
<th>votes1st</th>
<th>electo~e</th>
<th>gender</th>
<th>margin~d</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>ind</td>
<td>6544.23</td>
<td>335</td>
<td>95060</td>
<td>m</td>
<td>Safe</td>
<td>5</td>
</tr>
<tr>
<td>ind</td>
<td>14558.26</td>
<td>1614</td>
<td>95060</td>
<td>m</td>
<td>Safe</td>
<td>5</td>
</tr>
<tr>
<td>fg</td>
<td>19153.71</td>
<td>5468</td>
<td>95060</td>
<td>m</td>
<td>Winnable</td>
<td>5</td>
</tr>
<tr>
<td>lab</td>
<td>10658.21</td>
<td>4272</td>
<td>95060</td>
<td>m</td>
<td>Winnable</td>
<td>5</td>
</tr>
<tr>
<td>ff</td>
<td>19648.3</td>
<td>9343</td>
<td>95060</td>
<td>m</td>
<td>Safe</td>
<td>5</td>
</tr>
<tr>
<td>ff</td>
<td>16968.18</td>
<td>12489</td>
<td>95060</td>
<td>m</td>
<td>Winnable</td>
<td>5</td>
</tr>
<tr>
<td>ff</td>
<td>24100.27</td>
<td>8711</td>
<td>95060</td>
<td>m</td>
<td>Unlikely</td>
<td>5</td>
</tr>
<tr>
<td>gp</td>
<td>12110.11</td>
<td>4961</td>
<td>95060</td>
<td>f</td>
<td>Unlikely</td>
<td>5</td>
</tr>
<tr>
<td>lab</td>
<td>8404.43</td>
<td>3732</td>
<td>95060</td>
<td>m</td>
<td>Unlikely</td>
<td>5</td>
</tr>
<tr>
<td>fg</td>
<td>19743.1</td>
<td>7841</td>
<td>95060</td>
<td>m</td>
<td>Unlikely</td>
<td>5</td>
</tr>
<tr>
<td>lab</td>
<td>.</td>
<td>.</td>
<td>95060</td>
<td>m</td>
<td>Unlikely</td>
<td>5</td>
</tr>
<tr>
<td>sf</td>
<td>6633.45</td>
<td>2078</td>
<td>95060</td>
<td>m</td>
<td>Unlikely</td>
<td>5</td>
</tr>
<tr>
<td>fg</td>
<td>11217.01</td>
<td>4819</td>
<td>87087</td>
<td>m</td>
<td>Unlikely</td>
<td>5</td>
</tr>
<tr>
<td>ff</td>
<td>22383.53</td>
<td>10679</td>
<td>87087</td>
<td>m</td>
<td>Unlikely</td>
<td>5</td>
</tr>
<tr>
<td>sf</td>
<td>28953.32</td>
<td>10832</td>
<td>87087</td>
<td>m</td>
<td>Safe</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>party</td>
<td>spend_~l</td>
<td>votes1st</td>
<td>electo~e</td>
<td>gender</td>
<td>margin~d</td>
</tr>
<tr>
<td>---</td>
<td>-------</td>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td>16.</td>
<td>lab</td>
<td>8756.67</td>
<td>550</td>
<td>87087</td>
<td>m</td>
<td>Safe</td>
</tr>
<tr>
<td>17.</td>
<td>pd</td>
<td>30573.12</td>
<td>1131</td>
<td>87087</td>
<td>m</td>
<td>Safe</td>
</tr>
<tr>
<td>18.</td>
<td>ind</td>
<td>17196.73</td>
<td>1026</td>
<td>87087</td>
<td>m</td>
<td>Safe</td>
</tr>
<tr>
<td>19.</td>
<td>gp</td>
<td>10699.87</td>
<td>1100</td>
<td>87087</td>
<td>m</td>
<td>Unlikely</td>
</tr>
<tr>
<td>20.</td>
<td>fg</td>
<td>12839.2</td>
<td>4639</td>
<td>87087</td>
<td>m</td>
<td>Unlikely</td>
</tr>
<tr>
<td>21.</td>
<td>ind</td>
<td>17934.79</td>
<td>7722</td>
<td>87087</td>
<td>m</td>
<td>Unlikely</td>
</tr>
<tr>
<td>22.</td>
<td>ff</td>
<td>20122.19</td>
<td>3731</td>
<td>87087</td>
<td>m</td>
<td>Unlikely</td>
</tr>
<tr>
<td>23.</td>
<td>ff</td>
<td>21483.37</td>
<td>7204</td>
<td>87087</td>
<td>m</td>
<td>Unlikely</td>
</tr>
<tr>
<td>24.</td>
<td>fg</td>
<td>11124</td>
<td>6113</td>
<td>87087</td>
<td>m</td>
<td>Unlikely</td>
</tr>
<tr>
<td>25.</td>
<td>csp</td>
<td>3141.27</td>
<td>358</td>
<td>87087</td>
<td>m</td>
<td>Unlikely</td>
</tr>
<tr>
<td>26.</td>
<td>ind</td>
<td>34542.73</td>
<td>1943</td>
<td>87087</td>
<td>m</td>
<td>Unlikely</td>
</tr>
<tr>
<td>27.</td>
<td>ff</td>
<td>13120.48</td>
<td>6717</td>
<td>78643</td>
<td>m</td>
<td>Unlikely</td>
</tr>
<tr>
<td>28.</td>
<td>gp</td>
<td>7771.53</td>
<td>2903</td>
<td>78643</td>
<td>m</td>
<td>Safe</td>
</tr>
<tr>
<td>29.</td>
<td>csp</td>
<td>140</td>
<td>176</td>
<td>78643</td>
<td>m</td>
<td>Safe</td>
</tr>
<tr>
<td>30.</td>
<td>fg</td>
<td>14195.46</td>
<td>4015</td>
<td>78643</td>
<td>m</td>
<td>Safe</td>
</tr>
</tbody>
</table>
We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. The fight against increases in the cost of living is the most important single issue in economic management.

People without jobs represent waste of productive effort: National supports a policy of full employment and the dignity of labour. We do not accept unemployment as a balancing factor in economic management.

Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.
We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. The fight against increases in the cost of living is the most important single issue in economic management.

People without jobs represent waste of productive effort: National supports a policy of full employment and the dignity of labour. We do not accept unemployment as a balancing factor in economic management.

Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.
We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. The fight against increases in the cost of living is the most important single issue in economic management.

People without jobs represent waste of productive effort. National supports a policy of full employment and the dignity of labour. We do not accept unemployment as a balancing factor in economic management.

Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.
We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. The fight against increases in the cost of living is the most important single issue in economic management.

People without jobs represent waste of productive effort: National supports a policy of full employment and the dignity of labour. We do not accept unemployment as a balancing factor in economic management.

Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.

408 Economic Goals: Statements of intent to pursue any economic goals not covered by other categories.

Sotkamossa sijaitsevan Talvivaaran kaivoksen sivutuotteena tulee myös uraania, joka aiotaan ottaa jätelietteestä talteen. Tässä uraanin talteenotossa syntyy niin paljon ydinvoimalaitosten polttoainetta, että se riittäisi noin 80 prosenttisesti Suomessa toimivien ydinvoimalaisten tarpeisiin.

Talvivaaran tapauksessa ei kaivoksen johdon mukaan ole kysymys varsinaisen uraanikaivoksen avaamisesta, vaan vain sivutuotteen talteenotosta. Valtioneuvosto tullee päätämään Talvivaara-asiasta uraanin osalta tämän vuoden aikana.
and this one???
Problem (last): Interpret these text scaling results!

**Work-Horse/Show-Horse**

**Left-Right**

Figure 4: Rhetorical Ideal Points with Partisan Means and Cutline
Scaling models:
Wordscores; “Wordfish”; others

Classification methods
Naive Bayes (e.g. McIntosh et al)
Readme (Hopkins and King)
Dynamic Topic Models (e.g. Quinn et al.)
Support Vector Machines (Hillard et al, Yu et al.)
Expressed Agenda Model (Grimmer 2010)
Wordscores
(Laver, Benoit and Garry 2003)

“Reference” texts: texts about which we know something (a scalar dimensional score)

“Virgin” texts: texts about which we know nothing (but whose dimensional score we’d like to know)

Basic procedure:
Analyze reference texts to obtain word scores
Use word scores to score virgin texts
Wordscores

The Wordscore Procedure
(Using the British manifestos 1992-1997 as an illustration)

1. Labour 1992 5.35
2. Liberals 1992 8.21
3. Cons. 1992 17.21
4. Labour 1997 9.17 (.33)
   Liberals 1997 5.00 (.36)
   Cons. 1997 17.18 (.32)

Reference Texts

Scored word list

- drugs 15.66
- corporation 15.66
- inheritance 15.48
- successfully 15.26
- markets 15.12
- motorway 14.96
- nation 12.44
- single 12.36
- pensionable 11.59
- management 11.56
- monetary 10.84
- secure 10.44
- minorities 9.95
- women 8.65
- cooperation 8.64
- transform 7.44
- representation 7.42
- poverty 6.87
- waste 6.83
- unemployment 6.76
- contributions 6.68

Scored virgin texts
“Wordfish” scaling model
(Slapin and Proksch 2008; Monroe and Maeda 2004)

Assumptions about $P(W_1 \ldots W_V | \theta)$

$$\log E(W_i | \theta_j) = \alpha_j + \psi_i + \beta_i \theta_j$$

$\alpha_j$ is a constant term controlling for document length (hence it’s associated with the party or politician)

The *sign* of $\beta_i$ represents the *ideological direction* of $W_i$

The *magnitude* of $\beta_i$ represents the *sensitivity* of the word to ideological differences among speakers or parties

$\Psi$ is a constant term for the word (larger for high frequency words).
“Wordfish” scaling model

Data:
- $Y$ is $N$ (speaker) $\times$ $V$ (word) term document matrix
  - $V \gg N$

Model:

\[
P(Y_i \mid \theta) = \prod_{j=1}^{V} P(Y_{ij} \mid \theta_i)
\]

\[
Y_{ij} \sim \text{Poisson}(\lambda_{ij}) \quad \text{(POIS)}
\]

\[
\log \lambda_{ij} = (g +) \alpha_i + \theta_i \beta_j + \psi_j
\]

Estimation:
- Easy to fit for large $V$ ($V$ Poisson regressions with $\alpha$ offsets)
THE FUTURE

Integration of diverse approaches
FUTURE

Integrate *ex ante* scaling with purely inductive scaling

Integrate scaling techniques with classification methods

Multidimensionality

Better parametric models

Non- or semi-parametric methods for uncertainty

Textbook