Text Analysis in Political Science

Akitaka Matsuo

15 Feb, 2022

Overview

This lecture is a review of the application of text analysis in political science, with a focus on the legislature/legislators.

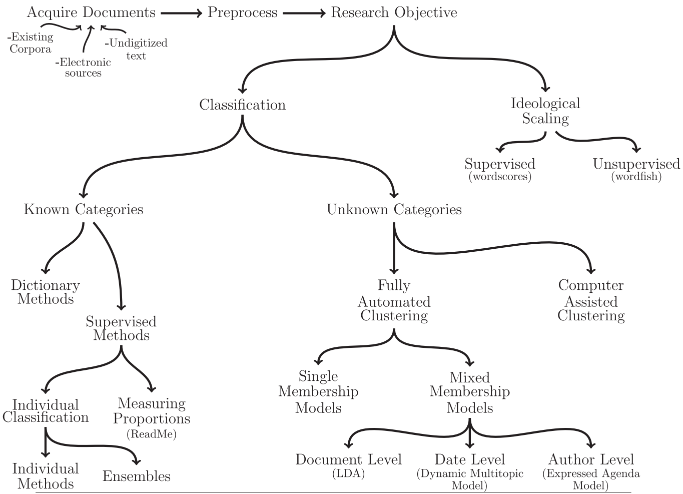
1. Bag-of-Words (BOW) methods
    - Scaling Methods
    - Topicmodels
    - Trends in text analysis
2. Non BOW methods
    - Word embedding
    - BERT and GPT
3. Final words

Bag-of-Words Methods

Bag-of-Words

- In most text analysis in political science, we use Bag-of-Words methods where only word frequencies in documents matter
  - No syntactic information
- The input data takes a form of a document-future matrix (DFM, i.e. document-term matrix (DTM)).
- DFM:
  - rows: documents (i.e. news articles, parliamentary speeches, government reports, Tweets)
  - columns: words (or other features of the text)
- Due to the development of methodology in computer science, political scientists also started using non-BOW methods, but still BOW methods are dominant.

Available BOW methodology



Acquire Documents → Preprocess → Research Objective

-Existing Corpora

-Electronic sources

-Undigitized text

Classification

Ideological Scaling

Supervised (wordscores)

Unsupervised (wordfish)

Known Categories

Unknown Categories

Dictionary Methods

Supervised Methods

Fully Automated Clustering

Computer Assisted Clustering

Individual Classification

Measuring Proportions (ReadMe)

Single Membership Models

Mixed Membership Models

Individual Methods

Ensembles

Document Level (LDA)

Date Level (Dynamic Multitopic Model)

Author Level (Expressed Agenda Model)

Fig. 1 An overview of text as data methods.

From Grimmer and Stewart (2013)

Stardard steps for BOW methods

This is how most text analysis with BOW methods works

1. Text cleaning
2. Tokenization (stemming/lemmatization, feature selection)
3. Construct DFM (feature selection)
4. Model estimation (or other methods, such as dictionary use)
5. Inference

Scaling

Scaling is a dimension reduction methods, popular among political science scholars.

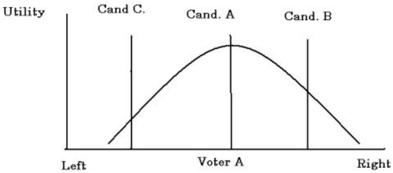
In the 2000s, text analysis was dominated by scaling, which reduces a text to a low-dimensional space. Specifically, it estimates positions in policy space and ideological positions from texts.

Example

- Wordscores (Laver, Benoit, and Garry 2003).
- Wordfish (Slapin and Proksch 2008).

Why?

- Affinity with spatial voting theory
  - There was an accumulation of theories that could be applied as long as the actor positions were known
  - In Congress with roll call voting, there was an established method of measurement.



Scaling with supervision: Wordscores

Computer text analysis in political science has been in vogue since the 2000's. One of the earliest methods is Wordscores (Laver, Benoit, and Garry 2003).

- Originally, the development was motivated by dissatisfaction with the reliability of the Comparative Manifesto Project (CMP) coding.
- **Setting**
  - Provide a **reference text** (a document with known positions, extreme actors, and a representative of those positions).
  - Calculate a score for each word from a comparison of word use in the reference text, and calculate the positions of other actors as an average of the scores.
- Although there is no mentions about ML, it can be considered as a supervised learning model.

Pros and cons of Wordscores

- Words that do not exist in the reference text are excluded from the analysis
  - Disadvantages: positions of other actors tend to be more central (Lowe 2008) (c.f. modified by Martin and Vanberg (2008))
  - Pros: easy to fix context
- Cannot be used in situations where there are no actors with known positions

Scaling without supervision: Wordfish

Based on these features of Wordscores, but without the need for references, Wordfish (Slapin and Proksch 2008) was developed

- Using only the document feature matrix as input, positions are reduced to one dimension
- Estimate word scores and positions simultaneously
  - Similar to methods for estimating positions from roll call votes (W-Nominate (Poole and Rosenthal 2000), Ideal (Clinton, Jackman, and Rivers 2004)).
- Reference to Computer Science. Position as unsupervised learning

Pros and cons of Wordfish.

- Unsupervised, so what the axes mean depends on the list of words that come up (e.g.. Curini, Hino, and Osaka 2018)
- Corpus with diverse topics is likely to produce strange results (c.f. Wordshoal Extension by Lauderdale and Herzog (2016))

Unsuvervised classification: Topicmodels

Even though Scaling has attracted a great deal of interest, the growing interest in other methods has led to the use of other models. Especially widely used are topicmodels.

Basic Idea

- Classify a corpus of documents into a number of topics.
- Unsupervised learning models are often used, but supervised models also exist (although in this case they are not much different from ordinary classification models).

Topicmodels: LDA

There are many variations of topic models, but the basic model is Latent Dirichlet Allocation (LDA) model (Blei, Ng, and Jordan 2012).

Features of the LDA model

- Think of each document as a Dirichlet distribution of topics
- For example: [Topic 1: 0.3, Topic 2: 0.3, Topic 3: 0.4]
- For each word slot, we first draw a topic from this topic distribution
- Since topics are a multinomial distribution of words, draw words from this distribution.

Actual steps when using the LDA model

- Determine the number of topics k
- Estimate
- Determine what each topic is talking about
  - Look at words that are characteristic of the topic (highest probability, frequency, exclusivity, etc.)
  - Read documents with high probability on each topic

Structural Topic Model (STM)

STM (Roberts, Stewart, and Tingley 2015; Roberts et al. 2014) is a model that is similar to LDA in topic and word draw, but more suitable for the purpose of explanation by introducing variables related to the topic.

STM Features

- Allows for correlation between topics
- Can use variables that affect the distribution of topics (e.g., time, attributes, gender)
- Easy to use stm package (searchK, plot.stm)

Winner of 2018 Polmeth Software Award. Widely used (58 published papers, https://www.structuraltopicmodel.com/).

Topicmodels

Why are they so popular? (especially STM).

- Rich stories can be constructed by interpreting topics
- Easy to analyze with R package (quanteda + stm/topicmodels)
- Used to be cool

difficulty? .

- Interpretation of topics is up to the researcher (Chang et al. 2009)
- There is no way to uniquely determine the number of topics
- Model output varies widely from one estimation to the next, depending on initial values
  - Problems with reproducibility of results
  - Room for cherry picking?
  - Each time we estimate, we have to check the content of the topics
  - searchK is no exception
- What to do with topics that cannot be interpreted?
  - Is it legitimate to use a large k and only a small one?
- `keyATM` attempts to fix some of them (Eshima, Imai, and Sasaki 2020)

Machine Learning

Natural language processing is a big field in computer science. How our work is related to what they do?

Very roughly speaking, differences are like this:

Discipline	Political Science	Machine Learning
Objective	Explanation	Prediction
Outcome Variable	Various	Categorical

- Political scientists love scaling, but CS scholars not so much.
    - Unsupervised: Dimension reduction
    - Regression problem
- Typically, machine learning tackles the problem of how to increase the accuracy of predictions based on data that is already labeled with categories.
- Avoiding the problem of overfitting by using test-train splitting

Will political science move in this direction? $\Rightarrow$ Probably not.

A study that makes good use of Machine Learning-like methods

Peterson, Andrew Arthur Spirling. 2018. "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems."*Political Analysis* 26(01):120–128.

- The idea of using prediction error as a measure of polarization.
- Using the content of congressional speeches to predict the party affiliation of the speaker.
- The more polarization, the better the prediction.

Slapin, Jonathan B. and Justin H. Kirkland. 2019. "The Sound of Rebellion: Voting Dissent and Legislative Speech in the UK House of Commons." *Legislative Studies Quarterly* (June):lsq.12251.

- Can we predict dissent from the content of speech?
- First person, simple speech predicts sedition to some extent.

The first person, simple speech predicts rebellion to some extent. Rather, the direction is to see if predictions can be made, and if so, what determines them. This may be the potential for machine learning to be seen as a valuable research tool in political science.

Multi-linugal Comparisons

One of the barriers to applying text analysis to the study of general topics in comparative politics is the question of how to analyze texts written in different languages. It has to be bridged in some way, and there is a lot of research being done to approach this problem.

- The possibility of using machine translation (Lucas et al. 2015).
- Use corpora that have official translations (e.g., European Parliament).
- If this is not available, validation using external data is basically required.
    - Comparison with expert surveys
    - Human coding
        - Expert
        - Crowd

Examples of Multi-linugal Comparisons

de Vries, Erik, Martijn Schoonvelde Gijs Schumacher. 2018. “No Longer Lost in Translation. Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications.”*Political Analysis*. pp. 1–31.

- Methodological study, using a corpus with official translations.
- Using europerl dataset, with some machine translation
- The results of topicmodel can be reproduced quite well with machine translated text

Proksch, Sven Oliver, Will Lowe, Jens Wäckerle Stuart Soroka. 2018. “Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches.”*Legislative Studies Quarterly*. (September):1–35.

- A Comparative Study of Sentiment in Congressional Speeches
- A dictionary-based sentiment analysis in which the dictionary is machine-translated and applied to texts from different countries. Analyzed differences in sentiment among members of the ruling and opposition parties toward bills proposed by the cabinet.
  - Considerable effort before and after machine translation (dictionary expansion, human checks)
- Used europerl for validation (EUP, the State of the Union Debate)

Beyond Bag-of-Words

Issues with Bag-of-Words

The methods we have seen so far are using BOW as input. There are several potential issues:

1. Ignoring context (obviously)
    - Negation etc
    - Homograph
2. Extreme sparsity of input data
    - If the training data covers small subset of the words in a corpus, we can't make informed inferences for the test/new data

Word Embeddings

- Word embedding is a dense vector for each word
  - Dimension of the vector is fixed
  - Similar word can have a similar vector (finding synonyms in multidimensional space)
  - Word arithmetic
    - King - Man + Woman = ? (Mikolov et al. 2013)
    - “Man is to Computer Programmer as Woman is to Homemaker” (Bolukbasi et al. 2016)
- See Rodriguez and Spirling (2022)

Methodology for estimating wordembeddings

- word2vec (Mikolov et al. 2013)
- Glove (Pennington, Socher, and Manning 2014)
- Both use the contextual information for estimation (e.g. sequences of words (word2vec), or word cooccurrence with narrow window (Glove))

What can we do with Word Embeddings

- The basic idea is to measure some specific words with other words
    - Having a small set of words that measures substantive interests, measure the relative distance from these words to other words in the corpus
- For two research questions, word embeddings particularly useful (Rodriguez and Spirling 2022)
    - Measuring
        - How the concept evolved over time (Rodman 2019)
        - When politicians use emotive rhetoric (Osnabrügge, Hobolt, and Rodon 2021)
        - Measuring styles (Hargrave and Blumenau 2022)
    - Finding the nature of political conflict
        - Finding the focus of partisan rhetoric (Rheault and Cochrane 2019)

Pretrained or train by yourself

When we use word embedding, there are esstially two ways to get word vectors:

1. Use pre-trained model
2. Train by yourself using your corpus

Pretrained vector

- Very small computation cost
- Usually using much larger corpus (reliable estimates)
- May not fit for some purpose (e.g. Rodman (2019))

Train by yourself

- The corpus should be large enough
- You can incorporate the domain specificity (e.g. Parliamentary speeches in the UK and Ireland, are they the same?)
- When corpus is big, the cost is high

Issues with word embedding

- One vector for each words
  - Contexts are not fully incorporated
- Limited benefit for aggregated level analysis
  - It is typical to measure the document level scale with the similarity measure of all words (e.g. Osnabrügge, Hobolt, and Rodon (2021) and Hargrave and Blumenau (2022), also Watanabe (2021))
  - But, is this really the best? There is no proper way to calculate the score at the aggregate level
    - Does average distance matter?
    - Estimating embedding with contextual information but ignore in the inference. . .

Sentence encoders: BERT and GPT

- The trend in the research is to employ transfer learning with large-scale pre-trained language models such as BERT and GPT
- BERT and GPT
    - Both Transformer based technology
        - Self-attention based mechanism
- BERT (Devlin et al. 2019)
    - Bidirectional Encoder Representation from Transformer
    - Works well for many NLP tasks
        - Machine translation
        - Text summary
        - Question and Answers
        - Text classification
- GPT-2/3
    - Autregressive language model
    - Very capable of text generation
    - Can do NLP tasks, but not with transfer learning, but with few-shot leanring (Brown et al. 2020)

So what can we do with BERT/GPT?

- BERT is handy for us:
  - Not too big
    - You can work with it at reasonable costs
  - Fine tuning for transfer learning with your labelled data is easy
    - Numerous instructions are available online
    - You can fine-tune with a single GPU instance (e.g. a desktop computer with dedicated GPU, Google Colab or maybe CERES)
- GPT-2/3 is a much bigger model
  - API access
  - Very capable
  - Parameter update is not possible/impractical

BERT Application: Position detection in parliamentary speeches

- Corpus: Parliamenary “debates" in Japan
  - Period: 1955-2015
  - N: 6800 (1360 Train, 5440 Test)
- Output variable: position expressed in the speech (Expressed approval or disapproval of a bill)
  - Words with clear positions are removed
- Models to compare:
  - Pre-trained BERT (base) with transfer learning
  - Naive Bayes
- This is not a particularly difficult task for Naive Bayes but the performance improvement is obvious.

Model	Data	Accuracy	positive-F1
BERT	Train	0.995	0.991
BERT	Test	0.939	0.892
Naive Bayes	Train	0.931	0.887
Naive Bayes	Test	0.913	0.847

Final words

- Domain knowledge is extremely important. We should know about the context in which the text was generated. This will affect:
  - Choice of method (do we need a topicmodel for a newspaper article?)
  - Interpreting the results (e.g., Wordfish position estimation)
    - What do the axes mean?
    - Does the position reflect the true intention of the speaker?
  - User information for Twitter accounts (e.g. text can be used as a scaling method, but so can the followership network (Barberá 2015))
- Change the method used depending on the size of the corpus and the size of each document. For example
  - If the corpus size is not very large, the original word embedding model is not very useful unless the corpus size is quite large.
  - Consider how many topics can be assumed for a short document
  - Dictionary-based analysis?
- In the end, the purpose of the analysis and the nature of the corpus will determine what to do with the text.
  - Computers still do not understand the language as humans do.

References I

Barberá, Pablo. 2015. “Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data.” *Political Analysis* 23: 76–91. https://doi.org/10.1093/pan/mpu011.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. “Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data.” *American Political Science Review* 110 (2): 278–95. https://doi.org/10.1017/S0003055416000058.

Benoit, Kenneth, Kevin Munger, and Arthur Spirling. 2019. “Measuring and Explaining Political Sophistication through Textual Complexity.” *American Journal of Political Science* 63 (2): 491–508. https://doi.org/10.1111/ajps.12423.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2012. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3 (4-5): 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993.

References II

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." *30th Conference on Neural Information Processing Systems*, no. NIPS 2016 (July): 1–9. https://code.google.com/archive/p/word2vec/%20http://arxiv.org/abs/1607.06520.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language models are few-shot learners." *Advances in Neural Information Processing Systems* 2020-December. http://arxiv.org/abs/2005.14165.

Catalinac, Amy. 2016. "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections." *The Journal of Politics* 78 (1): 1–18. https://doi.org/10.1086/683073.

Chang, Jonathan, Sean Gerrish, Chong Wang, and David M Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." *Advances in Neural Information Processing Systems 22*, 288——296. https://doi.org/10.1.1.100.1089.

References III

Clinton, Joshua D., Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *The American Political Science Review* 98 (2): 355–70. https://doi.org/http://dx.doi.org/10.1017/S0003055404001194.

Curini, Luigi, Airo Hino, and Atsushi Osaka. 2018. "The Intensity of Government-Opposition Divide as Measured through Legislative Speeches and What We Can Learn from It: Analyses of Japanese Parliamentary Debates, 1953-2013." *Government and Opposition*, 1–18. https://doi.org/10.1017/gov.2018.15.

Devlin, Jacob, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of deep bidirectional transformers for language understanding." *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (Mlm): 4171–86. http://arxiv.org/abs/1810.04805.

Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki. 2020. "Keyword Assisted Topic Models," April. http://arxiv.org/abs/2004.05964.

References IV

Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Hargrave, Lotte, and Jack Blumenau. 2022. "No Longer Conforming to Stereotypes? Gender, Political Style and Parliamentary Debate in the UK." *British Journal of Political Science*, 1–18. https://doi.org/10.1017/s0007123421000648.

Lauderdale, Benjamin E., and Alexander Herzog. 2016. "Measuring Political Positions from Legislative Speech." *Political Analysis* 24 (03): 374–94. https://doi.org/10.1093/pan/mpw017.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97 (02): 311–31. https://doi.org/10.1017/S0003055403000698.

Lowe, Will. 2008. "Understanding Wordscores." *Political Analysis* 16: 356–71. https://doi.org/10.1093/pan/mpn004.

References V

Lucas, C., R. a. Nielsen, Margaret E. Roberts, B. M. Stewart, A. Storer, and D. Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis*, 254–77. https://doi.org/10.1093/pan/mpu019.

Martin, Lanny W., and Georg Vanberg. 2008. "A robust transformation procedure for interpreting political text." *Political Analysis* 16 (1): 93–100. https://doi.org/10.1093/pan/mpm010.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space," 1–12. http://arxiv.org/abs/1301.3781.

Moser, Scott, and Andrew Reeves. 2014. "Taking the Leap: Voting, Rhetoric, and the Determinants of Electoral Reform." *Legislative Studies Quarterly* 39 (4): 467–502. https://doi.org/10.1111/lsq.12055.

Osnabrügge, Moritz, Sara B. Hobolt, and Toni Rodon. 2021. "Playing to the Gallery: Emotive Rhetoric in Parliaments." *American Political Science Review* 115 (3): 885–99. https://doi.org/10.1017/S0003055421000356.

References VI

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "Glove: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 19:1532–43. 5. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162.

Peterson, Andrew, and Arthur Spirling. 2018. "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems." *Political Analysis* 26 (01): 120–28. https://doi.org/10.1017/pan.2017.39.

Poole, Keith T., and Howard Rosenthal. 2000. *Congress : A Political-Economic History of Roll Call Voting*. Oxford University Press.

Proksch, Sven Oliver, Will Lowe, Jens Wäckerle, and Stuart Soroka. 2019. "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches." *Legislative Studies Quarterly* 44 (1): 97–131. https://doi.org/10.1111/lsq.12218.

Rheault, Ludovic, and Christopher Cochrane. 2019. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis*, July, 1–22. https://doi.org/10.1017/pan.2019.26.

References VII

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2015. “stm: R Package for Structural Topic Models.” *Journal of Statistical Software* VV.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. “Structural Topic Models for Open-Ended Survey Responses.” *American Journal of Political Science* 58 (4): 1064–82. https://doi.org/10.1111/ajps.12103.

Rodman, Emma. 2019. “A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors.” *Political Analysis*, July, 1–25. https://doi.org/10.1017/pan.2019.23.

Rodriguez, Pedro L., and Arthur Spirling. 2022. “Word Embeddings: What Works, What Doesn’t, and How to Tell the Difference for Applied Research.” *Journal of Politics* 84 (1): 101–15. https://doi.org/10.1086/715162.

Slapin, Jonathan B., and Justin H. Kirkland. 2019. “The Sound of Rebellion: Voting Dissent and Legislative Speech in the UK House of Commons.” *Legislative Studies Quarterly*, no. June (July): lsq.12251. https://doi.org/10.1111/lsq.12251.

References VIII

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–22. https://doi.org/10.1111/j.1540-5907.2008.00338.x.

Vries, Erik de, Martijn Schoonvelde, and Gijs Schumacher. 2018. "No Longer Lost in Translation. Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications." *Political Analysis*, 1–31.

Watanabe, Kohei. 2021. "Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages." *Communication Methods and Measures* 15 (2): 81–102. https://doi.org/10.1080/19312458.2020.1832976.