ECPR Methods Summer School: Big Data Analysis in the Social Sciences

Pablo Barberá

London School of Economics

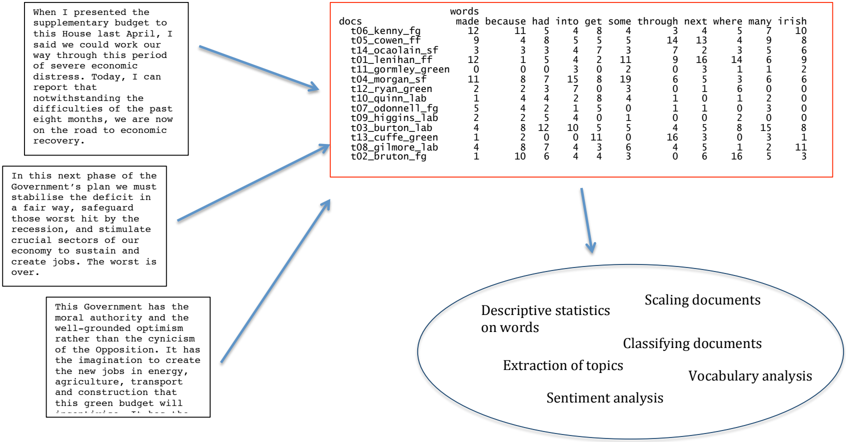
pablobarbera.com

Course website:

pablobarbera.com/ECPR-SC105

Automated Analysis of Large-Scale Textual Data

Workflow: analysis of text



When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.

In this next phase of the Government’s plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.

This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will incentivise. It has the

```
                    words
docs                 made because had into get some through next where many irish
  t06_kenny_fg         12       11   5    4   8    4       3    4     5    7    10
  t05_cowen_ff          9        4   8    5   5    5      14   13     4    9     8
  t14_ocaolain_sf       3        3   3    4   7    3       7    2     3    5     6
  t01_lenihan_ff       12        1   5    4   2   11       9   16    14    6     9
  t11_gormley_green     0        0   0    3   0    2       0    3     1    1     2
  t04_morgan_sf        11        8   7   15   8   19       6    5     3    6     6
  t12_ryan_green        2        2   3    7   0    3       0    1     6    0     0
  t10_quinn_lab         1        4   4    2   8    4       1    0     1    2     0
  t07_odonnell_fg       5        4   2    1   5    0       1    1     0    3     0
  t09_higgins_lab       2        2   5    4   0    1       0    0     2    0     0
  t03_burton_lab        4        8  12   10   5    5       4    5     8   15     8
  t13_cuffe_green       1        2   0    0  11    0      16    3     0    3     1
  t08_gilmore_lab       4        8   7    4   3    6       4    5     1    2    11
  t02_bruton_fg         1       10   6    4   4    3       0    6    16    5     3
```

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

Vocabulary analysis

Sentiment analysis

Why quantitative analysis of text?

Justin Grimmer's haystack metaphor: automated text analysis improves reading

- Analyzing a straw of hay: understanding meaning
  - Humans are great! But computer struggle
- Organizing the haystack: describing, classifying, scaling texts
  - Humans struggle. But computers are great!
  - (What this course is about)

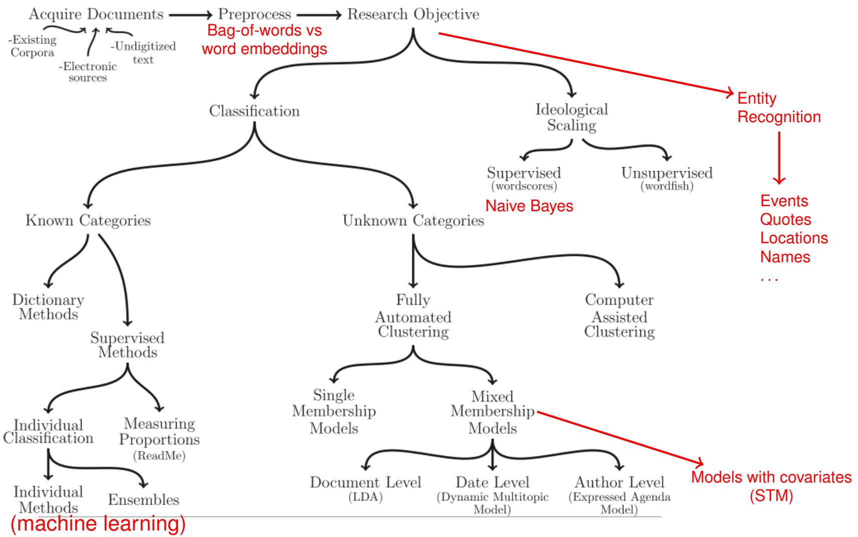
Principles of automated text analysis (Grimmer & Stewart, 2013)

1. All quantitative models are wrong – but some are useful
2. Quantitative methods for text amplify resources and augment humans
3. There is no globally best method for text analysis
4. Validate, validate, validate

Quantitative text analysis requires assumptions

1. Texts represent an observable implication of some underlying characteristic of interest
   - An attribute of the author of the post
   - A sentiment or emotion
   - Salience of a political issue
2. Texts can be represented through extracting their *features*
   - most common is the bag of words assumption
   - many other possible definitions of “features” (e.g. n-grams)
3. A document-feature matrix can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

Overview of text as data methods



Bag-of-words vs word embeddings

Entity Recognition

Events
Quotes
Locations
Names
. . .

Naive Bayes

Models with covariates (STM)

(machine learning)

Fig. 1 in Grimmer and Stewart (2013)

Some key basic concepts

(text) corpus a large and structured set of texts for analysis

document each of the units of the corpus (e.g. a FB post)

types for our purposes, a unique word

tokens any word – so token count is total words

e.g.

```
A corpus is a set of documents.
This is the 2nd document in the corpus.
```

is a corpus with 2 documents, where each document is a sentence. The first document has 6 types and 7 tokens. The second has 7 types and 8 tokens. (We ignore punctuation for now.)

Some more key basic concepts

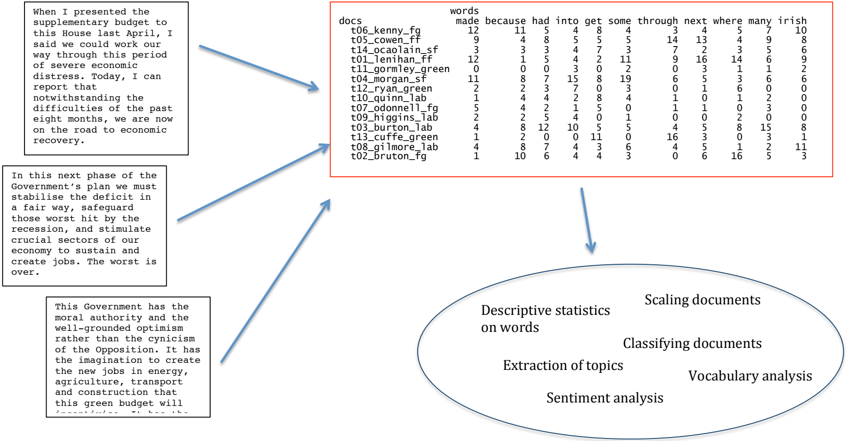
stems words with suffixes removed (using set of rules)

lemmas canonical word form (the base form of a word that has the same meaning even when different suffixes or prefixes are attached)

word	win	winning	wins	won	winner
stem	win	win	win	won	winner
lemma	win	win	win	win	win

stop words Words that are designated for exclusion from any analysis of a text

We generally adopt a bag-of-words approach



When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.

In this next phase of the Government’s plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.

This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will incentivise. It has the

```
                    words
docs                 made because had into get some through next where many irish
  t06_kenny_fg         12       11   5    4   8    4       3    4     5    7    10
  t05_cowen_ff          9        4   8    5   5    5      14   13     4    9     8
  t14_ocaolain_sf       3        3   3    4   7    3       7    2     3    5     6
  t01_lenihan_ff       12        1   5    4   2   11       9   16    14    6     9
  t11_gormley_green     0        0   0    3   0    2       0    3     1    1     2
  t04_morgan_sf        11        8   7   15   8   19       6    5     3    6     6
  t12_ryan_green        2        2   3    7   0    3       0    1     6    0     0
  t10_quinn_lab         1        4   4    2   8    4       1    0     1    2     0
  t07_odonnell_fg       5        4   2    1   5    0       1    1     0    3     0
  t09_higgins_lab       2        2   5    4   0    1       0    0     2    0     0
  t03_burton_lab        4        8  12   10   5    5       4    5     8   15     8
  t13_cuffe_green       1        2   0    0  11    0      16    3     0    3     1
  t08_gilmore_lab       4        8   7    4   3    6       4    5     1    2    11
  t02_bruton_fg         1       10   6    4   4    3       0    6    16    5     3
```

Bag-of-words approach

From words to numbers:

1. Preprocess text: lowercase, remove stopwords and punctuation, stem, tokenize into unigrams and bigrams (bag-of-words assumption)

   "A corpus is a set of documents."

   "This is the second document in the corpus." "a corpus is a set of documents."

   "this is the second document in the corpus." "~~a~~ corpus ~~is a~~ set ~~of~~ documents~~.~~"

   "~~this is the~~ second document ~~in the~~ corpus~~.~~" "corpus set document~~s~~"

   "second document corpus" [corpus, set, document, corpus set, set document]

   [second, document, corpus, second document, document corpus]

2. Document-feature matrix:
   - **W**: matrix of N documents by M unique n-grams
   - w_{im}= number of times m-th n-gram appears in i-th document.

orpus	et	locument	orpus set	...	M n-grams

Word frequencies and their properties

Bag-of-words approach disregards grammar and word order and uses word frequencies as features. Why?

- *Context is often uninformative*, conditional on presence of words:
  - Individual word usage tends to be associated with a particular degree of affect, position, etc. without regard to context of word usage
- Single words tend to be the most informative, as co-occurrences of multiple words (*n*-grams) are rare
- Some approaches focus on occurrence of a word as a binary variable, irrespective of frequency: a binary outcome
- Other approaches use frequencies: Poisson, multinomial, and related distributions

Dictionary Methods

Dictionary methods

Classifying documents when categories are known:

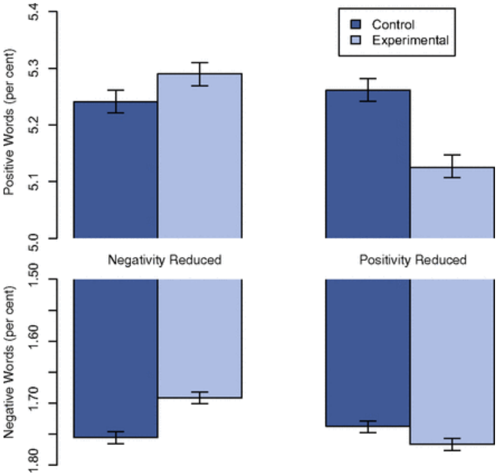
- Lists of words that correspond to each category:
    - Positive or negative, for sentiment
    - Sad, happy, angry, anxious... for emotions
    - Insight, causation, discrepancy, tentative... for cognitive processes
    - Sexism, homophobia, xenophobia, racism... for hate speech

      many others: see LIWC, VADER, SentiStrength, LexiCoder...
- Count number of times they appear in each document
- Normalize by document length (optional)
- Validate, validate, validate.
    - Check sensitivity of results to exclusion of specific words
    - Code a few documents manually and see if dictionary prediction aligns with human coding of document

Linquistic Inquiry and Word Count

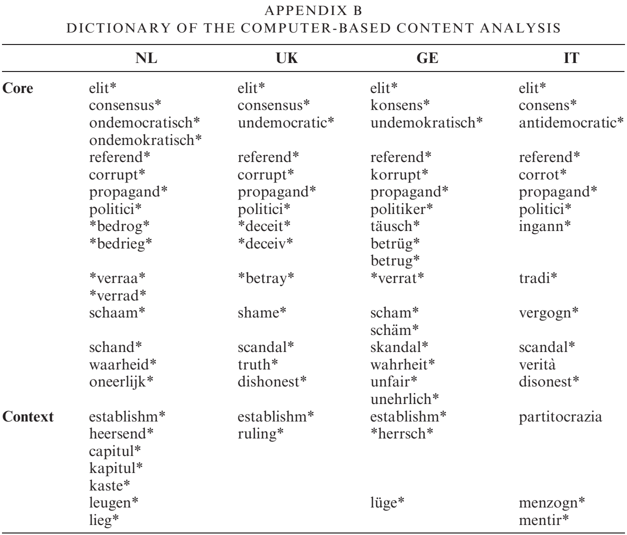
- Created by Pennebaker et al — see `http://www.liwc.net`
- Uses a dictionary to calculate the percentage of words in the text that match each of up to 82 language dimensions
- Consists of about 4,500 words and word stems, each defining one or more word categories or subdictionaries
- For example, the word *cried* is part of five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. So observing the token *cried* causes each of these five subdictionary scale scores to be incremented
- Hierarchical: so "anger" are part of an *emotion* category and a *negative emotion* subcategory
- You can buy it here: `http://www.liwc.net/descriptiontable1.php`

Example: Emotional Contagion on Facebook



Source: Kramer et al, PNAS 2014

Potential advantage: Multi-lingual



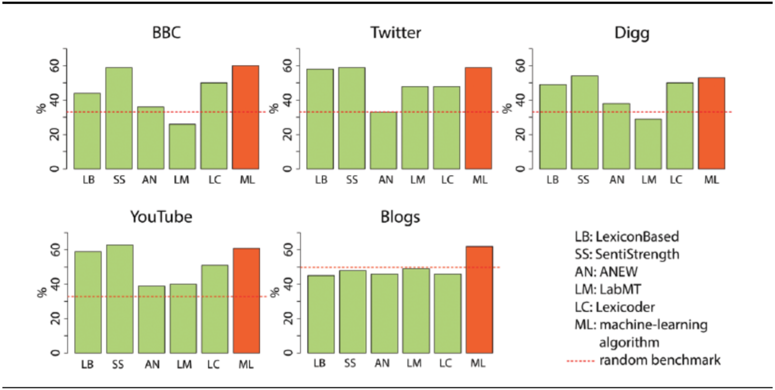
APPENDIX B
DICTIONARY OF THE COMPUTER-BASED CONTENT ANALYSIS

	NL	UK	GE	IT
Core	elit*	elit*	elit*	elit*
	consensus*	consensus*	konsens*	consens*
	ondemocratisch* ondemokratisch*	undemocratic*	undemokratisch*	antidemocratic*
	referend*	referend*	referend*	referend*
	corrupt*	corrupt*	korrupt*	corrot*
	propagand*	propagand*	propagand*	propagand*
	politici*	politici*	politiker*	politici*
	bedrog *bedrieg*	*deceit* *deceiv*	täusch* betrüg* betrug*	ingann*
	verraa *verrad*	*betray*	*verrat*	tradi*
	schaam*	shame*	scham* schäm*	vergogn*
	schand*	scandal*	skandal*	scandal*
	waarheid*	truth*	wahrheit*	verità
	oneerlijk*	dishonest*	unfair* unehrlich*	disonest*
Context	establishm*	establishm*	establishm*	partitocrazia
	heersend* capitul* kapitul* kaste*	ruling*	*herrsch*	
	leugen* lieg*		lüge*	menzogn* mentir*

(from Rooduijn and Pauwels 2011)

Potential disadvantage: Context specific

Lexicons' Accuracy in Document Classification
Compared to Machine-Learning Approach



Source: González-Bailón and Paltoglou (2015)

How to build a dictionary

- The ideal content analysis dictionary associates all and only the relevant words to each category in a perfectly valid scheme
- Three key issues:

Validity	Is the dictionary's category scheme valid?
Recall	Does this dictionary identify *all* my content?
Precision	Does it identify *only* my content?

- Imagine two logical extremes of including all words (too sensitive), or just one word (too specific)

How to build a dictionary

1. Identify “extreme texts” with “known” positions. Examples:
   - Tweets by populist vs mainstream parties (for populism dictionary)
   - Facebook comments to news about natural catastrophes vs football victories (for sentiment dictionary)
   - Subreddits for white nationalist groups vs regular politics (for racist rhetoric)
2. Search for differentially occurring words using word frequencies
3. Examine these words in context to check their precision and recall
4. Use regular expressions to see whether stemming or wildcarding is required