HARVARD | BUSINESS | SCHOOL

Using Textual Data in Research

Ian Gow

HBS Common Connection

6 October 2016

Big picture

- We live in a time of exponential growth in the amount of data
    - Much of this data is textual, either in nature or form
- This growth creates opportunities for researchers
- But capitalizing on these opportunities requires new skills …
- … and new ways of working

Goals of this talk

- Highlight a few interesting sources of textual data
  - How to collect
  - How to organize and store
- Describe a few ways that textual data can be used
  - As sources of discrete data
  - As linguistic data
- Showcase a few tools
- Provide a few examples of how I have used textual data in my research

Limits

- “Deep” linguistic analysis
- Audio or video
- Statistical (machine) learning (hints only)
- Fancy proprietary software
- “Big data”

Limits: My background

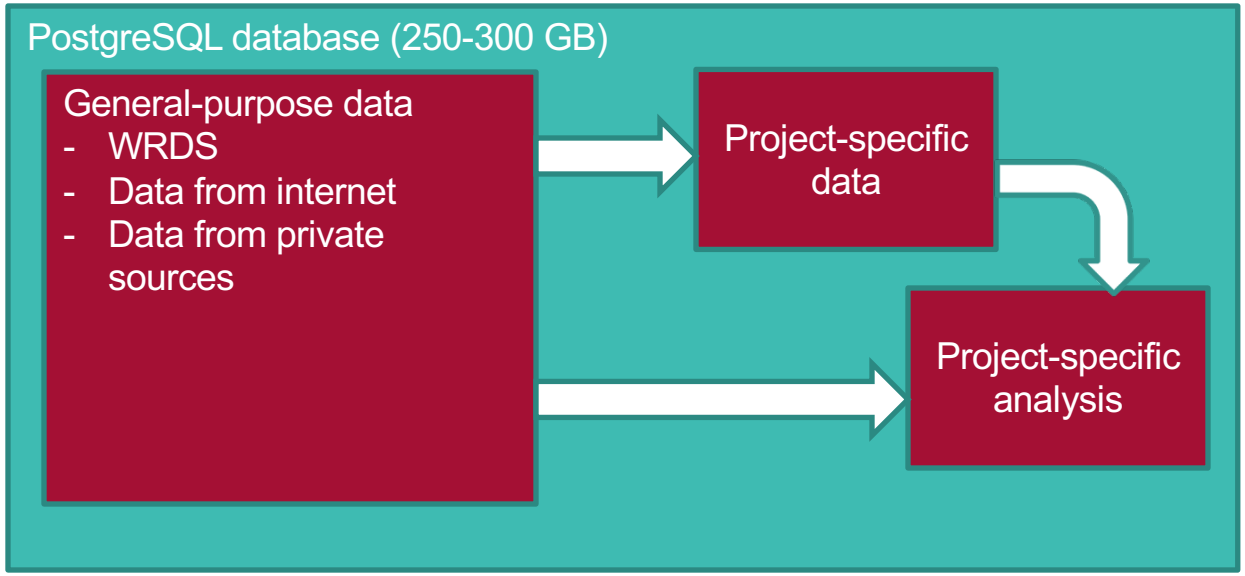
Sources of textual data

- StreetEvents: 300,000+ conference calls, etc.
- SEC filings
- Press filings
- Newspapers
- … the internet

Tools: Programming languages

- Regular expressions
- SAS
- Stata
- Perl
- Python
- R
- SQL

Database



Some data is kept in system files (e.g., raw SEC filings).

Database: Value

- Tends to force you to address messy data issues upfront
  - Encoding! (Most text comes in UTF-8 nowadays.)
- Makes tool re-use easier
- Facilitates using the best tools for the job
  - Stanford NLP tools are in Java, but other tools are in Perl or Python and analysis might be done in R or Stata
  - Finding Python experts might be easier than finding Stata whizzes
- Facilitates parallel processing
  - Many tasks are “embarassingly parallel” but costly
- Makes sharing data easier
  - Text data can be “big”

Source: SEC filings

- Financial reports (10-K, 10-Q)
- Proxy statements
- Comment letters
- Court orders
- Investment holdings (13-F)

Example: Comment letters

- SEC staff issue comment letters in connection with their review of disclosure filings. May request that a company
  - Provide additional information so the staff can better understand the company's disclosure,
  - Amend an SEC filing to revise a disclosure or provide additional disclosure
  - Provide additional or different disclosure in a future filing
- The SEC began publicly releasing this correspondence in 2005 for filings made after August 1, 2004 that were reviewed by SEC staff.
  - But letters are in PDF format

Example: Director biographies

- Directors provide biographies to shareholders
- Some casual empiricism suggested some strategic choices in disclosure
- To examine this properly, we needed a panel data set of director biographies.
  - No vendor collects these data
- So we needed to go to the raw (text) data.
  - Collected bios using RAs (using Google Sheets)
  - Collected *some* directorships using RAs
  - Use regular expressions to identify directorships mentioned.

Source: StreetEvents

- Transcripts of more than 300,000 conference calls, etc.
- Data updates provided monthly
- Data provided as XML files
- Processed using Perl scripts into “clean” data stored in PostgreSQL
- Then easily subjected to analysis

StreetEvents data: Fog

- Fog is a measure of linguistic complexity.
- Used by Feng Li in his dissertation (JAE 2008)
- So I emailed Feng for the code.

StreetEvents data: Fog

Hi Ian,

I used Perl for the calculations. Here is the core part, it's actually very simple. You need to save the file to a .txt file and download and install a Perl module called "`Lingua::EN::Fathom`".

```
use Lingua::EN::Fathom;
my $text = new Lingua::EN::Fathom;
$text->analyse_file("file.txt");

$percent_complex_words = $text->percent_complex_words;
$num_sentences     = $text->num_sentences;
$words_per_sentence = $text->words_per_sentence;

$fog   = $text->fog;
Print "$cik, $filedate, $file, $percent_complex_words, $num_sentences, ";
Print "$words_per_sentence, $fog, \n";
```

Let me know if you need more info about this.

Life is fine, busy with the recruiting stuff lately. Hope all is well with you.

Feng

Note: `sudo cpan -i Lingua::EN::Fathom` installs this module.

StreetEvents data: Fog

Fortunately, PostgreSQL allows you to define functions using other languages, including Perl, Python, R, C, SQL, JavaScript.

```
CREATE OR REPLACE FUNCTION fog_original(text)
  RETURNS double precision AS $BODY$
  # Load Perl modules that calculate fog, etc.
  use Lingua::EN::Fathom;
  my $text = new Lingua::EN::Fathom;
  if (defined($_[0])) {
    $text->analyse_block($_[0]);
    return($text->fog);
  }
$BODY$ LANGUAGE plperlu;
```

StreetEvents: Other linguistic features

- LIWC, etc.
- Non-answers
- NER

Mapping features to personality

Feature	Description	**Panel A:** CEOs with data on Big Five				**Panel B:** Without Big Five	
		ghSMART		O'Reilly			
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
	Personal pronouns						
First per. singular	LIWC "I": I, me, mine, etc.	29.05	13.04	29.88	11.69	29.72	13.52
First per. plural	LIWC "we": we, us, our, etc.	97.21	23.48	85.33	18.45	93.75	22.87
	Negative emotion words						
Sadness	LIWC "sad": devastat*, disadvantage*, etc.	2.21	2.35	2.41	2.19	2.55	2.76
Anxiety	LIWC "anx": worried, fearful, nervous, etc.	1.48	2.08	1.37	1.47	1.69	2.24
Anger	LIWC "anger": hate, kill, annoyed, etc.	1.36	1.92	1.65	1.79	1.40	2.01
Extreme negative	LZ using LIWC "negemo": absurd, adverse, awful, etc.	3.58	3.37	3.70	2.66	3.64	3.24
	Positive emotion words						
Extreme positive	LZ using LIWC "posemo": fantastic, great, definitely, etc.	10.00	5.62	12.29	5.50	9.17	5.86
Nonextreme positive	LZ using LIWC "posemo": love, nice, accept, etc.	53.69	14.61	48.40	11.84	53.22	15.75
	Agreement						
Negations	LIWC "negate": no, not, never, etc.	18.52	8.01	18.12	6.92	21.02	9.14
Assent	LIWC "assent": agree, OK, yes, etc.	5.18	4.76	3.47	2.55	5.11	4.43
Thanks	Self-constructed: thank you, thanks, you're welcome, etc.	3.71	4.63	1.59	1.94	3.95	4.93

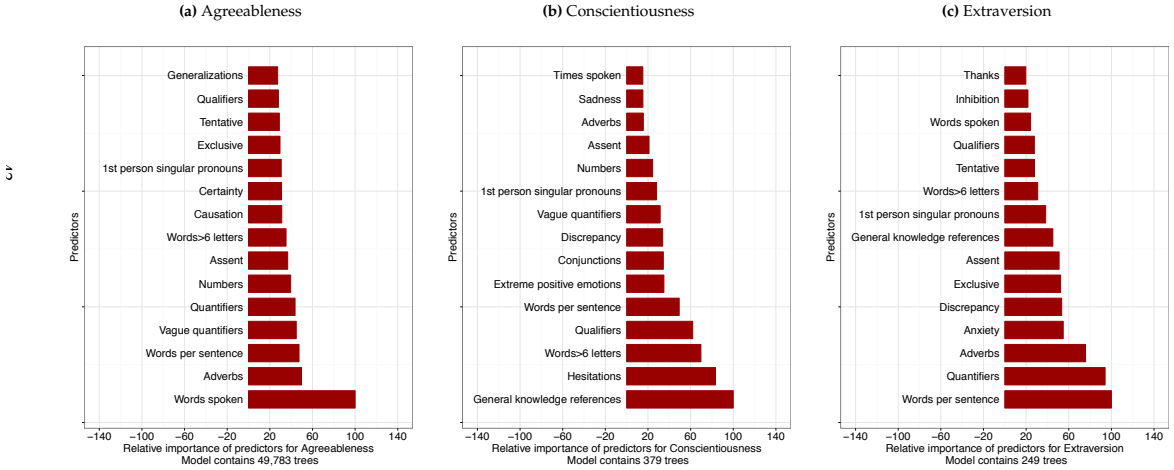
Mapping features to personality (cont)

Feature	Description	**Panel A:** CEOs with data on Big Five				**Panel B:** Without Big Five	
		ghSMART		O'Reilly			
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
		Certainty					
Certainty	LIWC "certain": always, never, etc.	23.71	8.49	24.22	8.10	23.09	8.60
Numbers	Number of numbers	7.92	7.07	8.62	7.22	10.32	8.93
Quantifiers	LIWC "quant": all, a lot, bit, etc.	58.88	12.99	59.67	11.47	58.67	13.51
Tentative	LIWC "tentat": maybe, perhaps, guess, etc.	47.70	15.05	44.15	11.70	50.20	14.89
Vague quantifiers	Self-constructed: a load of, a lot of, etc.	35.13	9.94	32.99	8.79	35.16	10.65
Qualifiers	Self-constructed: arguably, as a whole, etc.	44.70	14.21	42.12	13.23	45.70	14.37
Generalizations	Self-constructed: all that stuff, almost, etc.	28.88	9.46	29.84	8.44	28.93	9.57
		Cognitive process					
Insight	LIWC "insight": admitting, analy*, etc.	42.17	14.31	42.89	11.55	41.37	13.53
Causation	LIWC "cause": allow*, attribut*, based, etc.	30.90	10.28	34.51	9.18	29.35	10.69
Discrepancy	LIWC "discrep": besides, could, etc.	22.85	9.28	21.23	7.11	24.38	9.78
Inhibition	LIWC "inhib": abandon*, abstain*, etc.	5.21	3.75	5.42	3.54	5.66	4.27
Inclusive	LIWC "incl": each, inclu*, inside, etc.	149.26	24.44	140.91	20.25	143.36	24.22
Exclusive	LIWC "excl": either, except, exclu* , etc.	43.26	12.64	41.79	10.14	44.60	12.82
		Hesitations and fillers					
Hesitations	LZ using LIWC "filler": ah, um, etc.	0.08	0.36	0.23	2.32	0.21	1.32
Gen. knowledge	LZ: you know, investors well know, etc.	2.10	4.30	1.97	3.61	2.53	4.75
		Linguistic process					
Times spoken	Number of times spoken	3.75	1.38	3.40	1.18	4.01	1.53
Words spoken	Number of words spoken ignoring articles	2290.63	934.30	2422.13	854.10	2411.02	1009.34
Words per sentence	Words per sentence	22.05	4.36	23.44	4.78	21.47	4.44
Words>6 letters	Words longer than 6 letters	428.29	47.37	433.43	41.88	420.97	50.61
Articles	LIWC "article": a, an, the	122.12	18.83	124.93	14.84	121.28	18.77
Conjunctions	LIWC "conj": although, and, as, etc.	120.39	19.09	123.12	16.26	117.37	18.26
Adverbs	LIWC "adverb": about, absolutely, etc.	98.48	19.70	100.22	17.58	95.94	18.71
Calls obs.		1,186		771		70,329	

Personality: Linguistic features

Figure 3. Importance of Linguistic Features for Predicting Big Five Personality Traits

This figure depicts the relative importance of the top fifteen linguistic features for predicting Big Five traits. For each linguistic feature, importance is computed as the reduction of the squared error attributable to this feature as described in Friedman (2001) using the training (ghSMART) sample. The bars indicate the relative importance of each linguistic features normalized by the relative importance of the most influential linguistic feature (on the bottom).



Personality: Constructing features

- Get LIWC data
- Process data (word lists essentially), put in database
- Extract data (word lists), construct regular expressions
- Match regular expressions to text

Non-answers

- We estimate that about 15% of questions asked on earnings conference calls are not answered.
- How to identify non-answers?
  - 1. Manually code a random sample of answers.
  - 2. Split sample into training and test sub-samples.
  - 3. Develop regular expressions (including word lists) to flag non-answers in training sample.
  - 4. Evaluate performance of regular expressions on test sample.
  - 5. Apply regular expressions to larger data set.