Text as Data

Amy Catalinac

New York University

June 18, 2018
Why a Talk on Text Analysis?

Text is the new frontier of . . .

1. **Data**: lots of it (literally petabytes) on the web... not to mention archives.

2. **Quantitative methods**: for the most part, it is unstructured. Needs to be harvested and modeled.

3. **Social science**: politicians give speeches, thinkers write articles, nations sign treaties, users post on Facebook, etc.

**Today**: overview of quantitative ‘text-as-data’ approaches as strategies to learn more about social scientific phenomena of interest.
Goal of Text Analysis

○ In many (most?) applications of text as data in social science, we are trying to make an inference about a **latent variable**.
  → something which we cannot observe directly but which we can make inferences about from things we can observe.
  → Examples include ideology, priorities, ambition, narcissism, propensity to vote, etc.

○ In **traditional** social science research, we use data that is structured.
  → For example, surveys we wrote to understand ideology, donation decisions to understand which party a voter supports, or roll-call votes to understand where a politician stands on a policy.

○ Here, the thing we **can** observe are the words spoken, the passages written, the issues debated, etc.
The latent variable of interest may pertain to the.. .

**author**  ‘what does this Senator prioritize?’, ‘where is this party in ideological space?’

**document**  ‘does this treaty represent a fair deal for American Indians?’, ‘how did the discussion of lasers change over time?’

**both**  ‘how does the way Japanese politicians talk about national security change in response to electoral reform?’
We need to think carefully about . . .

1. The appropriate **population** and **sample**.
   
   → “I have so many texts!”, but must remain cognizant of **selection biases** in our sample that influence our ability to draw valid inferences about the population.
   
   → must incorporate **uncertainty** around any estimates of the latent variable derived from a sample.

2. The **features** we care about.
   
   → how to **get at them** with unstructured text, how to **represent** them numerically, and how to **describe** them.

3. How to **model** our data (the texts with their relevant features measured/coded).
   
   → entails **statistical** decisions

4. What we can infer about our **latent variable** of interest.
   
   → **compare** our estimates across people, **validate** them
In general, we will...

Get Texts

An expert hospital consultant has written to my hon. Friend...

Order. The Minister must be allowed to reply without interruption.

I am grateful to my hon. Friend for her question. I pay tribute to her work with the International Myeloma Foundation...

My constituent, Brian Jago, was fortunate enough to receive a course of Velcade, as a result of which he does not have to...

→ Make Document Term Matrix

\[
\begin{pmatrix}
\text{a} & \text{an} & \ldots & \text{ze} \\
MP_{001} & 2 & 0 & \ldots & 1 \\
MP_{002} & 0 & 3 & \ldots & 0 \\
MP_i & \vdots & \vdots & \ldots & \vdots \\
MP_{654} & \vdots & \vdots & \ldots & \vdots \\
MP_{655} & 0 & 0 & \ldots & 2
\end{pmatrix}
\]

→ Model

- (dis)similarity
- diversity
- readability
- scale
- classify
- topic model
- burstiness
- sentiment

→ Make inferences

Amy Catalinac (New York University)
I. From Text to (Numeric) Data
Reducing Complexity

Language is extraordinarily complex, and involves great subtlety and nuanced interpretation.

**but** remarkably, in transforming text to (numeric) data, we can do very well by **simplifying**, and representing documents as straightforward mathematical objects.

→ makes the modeling problem much more **tractable**.

**by** ‘do very well’, we mean that much more complicated representations add (almost) nothing to the quality of our inferences, our ability to predict outcomes, and the fit of our models.

**NB**. Inevitably, the degree to which one simplifies is dependent on the **particular task** at hand.

→ there is no ‘one best way’ to go from texts to numeric data.
→ good idea to check **sensitivity** of one’s inferences to these decisions (see Denny & Spirling, 2018).

2. Strip away ‘superfluous’ material: HTML tags, capitalization, punctuation, stop words etc.

3. Cut document up into useful elementary pieces: **tokenization**.

4. Add descriptive annotations to the tokens that preserve context: **tagging**.

5. Map tokens back to **common form**: lemmatization, stemming.

6. Model.

“PREPROCESSING”

Model.
In some languages, tokenizing is a non-trivial problem because whitespace may not be used:

来年の事を言えば鬼が笑う

We may want to deal directly with multiword expressions in some contexts. There are rules which help us identify them relatively quickly and accurately.

e.g. ‘White House’, ‘traffic light’

NB. these words mean something ‘special’ (and slightly opaque) when combined.
We Don’t Care about Word Order

Generally, we are willing to **ignore the order of the words** in a document. This **simplifies** things. And we do (almost) as well without that information as when we retain it.

→ we are treating a document as a **bag-of-words** (BOW).

**but** we keep multiplicity—i.e. multiple uses of same token

**e.g.** “The leading Republican presidential candidate has said Muslims should be banned from entering the US.”

→ “lead republican presidenti candid said muslim ban enter us”

= “us lead said candid presidenti ban muslim republican enter”

Could use *n*-grams to take order into account, but typically not worth it.
Vector Space Model

A document can be thought of as a **collection of features** (tokens/words, etc.) if each feature can be placed on the real line, then the document can be thought of as a point in space.

**e.g.** “Bob goes home” can be thought of a vector in 3 dimensions: one corresponds to how ‘Bob’-ish it is, one corresponds to how ‘goes’-ish it is, one corresponds to how ‘home’-ish it is.

Features will typically be the *n*-gram (mostly unigram) **frequencies** of the words/tokens in the document, or some **function** of those frequencies.

**e.g.** ‘the cat sat on the mat’ becomes (2,1,1,1,1) if we define the dimensions as (the, cat, sat, on, mat) and use simple counts.

Later, we take the vectors for all of our documents and arrange them as a **term document matrix** or a document term matrix.
II. Supervised Learning
Overview of Supervised Learning

**Goal:** categorize documents (or sentences) as belonging to a certain class (mutually exclusive? Jointly exhaustive?)

1. **Decide** on the taxonomy

2. **Label** examples of documents belonging to each category
   - e.g. some movie reviews that were positive ($y = 1$), some that were negative ($y = 0$); some statements that were liberal, some that were conservative.

3. **Train** a ‘machine’ on these documents (e.g. logistic regression), using the features (DTM, other stuff) as the ‘independent’ variables.
   - e.g. could flag use of word ‘fetus’ relative to ‘baby’ in discussing abortion as indicative of certain categories.

4. Use the learned relationship —some $f(x)$—to **classify** the outcomes of documents ($y \in \{0, 1\}$, review sentiment) not in the training set.
In doing this, we learn **which words are associated with which category**. The model we trained tells us this.

We could use this to build a **dictionary** of words indicative of our categories (e.g. liberal, conservative, positive, negative, etc.).

Scholars have already done this. Provide **pre-determined lists of words**, sometimes with weights, indicating **which category they belong to** and sometimes, the extent to which they belong to that category.

We can rely on their dictionaries to classify our documents, too. Although, presents some problems.
Example: Barnes’ review of *The Big Short*

Director and co-screenwriter Adam McKay (Step Brothers) bungles a great opportunity to savage the architects of the 2008 financial crisis in *The Big Short*, wasting an A-list ensemble cast in the process. Steve Carrel, Brad Pitt, Christian Bale and Ryan Gosling play various tenuously related members of the finance industry, men who made a killing by betting against the housing market, which at that point had superficially swelled to record highs. All of the elements are in place for a lacerating satire, but almost every aesthetic choice in the film is bad, from the U-Turn-era Oliver Stone visuals to Carell’s sketch-comedy performance to the cheeky cutaways where Selena Gomez and Anthony Bourdain explain complex financial concepts. After a brutal opening half, it finally settles into a groove, and there’s a queasy charge in watching a credit-drunk America walking towards that cliff’s edge, but not enough to save the film.
Director and co-screenwriter Adam McKay (Step Brothers) bungles a great opportunity to savage the architects of the 2008 financial crisis in The Big Short, wasting an A-list ensemble cast in the process. Steve Carell, Brad Pitt, Christian Bale and Ryan Gosling play various tenuously related members of the finance industry, men who made made a killing by betting against the housing market, which at that point had superficially swelled to record highs. All of the elements are in place for a lacerating satire, but almost every aesthetic choice in the film is bad, from the U-Turn-era Oliver Stone visuals to Carell’s sketch-comedy performance to the cheeky cutaways where Selena Gomez and Anthony Bourdain explain complex financial concepts. After a brutal opening half, it finally settles into a groove, and there’s a queasy charge in watching a credit-drunk America walking towards that cliff’s edge, but not enough to save the film.
Director and co-screenwriter Adam McKay (Step Brothers) bungles a great opportunity to savage the architects of the 2008 financial crisis in The Big Short, wasting an A-list ensemble cast in the process. Steve Carell, Brad Pitt, Christian Bale and Ryan Gosling play various tenuously related members of the finance industry, men who made a killing by betting against the housing market, which at that point had superficially swelled to record highs. All of the elements are in place for a lacerating satire, but almost every aesthetic choice in the film is bad, from the U-Turn-era Oliver Stone visuals to Carell's sketch-comedy performance to the cheeky cutaways where Selena Gomez and Anthony Bourdain explain complex financial concepts. After a brutal opening half, it finally settles into a groove, and there's a queasy charge in watching a credit-drunk America walking towards that cliff's edge, but not enough to save the film.
Simple math.

negative 11
positive 2

total \( \frac{2 - 11}{13} = \frac{-9}{13} \)
Naive Bayes Classification

Is a way of **building our own dictionary**.

Motivation: emails $d$ arrive and must classified as belonging to one of two classes $c \in \{\text{spam,ham}\}$.

by using the frequencies of the features (words, tokens) the emails contain.

use Naive Bayes, also simple Bayes, or independence Bayes, is a family of classifiers which apply Bayes’s theorem and make ‘naive’ assumptions about independence between the features of a document.

→ fast, simple, accurate, efficient and therefore popular.
Set up

We’re interested in the probability that an email is in a given category, given its features—i.e. frequency of terms.

NB we assume (naively) terms basically occur randomly throughout the document/no position effects

When we see certain terms which are disproportionately prevalent in certain classes of document, the probability we are observing a document of that class increases.

e.g. perhaps 99% of spam emails make reference to dollars, but only 1% of non-spam emails do. If the machine sees dollars it should update that it’s likely a spam email.
Behind bars in Indonesia's version of Alcatraz, radical cleric Abu Bakar Bashir has declared his support for the Islamic State of Iraq and Sham (ISIS). After taking over Iraq's second-largest city Mosul last month, ISIS, a violent Muslim extremist group, declared areas it occupied in Syria and Iraq as an Islamic caliphate.

Nielsen (2012) investigates why certain scholars of Islam become Jihadi: i.e. why they encourage armed struggle (especially against the West).

Requires him to first classify scholars as Jihadi and ¬ Jihadi: has 27,142 texts from 101 clerics, so difficult to do by hand.
Training set: self-identified Jihadi texts (765), and sample from Islamic website as $\neg$ Jihadi (1951)

Preprocess: drops terms occurring in less than 10%, or more than 40% of documents, and uses ‘light’ stemmer for Arabic

Assigns a **Jihad Score** to each document: basically the logged likelihood ratio, $\sum_i \log \frac{\Pr(t_k|\text{Jihad})}{\Pr(t_k|\neg \text{Jihad})}$ (note: doesn’t know what ‘real world’ priors are, so drops them here)

Then for each cleric, he concatenates all works into one and gives this ‘document’/cleric a score.
Discriminating Words

Apostasy

Jihad

Word Frequency
\[ a = 1/250 \]
\[ a = 1/500 \]
\[ a = 1/1000 \]
\[ a = 1/2000 \]

Jihadi

The world

The nation

Stale

Infidel

Monotheism

Mujahideen

Excommunication

Not Jihadi

Narrate by

Age

Women

Her husband

Seffiu

Quran

Hajj

Money

Man

Imam

Human

Amy Catalinac (New York University)
III. Unsupervised Learning
Don’t have to label our data as belonging to categories we come up with, or decide how to recognize those categories.

→ while we know who gave a speech, we don’t yet know what that speech ‘represents’ in terms of its latent properties (what ‘kind’ of speech it is, what ‘topics’ it covers, what speeches it is similar to conceptually, etc.)

Goal is to take the documents and uncover their hidden structure and meaning. By this, we mean that we look for similarities across documents.

→ Because we did not label our data, it is then up to us to interpret why some documents are estimated to be similar to others and not to others (what the groups/dimensions/concepts represent).

• Typically, these techniques put documents on a scale or in a group.
We suspect that the German Greens and Social Democrats have moved steadily rightwards, post-reunification.

→ This is a **time series** problem, but extant techniques struggle.

**i.e.** hand-coding is expensive (manifestos are long) and hard to find reference texts for **Naive Bayes** over time

→ need to assume lexicon is pretty **stable**, and that you can identify texts that contain **all** relevant terms.
Helpful to have an **unsupervised approach**, which is not dependent on reference texts

Suggest **Wordfish** scaling technique ("A Scaling Model for Estimating Time-Series Party Positions from Text")

1. Begin with **naive Bayes assumption**: idea that each word’s occurrence is independent of all other words in the text.

   - surely false, but convenient starting point.

2. Need a (parametric) model for **frequencies** of words.

   - Choose **Poisson**: extremely simple because it has only one parameter—$\lambda$ (which is mean and variance).
The model

\[ \lambda_{ijt} = \exp(\alpha_{it} + \psi_j + \beta_j \times \omega_{it}) \]

\( \alpha_{it} \) fixed effect(s) for party \( i \) in time \( t \): some parties have longer manifestos in certain years (which boosts all counts)

\( \psi_j \) word fixed effect: some parties just use certain words more (e.g. their own name)

\( \beta_j \) word specific weight: importance of this word in discriminating between party positions.

\( \omega_{it} \) estimate of party’s position in a given year (so, this applies to specific party manifesto)

\( \rightarrow \) fit via expectation maximization (details in paper)
Question: theoretical work suggests that candidates adopt different ideological positions in different electoral systems. Little empirical evidence of this.

Empirical strategy: did candidates in elections to Japan’s Lower House adopt different positions after electoral reform in 1994?

Specifically:

H1 Did MMDs → SMDs bring candidates trying to win the district closer together ideologically?

H2 Did intra-party competition → no intra-party competition bring candidates of the same party closer together ideologically?
Likely to be representative of a candidate’s campaign strategy. Likely to be taken seriously. 40% of voters say they saw it (surveys, 1972-2005).


Used Wordfish to locate each manifesto on a unidimensional scale.
In all elections, JCP is to the **left** of all other parties and LDP is to the **right** of the JSP, SDP, NSP, and DPJ. In 1986, LDP and NLC positions are **same**.
Validation with ATES Data

→ 2003 and 2005: candidates asked to locate themselves on ideological scale (response rates: 95% and 91%). Correlate highly with estimates from Wordfish.
Results: Hypothesis 1

Predicted values of **within-district dispersion** with 95% confidence intervals. **Lower after reform.**
Results: Hypothesis 2

Predicted values of within-party dispersion with 95% confidence intervals. Lower after reform.
Okayama 1st, 2nd, 3rd Districts, 1996 (m=1 each)

Left

Aisawa Ichiro (LDP)

Kumashiro Akihiko (LDP)

Hiranuma Takeo (LDP)

Right

0.95
0.86
0.98
Topic Models

Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.

Blei, 2012

NB. in social science we often use the outputs from topic models as a measurement strategy:

“who pays more attention to education policy, conservatives or liberals?”
Clustering

Document 1

Document 2

Document 3

\vdots

Document $N$

Cluster 1

Cluster 2

\vdots

Cluster $K$
Clustering

Document 1
Document 2
Document 3
...
Document $N$

Cluster 1
Cluster 2
...
Cluster $K$
Clustering

Document 1
Document 2
Document 3
\vdots
Document $N$

\rightarrow

Cluster 1
Cluster 2
\vdots
Cluster $K$
Clustering

Document 1
Document 2
Document 3
...
Document \( N \)
Clustering

Document 1
Document 2
Document 3

...  Document N

Cluster 1
Cluster 2
...
Cluster K
Topic Modeling

Document 1
Document 2
Document 3
:
Document $N$

Topic 1
Topic 2
:
Topic $K$
Topic Modeling

Document 1

Document 2

Document 3

\vdots

Document N

\Rightarrow

Topic 1

Topic 2

\vdots

Topic K
Documents exhibit same topics, but in different proportions.

E.g. a speech by the Finance minister might be 50% drawn from the trade topic, 40% from the spending topic, 9.9% from the taxation topic, 0.1% from the health topic.

Think of a topic as a distribution over a fixed vocabulary.

E.g. the trade topic will have words like import and tariff with a high probability.

Technically we assume the topics are generated first, and the documents are generated second (from those topics).
Japan is a curious case for IR: wealthy post-war yet not very interested in foreign policy. Recent times have seen more interest in this area. Why?

1. Rise of China? Need to focus on security.
2. Change in Electoral System? Moved from promising pork to having to deliver policy as part of a Westminster-style political system.

To decide, need a data source that covers all Lower House legislators where they set out their policy priorities over time. See if/when they shift priorities.
Manifestos


- Poor quality of paper meant **difficulties extracting digitized text**. OCR didn’t work → transcribed by hand.

- Used MeCab implemented in R to create TDM → ran LDA.
<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>改革</td>
<td>年金</td>
<td>推進</td>
<td>区</td>
<td>政治</td>
<td>日本</td>
</tr>
<tr>
<td>郵政</td>
<td>円</td>
<td>整備</td>
<td>政策</td>
<td>改革</td>
<td>改革</td>
</tr>
<tr>
<td>民営</td>
<td>廃止</td>
<td>団る</td>
<td>地域</td>
<td>国民</td>
<td>外交</td>
</tr>
<tr>
<td>小泉</td>
<td>改革</td>
<td>つとめる</td>
<td>まち</td>
<td>企業</td>
<td>国家</td>
</tr>
<tr>
<td>構造</td>
<td>兆</td>
<td>社会</td>
<td>鹿児島</td>
<td>自民党</td>
<td>社会</td>
</tr>
<tr>
<td>政府</td>
<td>実現</td>
<td>対策</td>
<td>全力</td>
<td>日本</td>
<td>国民</td>
</tr>
<tr>
<td>官</td>
<td>無駄</td>
<td>振興</td>
<td>選挙</td>
<td>共産党</td>
<td>保障</td>
</tr>
<tr>
<td>推進</td>
<td>日本</td>
<td>充実</td>
<td>国政</td>
<td>献金</td>
<td>経済</td>
</tr>
<tr>
<td>民</td>
<td>増税</td>
<td>促進</td>
<td>作り</td>
<td>金権</td>
<td>地域</td>
</tr>
<tr>
<td>自民党</td>
<td>削減</td>
<td>安定</td>
<td>横浜</td>
<td>党</td>
<td>拉致</td>
</tr>
<tr>
<td>日本</td>
<td>一元化</td>
<td>確立</td>
<td>対策</td>
<td>禁止</td>
<td>守る</td>
</tr>
<tr>
<td>制度</td>
<td>政権</td>
<td>企業</td>
<td>中小</td>
<td>憲法</td>
<td>問題</td>
</tr>
<tr>
<td>民間</td>
<td>子供</td>
<td>実現</td>
<td>発電</td>
<td>腐敗</td>
<td>北朝鮮</td>
</tr>
<tr>
<td>年金</td>
<td>地域</td>
<td>小中</td>
<td>推進</td>
<td>団体</td>
<td>教育</td>
</tr>
<tr>
<td>実現</td>
<td>ひと</td>
<td>育成</td>
<td>エネルギー</td>
<td>企業</td>
<td>責任</td>
</tr>
<tr>
<td>進める</td>
<td>サラリーマン</td>
<td>制度</td>
<td>声</td>
<td>区</td>
<td>力</td>
</tr>
<tr>
<td>断行</td>
<td>制度</td>
<td>政治</td>
<td>実現</td>
<td>ソ連</td>
<td>創る</td>
</tr>
<tr>
<td>地方</td>
<td>議員</td>
<td>地域</td>
<td>活性</td>
<td>守る</td>
<td>安心</td>
</tr>
<tr>
<td>止める</td>
<td>金</td>
<td>福祉</td>
<td>自民党</td>
<td>平和</td>
<td>目指す</td>
</tr>
<tr>
<td>保障</td>
<td>民主党</td>
<td>事業</td>
<td>地方</td>
<td>反対</td>
<td>限り</td>
</tr>
<tr>
<td>財政</td>
<td>年間</td>
<td>改革</td>
<td>尽くす</td>
<td>真</td>
<td>憲法</td>
</tr>
<tr>
<td>作る</td>
<td>一掃</td>
<td>確保</td>
<td>商店</td>
<td>是正</td>
<td>可能</td>
</tr>
<tr>
<td>賛成</td>
<td>郵政</td>
<td>強化</td>
<td>いかす</td>
<td>一掃</td>
<td>道</td>
</tr>
<tr>
<td>社会</td>
<td>道路</td>
<td>教育</td>
<td>全国</td>
<td>政党</td>
<td>恵政</td>
</tr>
<tr>
<td>国民</td>
<td>交代</td>
<td>施設</td>
<td>政党</td>
<td>株本</td>
<td>未来</td>
</tr>
<tr>
<td>公務員</td>
<td>社会保険</td>
<td>生活</td>
<td>支援</td>
<td>支援</td>
<td>ひと</td>
</tr>
<tr>
<td>力</td>
<td>月額</td>
<td>支援</td>
<td>支援</td>
<td>支援</td>
<td>再生</td>
</tr>
<tr>
<td>経済</td>
<td>手当</td>
<td>環境</td>
<td>経済</td>
<td>経済</td>
<td>将来</td>
</tr>
<tr>
<td>国</td>
<td>談合</td>
<td>発展</td>
<td>福祉</td>
<td>改革</td>
<td>解決</td>
</tr>
<tr>
<td>安心</td>
<td>住居</td>
<td>締結</td>
<td>福祉</td>
<td>改革</td>
<td>耳士</td>
</tr>
</tbody>
</table>
Change in proportion of ‘Pork’ and ‘Policy’

Proportion of Discussion


Pork

Policy

Amy Catalinac (New York University)
Change in proportion of ‘National Security’ Topic

First Election Under MMM
IV. Next Steps in Text as Data
Next Steps

- Moving beyond unigram representations: embeddings look promising.

- Thinking carefully about how to handle non-Latin scripts and their encodings.

- Working on social science specific training of models (e.g. for sentiment dictionaries and ‘sophistication’ measures).

- Bringing more supervision to unsupervised problems, like (structural) topic models.
Thanks!

https://scholar.harvard.edu/amycatalinac/home

amy.catalinac@nyu.edu