Text as Data

Gov 51: Data Analysis and Politics

Scott Cunningham

Harvard University

Week 3, Tuesday
February 11, 2026

Why Treat Text as Data?

How do you measure what politicians believe?

Political speech is the richest record of political thought

The data is everywhere:

- Congressional Record: every floor speech since 1873
- Presidential papers: addresses, statements, executive orders
- State legislatures, court opinions, party platforms
- Social media, press releases, campaign materials

17 million congressional speeches alone—waiting to be analyzed

But text is unstructured—computers struggle with meaning

Some data is easy:

- Numbers: voter turnout = 0.65
- Categories: party = Democrat

Text is hard:

> *"Give me your tired, your poor, your huddled masses yearning to breathe free..."*

Is this pro-immigration, anti-immigration, or neutral?

The challenge: convert meaning into measurement

Three ways to turn text into data

Human Coding

Read and label manually

Gold standard
but expensive

Dictionary Methods

Count predefined words

Fast
but inflexible

Machine Learning

Train algorithms to classify

Scalable
but requires training data

Today: How machine learning revolutionized text analysis

What we'll cover:

1. The history of NLP (origins, key players)
2. A famous case study: Who wrote the Federalist Papers?
3. Modern applications in political science
4. Deep dive: 140 years of immigration rhetoric

Connects to: replication exercise and problem set

A Brief History of Text Analysis

NLP began as a Cold War project to translate Russian

1950s: US government funded machine translation research

- Goal: Automatically translate Soviet scientific papers
- Early optimism: "Five years, maybe three" (1954 prediction)
- Reality: Took 60+ years to get good translation

Apocryphal story:

"The spirit is willing but the flesh is weak"

→ *"The vodka is good but the meat is rotten"*

Two tribes developed text analysis: linguists and statisticians

Linguists (Chomsky)

- Rules-based approaches
- Grammar, syntax, parsing
- “Understand language structure”

Focused on *how* language works

Statisticians (Shannon)

- Probability-based approaches
- Word frequencies, patterns
- “Predict the next word”

Focused on *what* language does

Modern LLMs descend from the statistical tradition

Three breakthroughs enabled modern text analysis

1. **Word embeddings** (2013)
   Words as vectors in space (Word2Vec)
2. **Attention/Transformers** (2017)
   "Attention is all you need" (Google)
3. **Scale** (2018+)
   Billions of parameters, trained on the internet (BERT, GPT)

Result: Machines now "understand" text well enough to classify, summarize, translate

Social scientists now have powerful tools—but they require care

Software availability:

- Python: NLTK, spaCy, transformers
- R: tidytext, quanteda
- Pre-trained models: Can classify text without building from scratch

But:

- Validation against human judgment is essential
- Key question: Does the algorithm capture what humans mean?

The tool is only as good as your validation.

Case Study: The Federalist Papers

Who wrote the disputed Federalist Papers?

The Federalist Papers shaped American democracy—but authorship was disputed

The setup:

- 85 essays published 1787–1788 under pseudonym “Publius”
- Goal: Persuade New York to ratify the Constitution
- Authors: Alexander Hamilton, James Madison, John Jay

The problem:

- 51 known Hamilton, 14 known Madison, 5 known Jay
- **12 disputed**: Both Hamilton and Madison claimed them
- Hamilton died in 1804 duel; Madison lived until 1836

Mosteller and Wallace (1963) used statistics to solve a 175-year mystery

The researchers:

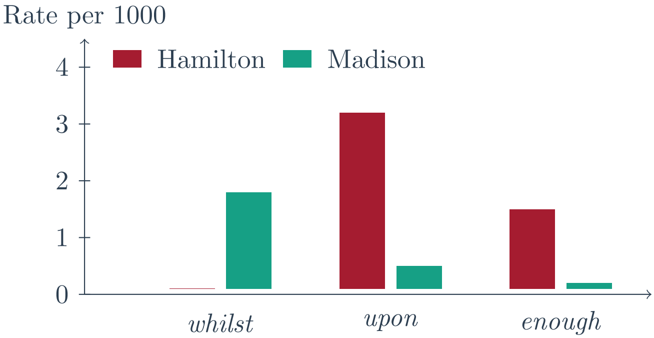
- Frederick Mosteller (Harvard Statistics)
- David Wallace (University of Chicago)

The method: Count "function words"

- Words like: *upon, whilst, by, to, enough*
- Not content words (which vary by topic)
- Function words are unconscious stylistic fingerprints

Published: *Inference and Disputed Authorship: The Federalist*

Hamilton and Madison had different word usage patterns



Key insight: Authors have unconscious stylistic fingerprints

All 12 disputed papers were written by Madison

Bayesian analysis: Probability ratios overwhelming

- Federalist 18–20: Odds ratio > 1000:1 for Madison
- Federalist 51 (“Ambition must be made to counteract ambition”): **Madison**
- Federalist 63: **Madison**

Statistical analysis resolved a question historians couldn’t

This is now the consensus among historians.

The Federalist study established key principles still used today

1. **Validation**
   Test method on known cases before applying to unknowns
2. **Feature selection**
   Choose features that distinguish authors, not topics
3. **Probabilistic reasoning**
   Express uncertainty, don’t claim certainty
4. **Replication**
   Make methods transparent for others to verify

Good text analysis is careful text analysis

Text as Data in Political Science Today

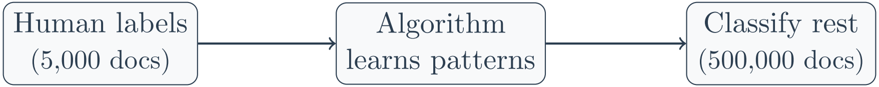
Researchers now use text analysis to study political behavior at scale

Examples:

- **Ideology measurement:** Scaling politicians from speeches
  Gentzkow, Shapiro & Taddy (2019)
- **Media bias:** Comparing news coverage across outlets
  Groseclose & Milyo (2005)
- **Public opinion:** Analyzing social media sentiment
  Barber et al. (2015)
- **International relations:** Measuring threats in UN speeches
  Baturo et al. (2017)

Classification is the workhorse method

Supervised learning: Train on labeled examples, apply to unlabeled


Human labels (5,000 docs) → Algorithm learns patterns → Classify rest (500,000 docs)

Trade-off: Human accuracy vs. machine scalability

Validation is non-negotiable

Inter-annotator agreement: Do humans agree with each other?

- ▷ Common metric: Krippendorff's α ($0 =$ random, $1 =$ perfect)
- ▷ Rule of thumb: $\alpha > 0.67$ is acceptable

Classifier accuracy: Does the machine agree with humans?

- ▷ Hold out test set
- ▷ Metrics: accuracy, precision, recall, F1

Warning: If humans disagree, the task may be inherently ambiguous

Deep Dive: Measuring Immigration Rhetoric

How do Americans talk about immigrants?
And how has it changed over 140 years?

Card et al. analyzed 305,000 political speeches spanning 1880–2020

Data sources:

Source	Records	Time Period
Congressional speeches	290,800	1880–2020
Presidential communications	14,195	1880–2021
Total	**304,995**	**140 years**

Span: Chinese Exclusion Act (1882) to present

Card, Boustan, Abramitzky et al. (PNAS 2022)

Human annotation created the training data

Annotation setup:

- 5 Princeton annotators
- 7,626 speech segments labeled

Two classification tasks:

1. **Relevance:** Is this about immigration? (binary)
2. **Tone:** Pro-immigration, Anti-immigration, Neutral (ternary)

Inter-annotator agreement:

- Relevance: $\alpha = 0.76$ (good)
- Tone: $\alpha = 0.48$ (moderate—tone is subjective!)

RoBERTa learned to classify from human examples

RoBERTa: A variant of BERT (Bidirectional Encoder Representations from Transformers)

Training process:

1. Start with pre-trained language model (trained on internet text)
2. Fine-tune on congressional speeches (domain adaptation)
3. Fine-tune as classifier using annotated examples

Performance:

- ▷ Relevance: ∼90% accuracy
- ▷ Tone: ∼65% accuracy

65% sounds low, but humans only agree at $\alpha = 0.48$

The output: probability scores for each speech

For each speech segment, the model outputs:

- P(Pro-immigration)
- P(Anti-immigration)
- P(Neutral)

Aggregate measure:

$$\text{Average tone} = \%\ \text{Pro} - \%\ \text{Anti}$$

Range: -100 (all anti) to $+100$ (all pro)

Can compute by: year, party, speaker, topic

What Did They Find?

Overall sentiment toward immigration is more positive today than a century ago

Surprising! Given current rhetoric, you might expect the opposite.

Three eras:

1. **1880–1940:** Consistently negative (quota era)
2. **1940–1965:** Dramatic shift toward positive (WWII to Immigration Act)
3. **1965–present:** Net positive on average

But this masks important variation...

Immigration rhetoric: Overall trend (1880–2020)

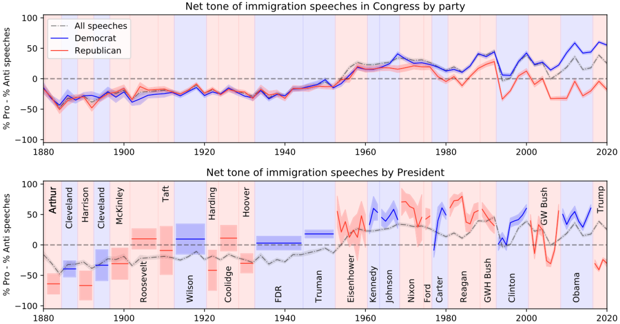


Fig. 1. Evolution of attitudes toward immigration expressed in congressional speeches and presidential communications. Average tone is computed as the percentage of proimmigration speeches minus the percentage of antiimmigration speeches, where proimmigration means valuing immigrants and favoring less restricted immigration and vice versa. *Top* and *Bottom* show the overall tone using all congressional speeches about immigration (black dashed line, with bands showing plus or minus two SDs based on the estimated proportions and number of speeches). *Top* also shows separate plots for speeches by Democrats and Republicans in Congress. (Due to limitations of the data, about 15% of speeches do not have a named speaker or party affiliation.) *Bottom* shows the corresponding estimates for each president, showing the overall average for a president's tenure when there are insufficient data to show annual variation. Note that most modern presidents have been more favorable toward immigration than the average member of Congress. By contrast, Donald Trump appears to be the most antiimmigration president in nearly a century. Similarly, congressional Republicans over the past decade have framed immigration approximately as negatively as the average member of Congress did a century earlier.

Average tone = % Pro − % Anti. Source: Card et al. (PNAS 2022)

But the parties have polarized dramatically since the 1970s

Through the 1970s: Both parties roughly similar on immigration

Then divergence:

- Democrats: increasingly positive
- Republicans: increasingly negative

Today:

- Democrats: unprecedentedly positive
- Republicans: as negative as the 1920s quota era

Partisan polarization on immigration is at historic levels

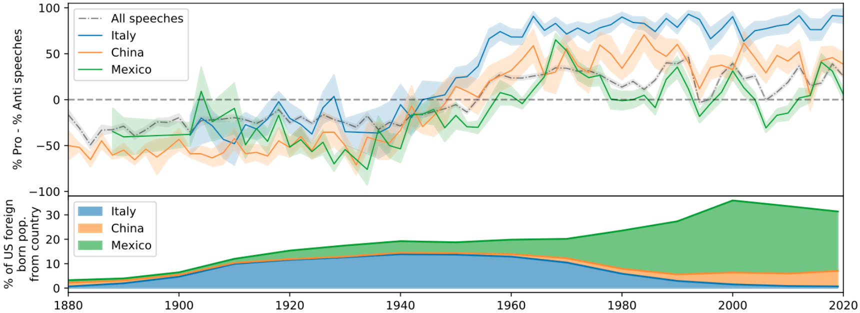


Fig. 2. Average tone of immigration speeches when considering only those speeches that mention the country or nationality for each of the three most frequently mentioned nationalities (*Top*) and the percent of the US foreign-born population from each of these countries over time (*Bottom*). Despite the midcentury increase in proimmigration attitudes applying to all groups, a gap in tone by group persists to the present day, with Mexican immigrants being consistently framed more negatively than others and Italian immigrants being framed especially positively. These trends are mirrored in broader regional patterns for Europe, Asia, and Latin American and the Caribbean (*SI Appendix*).

Average tone by party over time. Source: Card et al. (PNAS 2022)

Trump is the most anti-immigration president in 140 years

A historical first:

- First modern president more anti-immigration than average member of own party
- Presidential communications consistently more negative than Republican Congress
- Broke from historical pattern where presidents were moderating voices

Presidential rhetoric: Trump broke from historical patterns

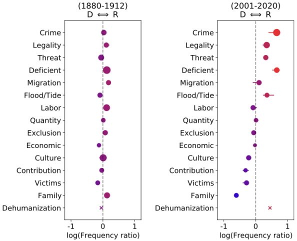


Fig. 3. Relative usage frequency for each of 14 frames by Republicans compared to Democrats, both for the late 19th/early 20th century (*Left*) and the past 2 decades (*Right*). Farther to the left on each plot represents more frequent usage by Democrats and vice versa (plotted as log frequency ratio). Circle size represents the overall prominence of the frame in speeches about immigration, relative to all speeches. To ensure the robustness of these findings, we leave out each word in turn from each frame and show the full range of possible values obtained using horizontal lines (not visible when the full range is contained within the circle). “Dehumanization” is an aggregation of metaphorical categories (see *Measuring Dehumanization*). Compared to the absence of polarization a century ago, certain frames today are disproportionately used by Republicans (“crime,” “legality,” “threats,” “deficiency,” and “flood/tide”) and Democrats (“family,” “victims,” “contributions,” and “culture”). Republicans also show significantly higher use of implicit dehumanizing metaphors like “animals” and “cargo.”

Presidential tone relative to Congress. Source: Card et al. (PNAS 2022)

Mexican immigrants are framed most negatively of any group

Rhetoric varies by nationality:

- **Chinese:** Overwhelmingly negative during exclusion era (1882–1943)
- **Italian:** Now overwhelmingly positive (was negative in early 1900s)
- **Mexican:** Persistently most negative framing

Key finding: Gap between Mexican and Italian mentions ≈ partisan gap

Rhetoric varies dramatically by immigrant nationality

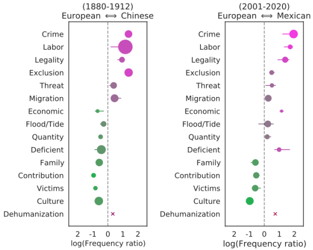


Fig. 4. Relative usage frequency for each of 14 frames in speeches mentioning Chinese vs. European immigrants in the late 19th/early 20th century (*Left*) and those mentioning Mexican vs. European immigrants in the 21st century (*Right*). Farther to the left on each plot represents greater usage in speeches mentioning European groups. Circle size represents the overall frequency of the frame in the relevant speeches relative to all speeches. Horizontal lines show the minimum and maximum values of the log ratio obtained when leaving out each term in the corresponding lexicon in turn. “Dehumanization” is an aggregation of the six metaphorical categories. There is a strong correlation between how Mexican immigrants are framed today and how Chinese immigrants were framed a century earlier, relative to European immigrants of the corresponding time period, in terms of both the explicit frames emphasized and a significantly higher usage of dehumanizing metaphors for mentions of the non-European groups.

Tone by immigrant nationality mentioned. Source: Card et al. (PNAS 2022)

Republicans use more dehumanizing language

The paper measured 6 metaphorical categories:

- Animals
- Cargo
- Disease
- Flood/Tide
- Machines
- Vermin

Method: BERT masked language model to detect metaphors

Finding: Republicans show **1.6× higher probability** of dehumanizing language

Difference is statistically significant ($p < 0.001$)

Connecting to Your Work

You will replicate part of this analysis using modern LLMs

What you'll do:

- **In-class exercise:** Use GPT-4o-mini to classify sample speeches
- **Problem set:** Compare LLM classifications to original RoBERTa predictions

Research question: Can a zero-shot LLM replicate a fine-tuned classifier?

Cost: ∼$11 to classify all 305,000 speeches

This is the future of text analysis—accessible and cheap

Key Takeaways

What We Learned

1. **Text is data:** Political speech can be measured and analyzed
2. **Classification is the workhorse:** Train on human labels, apply at scale
3. **Validation is essential:** Always check against human judgment
4. **Immigration rhetoric:** More positive overall, but sharply polarized by party
5. **Methods matter:** Good research design + good data = credible answers

Text as data transforms unstructured language into measurable evidence about political behavior

What questions do you have?

Discussion prompts:

- Does the 65% accuracy concern you? Why or why not?
- What other political questions could text analysis address?
- How might you validate these findings independently?

Reading: Card et al. (PNAS 2022)—available on course website