Text as Data: Text Regressions

Guest Course – January 2026

Germain Gauthier, Philine Widmer[1]

[1]Bocconi Unversity, Paris School of Economics

USI Lugano

Warm-up: dictionary methods (what we already know)

A dictionary score is typically:

$$s_i = \sum_{j \in \mathcal{V}} a_j \, x_{ij} \, ,$$

where

- x_{ij} = count (or tf-idf) of token j in document i
- a_j = *hand-chosen* weights (e.g., sentiment lexicon)

What changes with text regressions? We now estimate a_j from labeled data to optimize prediction.

Supervised learning setup

We observe documents $i = 1, \ldots, n$:

- Text → features $x_i \in \mathbb{R}^p$
- Target y_i (continuous outcome or class label)

Goal: learn a function $f(\cdot)$ such that $\hat{y}_i = f(x_i)$ generalizes out-of-sample.

Text → feature matrix X

Common choices:

- Bag-of-words (counts)
- tf-idf weighting
- n-grams (bigrams/trigrams) for short phrases

Practical notes:

- p can be huge (10^4–10^6), sparse matrices are essential
- Keep feature construction consistent across training/validation/test

Why penalization? ($p \gg n$ and multicollinearity)

In text, we often have:

- far more features than observations ($p \gg n$)
- correlated predictors (synonyms, topics, stylistic clusters)

Penalization controls overfitting and (sometimes) yields sparsity/interpretability.

Ridge, Lasso, Elastic Net (linear regression)

For squared loss:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 + \lambda \mathcal{P}(\beta) \right\}.$$

Penalties:

- Ridge: $\mathcal{P}(\beta) = \|\beta\|_2^2$ (shrinkage, no sparsity)
- Lasso: $\mathcal{P}(\beta) = \|\beta\|_1$ (sparsity / feature selection)
- Elastic Net: $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2$

Ridge, Lasso, Elastic Net (logistic regression)

Binary label $y_i \in \{0, 1\}$:

$$\Pr(y_i = 1 \mid x_i) = \sigma(x_i^\top \beta), \quad \sigma(t) = \frac{1}{1 + e^{-t}}.$$

Log-likelihood contribution:

$$\ell(y_i, x_i^\top \beta) = y_i \log \sigma(x_i^\top \beta) + (1 - y_i) \log\Big(1 - \sigma(x_i^\top \beta)\Big).$$

Estimate:

$$\hat{\beta} = \arg\min_{\beta} \Big\{ -\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i^\top \beta) + \lambda \mathcal{P}(\beta) \Big\}.$$

Application – Speech Polarization in the U.K.

- Predict party from speeches (supervised text classification) and then use out-of-sample classification accuracy as a substantive measure of polarization[1]

  → *"If parties are easy to tell apart from language, polarization is high."*

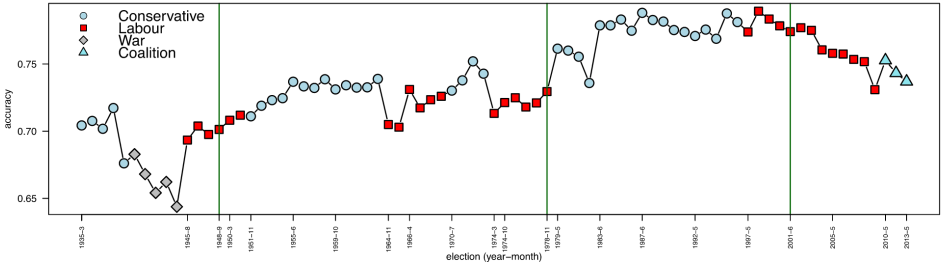


Figure 3. Estimates of parliamentary polarization, by session. Election dates mark *x*-axis. Estimated change points are [green] vertical lines.

Gentzkow, Shapiro & Taddy (2019)[3]

- Question: how different are two groups' language choices when the vocabulary is huge?
- Setting: two parties (Republicans and Democrats) choose among many possible phrases for their speeches (J very large).
- Core idea: measure partisanship as expected posterior accuracy — i.e., how well an observer can infer a speaker's party from one phrase?
- Their method is now used in various papers in political economy[2].

Data and multinomial model of speech

For speaker i in session t:

- counts $c_{it} = (c_{i1t}, \ldots, c_{iJt})$
- total phrases $m_{it} = \sum_j c_{ijt}$
- party $P(i) \in \{R, D\}$, covariates x_{it}

Model:

$$c_{it} \sim \mathrm{MN}\Big(m_{it},\ q_t^{P(i)}(x_{it})\Big) .$$

Multinomial logit parameterization

Choice probabilities:

$$q_{jt}^{P(i)}(x_{it}) = \frac{\exp(u_{ijt})}{\sum_{\ell=1}^{J} \exp(u_{i\ell t})}, \qquad u_{ijt} = \alpha_{jt} + x_{it}^{\top}\gamma_{jt} + \phi_{jt}\,\mathbf{1}\{i \in R_t\}.$$

Interpretation:

- α_{jt}: baseline popularity of phrase j in session t
- γ_{jt}: how covariates shift phrase use
- ϕ_{jt}: party loading (the key “partisan” parameter)

Partisanship as expected posterior accuracy (definition)

Define posterior belief after hearing phrase j:

$$\rho_{jt}(x) = \frac{q_{jt}^R(x)}{q_{jt}^R(x) + q_{jt}^D(x)}.$$

Partisanship at x:

$$\pi_t(x) = \frac{1}{2}\, q_t^R(x) \cdot \rho_t(x) + \frac{1}{2}\, q_t^D(x) \cdot (1 - \rho_t(x)).$$

Average partisanship in session t:

$$\pi_t = \frac{1}{|R_t \cup D_t|} \sum_{i \in R_t \cup D_t} \pi_t(x_{it}).$$

How to interpret π_t

- **Intuition:**
  - Pick a random speaker's party with probability $1/2$ each, then pick a phrase from that party's phrase distribution, then guess the party from the phrase.
- $\pi_t \in [1/2, 1]$
- $\pi_t = 1/2$: phrase choice gives no information about party
- larger π_t: an observer can infer party more accurately from a short utterance

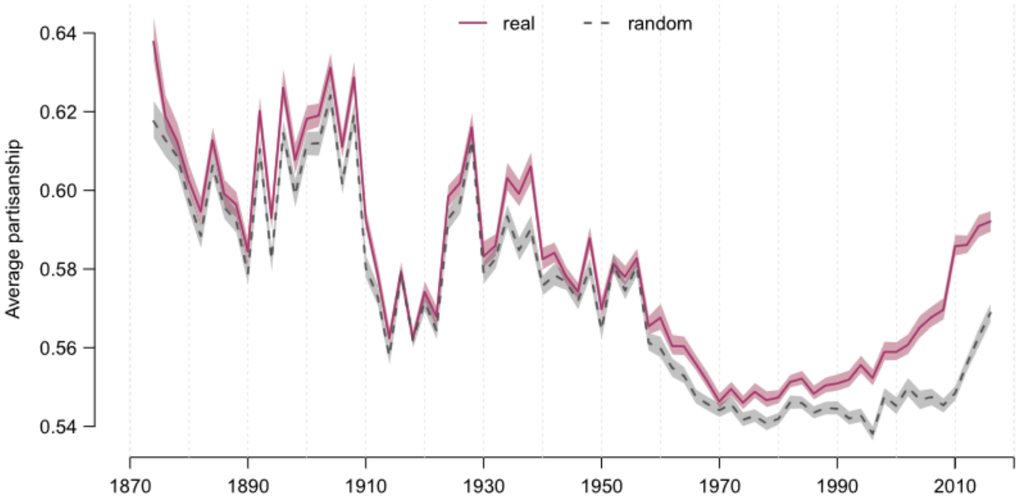
Naive MLE plug-in and why it is biased

Let $\hat{q}_t^R, \hat{q}_t^D$ be empirical phrase frequencies and $\hat{\rho}_t$ the empirical posterior. The plug-in MLE for partisanship is:

$$\hat{\pi}_t^{MLE} = \frac{1}{2}(\hat{q}_t^R) \cdot \hat{\rho}_t + \frac{1}{2}(\hat{q}_t^D) \cdot (1 - \hat{\rho}_t).$$

Intuition for bias: with many phrases, some will look “party-exclusive” by chance, inflating dispersion of $\hat{\rho}_{jt}$.

Panel A: Partisanship from Maximum Likelihood Estimator ($\hat{\pi}_t^{MLE}$)



Penalized estimator: estimating the full model

They estimate $\{\alpha_t, \gamma_t, \phi_t\}$ by minimizing (paper notation):

$$\sum_j \Bigg\{ \sum_t \sum_i \Big[m_{it} \exp(\alpha_{jt} + x_{it}^\top \gamma_{jt} + \phi_{jt} \mathbf{1}\{i \in R_t\}) - c_{ijt}(\alpha_{jt} + x_{it}^\top \gamma_{jt} + \phi_{jt} \mathbf{1}\{i \in R_t\}) \Big]$$

$$+ \psi\Big(|\alpha_{jt}| + \|\gamma_{jt}\|_1\Big) + \lambda_j |\phi_{jt}| \Bigg\}.$$

Then compute $\hat{\pi}_t^*$ by plugging parameter estimates into π_t.

Nitty gritty details: Poisson trick + L1 shrinkage

Two key moves:

- **Poisson approximation:** treat $c_{ijt} \sim \mathrm{Pois}(\exp[\mu_{it} + u_{ijt}])$ with plug-in $\hat{\mu}_{it} = \log m_{it}$, which makes the objective tractable.
- **L1 penalty on ϕ_{jt}:** shrinks many party loadings toward zero, limiting dispersion in $\hat{\rho}_{jt}$ and reducing finite-sample bias.

Panel B: Partisanship from Preferred Penalized Estimator ($\hat{\pi}_t^*$)

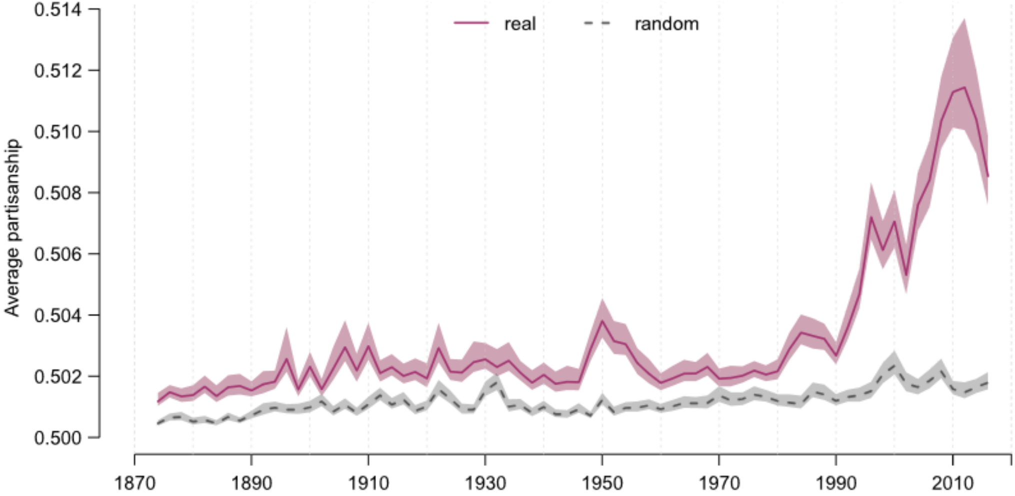
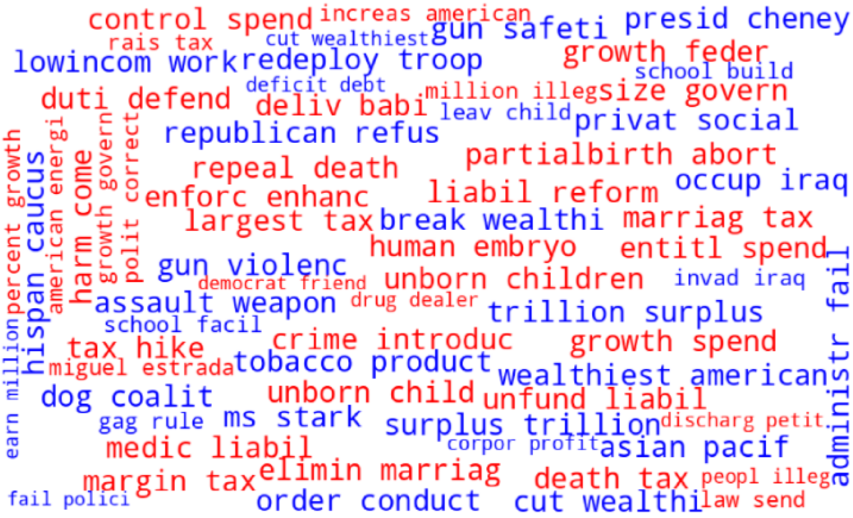


Figure 1: Democrat (Blue) and Republican (Red) Bigrams, 2005-2008



Notes: The word cloud shows the most partisan bigrams for Democrats (blue) and Republicans (red) across newspapers during the period 2005–2008. We restrict the computation to the top 1,000 most partisan bigrams. Font size represents the relative partisanship of each bigram, with larger text indicating greater partisanship. Procedural bigrams have been filtered out.

Why go beyond linear models?

Linear text regressions assume:

$$\hat{y} = g(x^\top \beta)$$

so effects are additive in features.

But language phenomena include:

- interactions (“not good”)
- non-linear intensity (diminishing returns of repeated words)
- more complex composition with richer representations

Feed-forward neural network (MLP) on text features

Simple MLP:

$$h = \phi(Wx + b), \qquad \hat{y} = g(Vh + c)$$

- x can be tf-idf, counts, or embeddings
- $\phi =$ ReLU/tanh; $g =$ identity (regression) or sigmoid/softmax (classification)

Interpretation: a flexible function of the same “bag-of-features” input.

Regularization and training (NNs)

What prevents overfitting:

- early stopping on validation loss
- weight decay (L2)
- dropout

When MLPs shine:

- lots of labeled data
- signal depends on feature interactions or non-linearities

When linear models often win:

- small-to-medium labeled data, need interpretability, stable estimates

A pragmatic modeling ladder

1. Start with a linear baseline (ridge/logistic)
2. Add sparsity (lasso/elastic net) for interpretability
3. Try non-linear models (MLP) if baseline saturates
4. Always evaluate the same way (held-out test; avoid tuning on test)

The typical pipeline

- Define target y (what exactly is the label?)
- Create annotation protocol + train annotators
- Split data: train / validation / test (or CV)
- Train models + tune hyperparameters on validation
- Final evaluation on test set (once)

Human annotation: what to emphasize

Key choices:

- sampling (representative? balanced classes?)
- label quality: inter-annotator agreement, adjudication rules
- unit of annotation (sentence? speech? document?)

Remember: model performance is capped by label noise and ambiguity.

Metrics for classification (1/2): confusion matrix intuition

Confusion matrix terms (for the “positive” class):

- TP: true positives (predicted positive and truly positive)
- FP: false positives (predicted positive but truly negative)

  → *false alarm*
- FN: false negatives (predicted negative but truly positive)

  → *missed case*
- TN: true negatives (predicted negative and truly negative)

Which errors matter depends on the application:

- costly false alarms → care about FP (precision)
- costly misses → care about FN (recall)

Metrics for classification (2/2): precision, recall, F1

$$\text{Precision} = \frac{TP}{TP + FP}$$

→ When we predict *positive*, how often are we correct?

$$\text{Recall} = \frac{TP}{TP + FN}$$

→ Of all *true positives*, how many did we find?

$$F1 = \frac{2PR}{P + R}$$

→ Single score that balances precision and recall (harmonic mean).

Takeaways

- Dictionary methods $\subset$ supervised text regression (fixed vs. learned weights)
- Penalization makes high-dimensional text regression feasible and often interpretable
- Neural nets relax linearity but need more data $+$ careful regularization
- Validation (train-test splits, labels, metrics) is essential

Why unsupervised learning next?

Supervised learning needs labels:

- great for prediction and measurement once target is defined
- limited for discovery: "what are the themes in this corpus?"

Topic models aim to uncover latent structure:

- documents as mixtures of topics
- topics as distributions over words

Next time: LDA-style models, interpretation, and how topics can feed into regression.

References I

- Cagé, Julia, Caroline Le Pennec, and Elisa Mougin (2024). "Firm donations and political rhetoric: Evidence from a national ban". In: *American Economic Journal: Economic Policy* 16.3, pp. 217–256.
- Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy (2019). "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech". In: *Econometrica* 87.4, pp. 1307–1340. DOI: 10.3982/ECTA16566.
- Peterson, Andrew and Arthur Spirling (2018). "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems". In: *Political Analysis* 26.1, pp. 120–128. DOI: 10.1017/pan.2017.39.
- Widmer, Philine, Sergio Galletta, and Elliott Ash (2022). "Media slant is contagious". In: *arXiv preprint arXiv:2202.07269*.