UNIVERSITÄT
LEIPZIG

Forschungsseminar CSS – Supervised ML

GWZ H2.1.15, 20.11.2025

Felix Lennert, M.Sc.

OUTLINE

- Intro
- Supervised ML
  - Motivation
  - “Text Regression”
  - The Procedure
- Unsupervised ML; Topic Modeling
  - Value for Social Sciences
  - LDA
  - In Practice
  - Evaluation Strategy

BEFORE WE START

- Today is going to be about the logic behind and the steps you researchers have to do when using supervised text classification
- Caveat:
  - models based on the bag of words-assumption (which we will use today) are becoming increasingly outdated
  - new models are there and incredible, but they remain black boxes we cannot open
  - yet they are fairly user-friendly, about 5 lines of code (and a lot of waiting time depending on your computer)
  - and: the training and evaluation process is basically the same

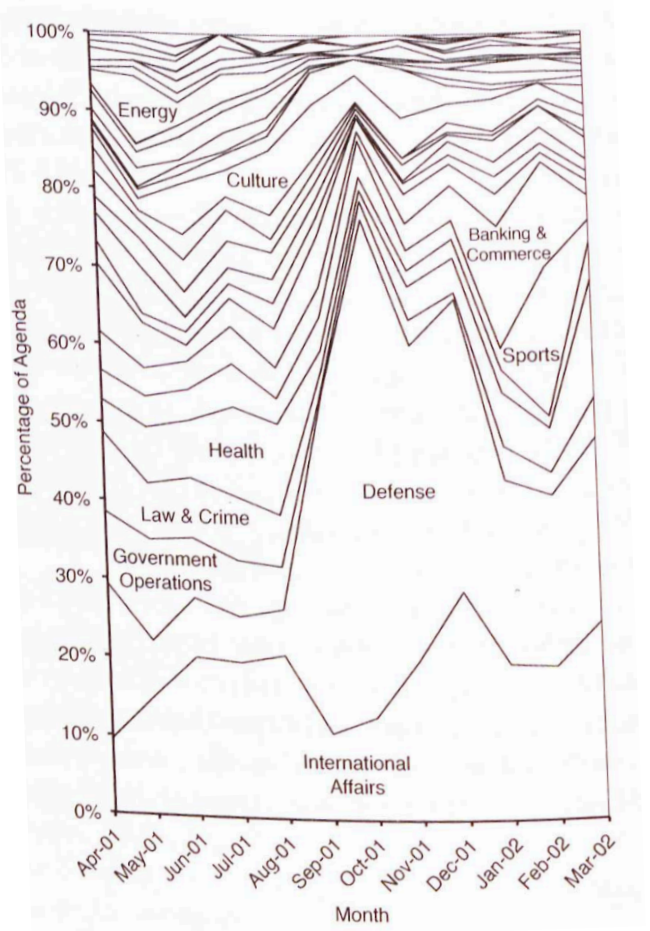
RECAP: MEASUREMENT USING TEXT DATA

- Text mining is often about “producing data” – a (numerical) **summary** of the documents in question
- With the methods we’re using today, these produced data can look like…
  - A discrete label from binary classification (e.g., “positive/negative”, being about a certain topic, “sexist/non-sexist”)
  - A discrete label from multinomial classification (e.g., multiple topics, authors)
  - A continuous value (sentiment, probability of having a certain label, ideological scaling)

⟹ We can then eventually use these values/label counts to test hypotheses

RECAP: MEASUREMENT USING TEXT DATA

- Example: labels counted over time
- International politics frames that made it to NYT headlines in 2001 (Boydstun 2013; taken from Grimmer et al. 2022)



RECAP: MEASUREMENT USING TEXT DATA

- Example: using classification accuracy as **continuous indicator** for speech polarization

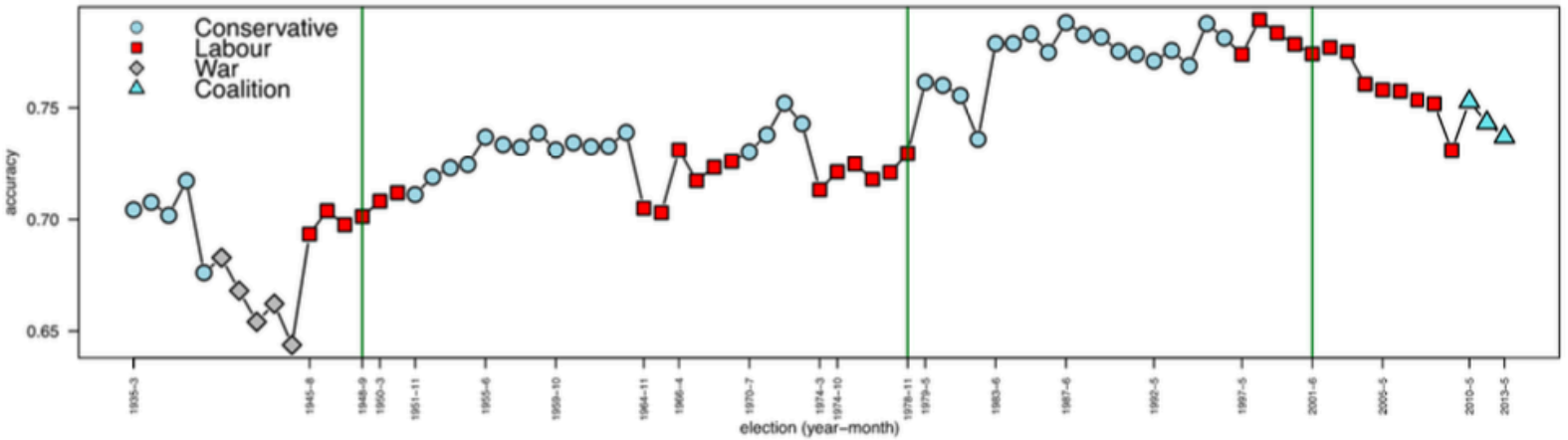


Figure 3. Estimates of parliamentary polarization, by session. Election dates mark *x*-axis. Estimated change points are [green] vertical lines.

HOW TO PRODUCE THESE DATA?

Most basic approach: read the text

1. Develop a coding scheme (based on prior theory)
2. Read text, decide on annotation based on coding scheme
3. Do it for all your documents
4. …
5. …there is plenty of text available now, so it takes forever…
6. **Consider different career paths over and over again as this process sucks so hard**

⟹ Luckily, there are computational tools we can harness to take away some of the pain

⟹ **MACHINE LEARNING**

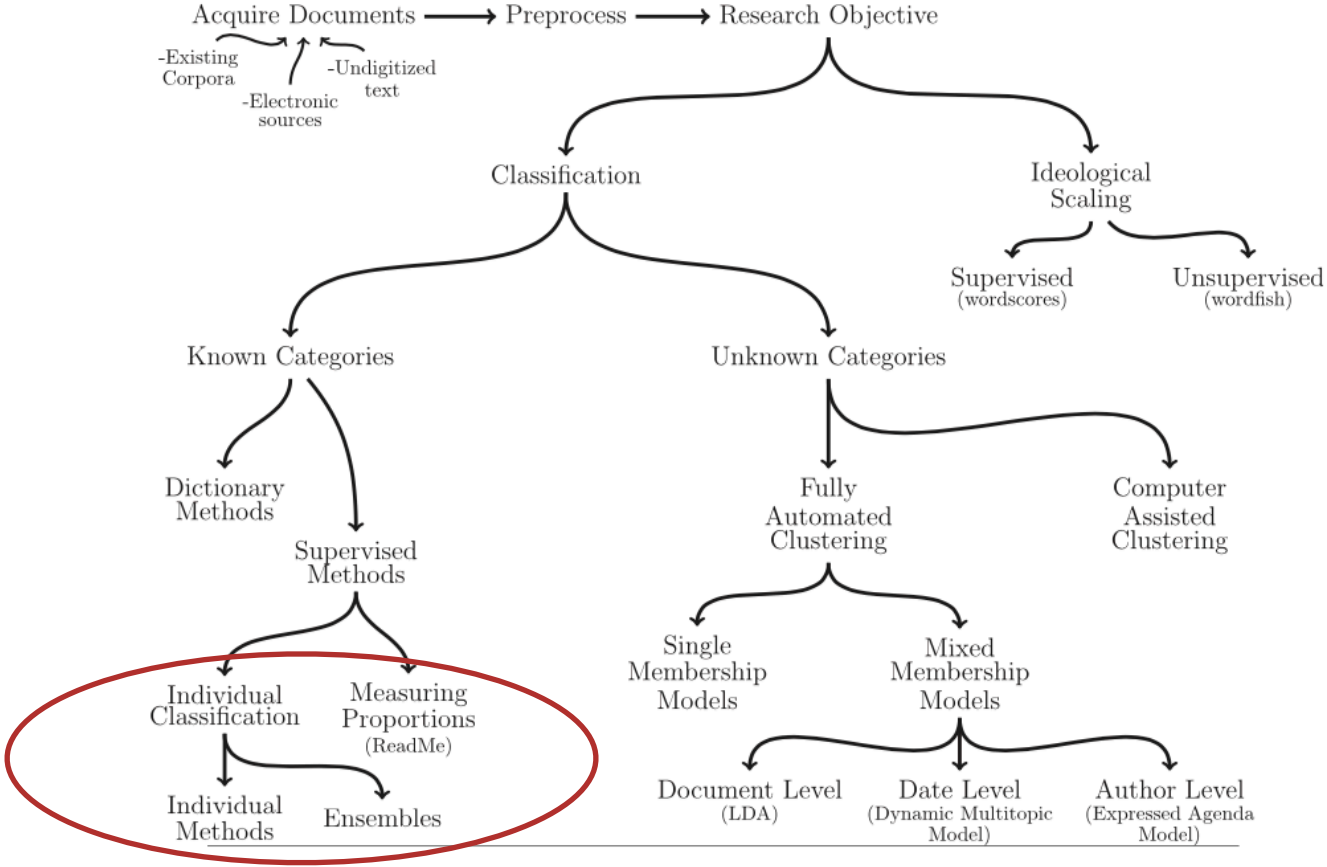
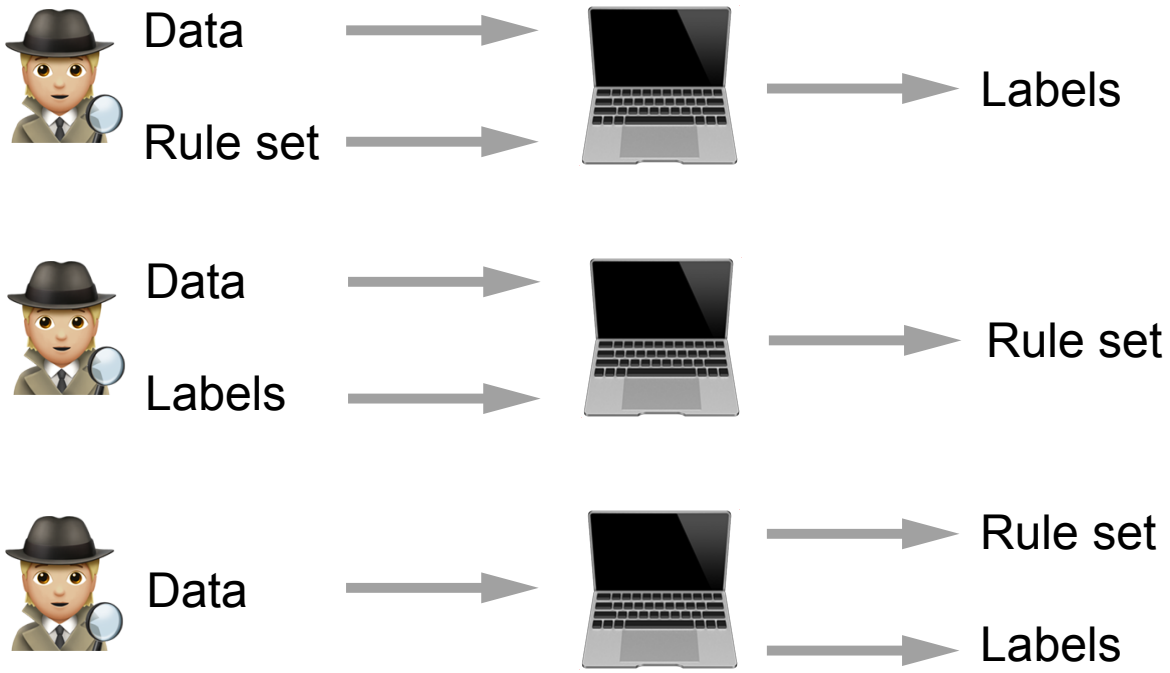


Fig. 1 An overview of text as data methods.

inspired by Ash (2018)

HOW TO PRODUCE THESE DATA?

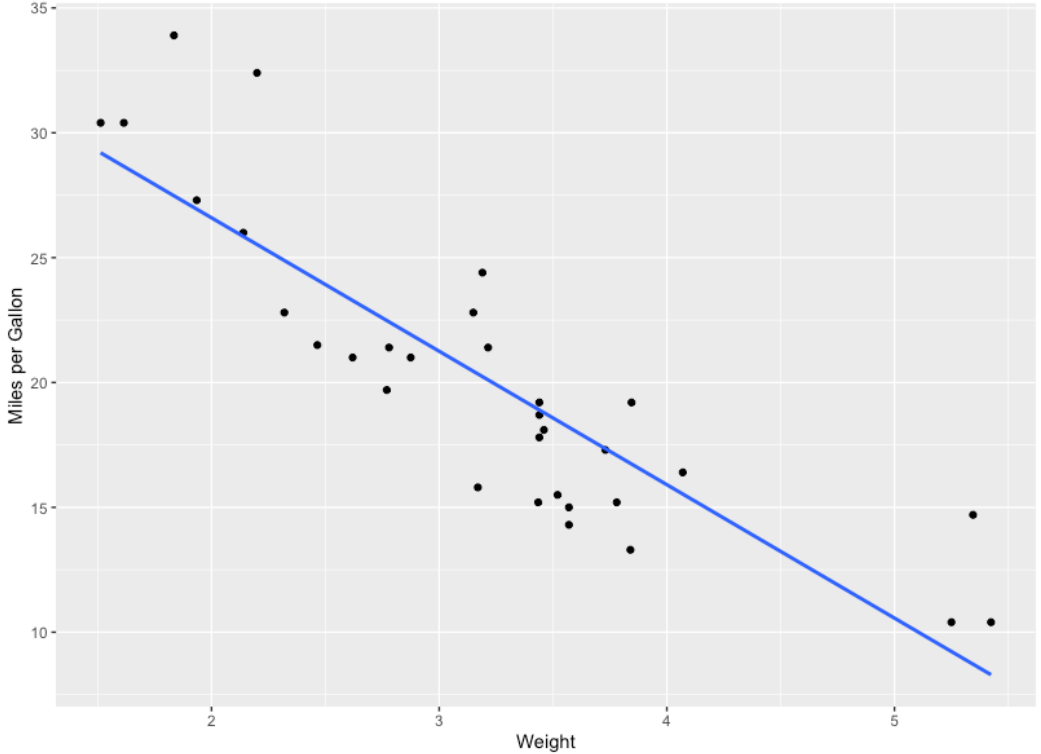


Dictionary-based analysis
Computer applies rules

Supervised ML
Computer learns relationship (“rules”) between data and answers

Unsupervised ML
Computer suggests rules and answers based on patterns in data

EVEN OLS IS MACHINE LEARNING IF YOU WILL



$$MPG = \beta_0 + \beta_1 Weight + \epsilon$$

inspired by Ash (2018)

HOW DOES IT LOOK FOR TEXT – "TEXT REGRESSION"

Objective: to learn a model that maps an outcome Y (label) to the features W' (words)

$$Y_i = \beta W_i' + \epsilon_i$$

⟹ Requires labeled documents

⟹ Features (words) are treated as predictors

⟹ Algorithms will not accept words – we use word counts (alternatives: "one-hot encoding" (1 if word is present in document, 0 if not), tf-idf values, embedding vectors)

HOW DOES IT LOOK FOR TEXT – "TEXT REGRESSION"

Objective: to learn a model that maps an outcome Y to the features W'

⟹ Eventually, predictions can be made on unseen documents

⟹ Different approaches/algorithms exist – which one to choose depends on computational capabilities and desired outcome (i.e., discrete label – binary or multinomial – or continuous value)

SUPERVISED LEARNING WITH TEXT – THE PROCESS

- Choose a set of documents (corpus)
- Annotate a sub-set of the corpus
- Split the annotated set into training and test set (for validity assessment)
- Preprocess the documents
  ⇨ e.g., tokenization (also: bi- and trigrams), weighting, stemming/lemmatization, etc. – whatever works best
- Train a classifier on training set
  ⇨ tuning with cross-validation
- Evaluate classifier using test set and confusion matrix
- If sufficient, apply it to unlabeled data

(for a hands-on guide, see Barberá et al. 2021)

CHOICE OF CORPUS

- Must fit the question
- Usual approach: keyword-based search (e.g., using regular expressions)
  ⟹ has its own pitfalls though, see Barberá et al. (2021) and King, Lam, and Roberts (2017)

CHARACTERISTICS OF A GOOD ANNOTATED SET (GRIMMER ET AL. 2022, P. 190)

- **Objective–intersubjective**: categories are *objectively* measured; researchers have a *shared understanding* of them
- **A priori**: codebook is derived from theory
- **Reliable**: annotation process is repeatable across coders – will yield same results
- **Valid**: concept of interest is clearly measured
- **Generalizable**: the training set is a representative sample of the underlying texts (and also the final population)
- **Replicable**: approaches should replicate with same and different data

ANNOTATION OF TRAINING AND TEST SET OF CORPUS

Step 1: Randomly sample documents from corpus

- Sample should be representative (e.g., if corpus spans a long time period, has different authors, etc.)
- Usually, algorithm can only derive rules for terms it has seen

ANNOTATION OF TRAINING AND TEST SET OF CORPUS

Step 2: Define your codebook

- Usually: rules depend on your theory
  - They need to be stated explicitly (in paper and/or appendix)
  - Ideally, you find examples from the data for each rule
    ⟹ To guide your reader

    ⟹ But also for yourself
- Sometimes, codebooks are already available (e.g., from related studies)

ANNOTATION OF TRAINING AND TEST SET OF CORPUS

Step 3: Get other coders/get ready to annotate multiple times

- Needed to assess the reliability of the coding process
  ⟹ Either between raters

  ⟹ If only one rater exists: multiple timepoints
- Also a test for the codebook
- Finally, agreement between coders needs to be assessed
- Ideally: make a test run with a set that will be later discarded to ensure that concepts are understood; discuss cases of disagreement
- More on this: Barberá et al. (2021)

ANNOTATION OF TRAINING AND TEST SET OF CORPUS

Step 4: Determine training and test set

- Training set: used to train the model
- Test set: used to evaluate performance
- Usual split: 80/20
- Important: classes should be equally represented in training and test set (can be mitigated using upsampling or downsampling)

PREPROCESSING

- No one-fits-all solution
- *recipes* and the *tune package* make it easy to experiment a bit
- Common steps:
  - Using bi- and trigrams
  - Weighting by TF or TF-IDF
  - Stemming/Lemmatization
  - Removal of rare/common words or stopwords (feature reduction)

TRAINING THE CLASSIFIER

Step 1: Choose a classifier

⟹ Depends on question and computational capabilities

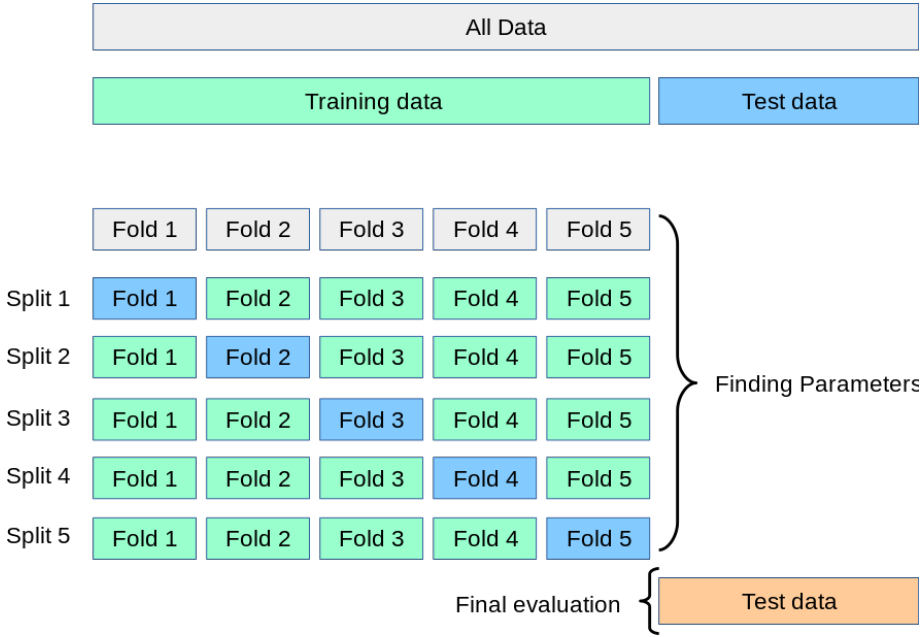
- Do you want to predict continuous or categorical value?
- Will you run the models on a server or your own laptop?

Step 2: Train classifier(s) using training set

- Use different specifications of training set
- Use different classifiers

Step 3: Cross-validate and tune different specifications to find optimal solution

CROSS-VALIDATION



https://scikit-learn.org/stable/modules/cross_validation.html

FINAL EVALUATION

How well does the classifier compare to gold standard data?

Example: Sentiment Analysis

PREDICTED VALUES	ACTUAL VALUES: POSITIVE	ACTUAL VALUES: NEGATIVE
POSITIVE	TRUE POSITIVE	FALSE POSITIVE
NEGATIVE	FALSE NEGATIVE	TRUE NEGATIVE

FINAL EVALUATION

How well does the classifier compare to gold standard data?

Accuracy: $\frac{TP + FN}{TP + FP + FP + FN}$ – how many predictions are correct (reasonable if labels are balanced!)

Precision: $\frac{TP}{TP + FP}$ – how many positive predictions are correct

Recall/Sensitivity: $\frac{TP}{TP + FN}$ – how many actual positives are predicted properly

F1-score: $2 \times \frac{Precision \times Recall}{Precision + Recall}$ – harmonic mean of precision and recall

FINAL EVALUATION

Table 1: Classification Performance Metrics

Class	Precision	Recall	F1-Score	Support
Unpolarised Rhetoric	0.97	0.95	0.96	87
Polarised Rhetoric	0.71	0.77	0.74	13
Macro Avg.	0.84	0.86	0.85	100
Weighted Avg.	0.93	0.93	0.93	100
Overall Accuracy: 0.93				

REMARKS

These methods are great and robust, but (unfortunately) will be outdated in the near future: transfer learning using large language models is going to replace them – for more on this, wait for TAD IV

The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy

Salomé Do[1,2], Étienne Ollion[3], and Rubing Shen[2,3]

REMARKS

Idea behind transfer learning:

- problem with “local classifiers:” they can only predict based on the relationships they observe – and text is often very ambiguous and vague
- *learn* the relationships between words on huge corpora – the model
- *transfer* this knowledge to the task at hand, enjoy increased performance

REMARKS

There's also crazy new stuff powered by the same idea (one large model in the background)

- zero-shot learning (no training examples shown, just the labels)
- one-/few-shot learning (one/few training examples shown)
- passing your coding scheme and some text to ChatGPT, ask it for classification of text based on the scheme (– works fairly well with the newer models, not so well with the old ones)

⟹ However, you will still have to have your own annotated data to check for the performance of the classifier

REMARKS

- This stuff is mostly implemented in Python
- Huggingface 🤗 is your friend here
- But it’s quite user-friendly, we’ll have a look soon

BONIKOWSKI, LUO, AND STUHLER 2022

Politics as Usual? Measuring Populism, Nationalism, and Authoritarianism in U.S. Presidential Campaigns (1952–2020) with Neural Language Models

MOTIVATION

- Radical-right politics contain populism, exclusionary and declinist nationalism, and authoritarianism
- Present on both demand (voter holding opinions) as well as on supply side (politicians using frames to motivate voters)
- But how pervasive are these in the mainstream?
  - Donald Trump as first RRP candidate
  - They compare his rhetorics against prior candidates

THEORY

RRP's rhetoric consists of populism, nationalism, and authoritarianism

- Populism: "a form of moral claims-making that juxtaposes a fundamentally corrupt elite with the virtuous people and promises to restore political power to the latter" (p. 1727)
- Nationalism: "articulation of distinct conceptions of nationhood … to either highlight the nation's present-day virtues or to offer a critique of the nation's decline and an alternative vision for its future" (pp. 1727-8)
- Authoritarianism: "the targeted use of state power against alleged domestic enemies … in a manner that undermines liberal rights regimes and democratic norms and institutions" (p. 1728)

⟹ shall be more present in D. Trump's speeches in 2016 and 2020 than in more mainstream candidates' speeches

⟹ when it comes to nationalism, both exclusionary nationalist claims and low levels of national pride are increasingly used by Trump

THEORY – EXAMPLES

Populism:

- Trump: “It’s going to be a victory for the people. A victory for the everyday citizen whose voice hasn’t been heard. It will be a win for the voters, not the pundits, not the journalists, not the lobbyists, not the global special interests funding my opponent’s campaign.”
- Obama: “Finally, the American people must be able to trust that their government is looking out for all of us—not the special interests that have set the agenda in Washington for eight years, and the lobbyists who run John McCain’s campaign. I’ve spent my career taking on lobbyists and their money, and I’ve won.”

THEORY – EXAMPLES

Nationalism:

- Excluding – Trump: “Expanding President Obama’s unconstitutional executive amnesty, including instant work permits for millions of illegal workers. Freeing even—there go your jobs—freeing even more criminal aliens by expanding Obama’s non-enforcement directives. And this is to me, the beauty of them all. Obama has allowed thousands and thousands and thousands of people to come in, Syrians from the Middle East. She wants an increase of 550 percent in Syrian refugees into our country.”
- Including – Carter: “We can have an America that provides excellence in education to my child and your child and every child. We can have an America that encourages and takes pride in our ethnic diversity, our religious diversity, our cultural diversity– knowing that out of this pluralistic heritage has come the strength and the vital- ity and the creativity that has made us great and will keep us great.”

THEORY – EXAMPLES

Nationalism – national pride:

- Low level – Trump: “We’ve lost 70,000 factories since China’s entry into the World Trade Organization. Another Bill and Hillary backed disaster. We are living through the greatest job theft in the history of the world. More jobs have been stolen from our country, so stupidly we let them go. We let our companies go so foolishly. We don’t know what we’re doing. A Trump administration is going to renegotiate NAFTA, stand up to the foreign cheating, and stop the jobs from leaving our country, and have jobs come back in the other direction.”
- High level – Stevenson: “America is a great, a strong, a wise, and most of all a good country. And I believe with all my heart that by these qualities, we can and we will safely in God’s good time win our way to a peaceful world.”

THEORY – EXAMPLES

Authoritarianism:

- Law and order – Nixon: “[W]hen we find [...] in city after city in this country that convicted murderers, convicted rapists, are turned free, confessed murderers, I mean, and confessed rapists, are turned free after they confess their crime because of a technicality, then I say that our courts in their decisions have gone too far in weakening the peace forces as against the criminal forces in this country.”
- Anti-immigrant – Trump: “That’s why I was so happy what we did to annihilate the enemy the other day. So happy. Because we’re dealing against a very dishonest system. But Hillary, so important, wants to have a radical, and this is very radical, immigration. She wants to radicalize immigration where you have people pouring in. Remember this, the border patrol agents, 16,500 gave me their endorsement. Last week, … these are great people, you don’t hear great things because they’re not allowed to do their job.”

THEORY – HYPOTHESES

Hypotheses:

- Populist claims are used by both parties
- Exclusionary nationalist claims were employed by main- stream Republican candidates, but less frequently so than populism, authoritarianism, and low pride.
- Authoritarian claims were used more frequently by mainstream Republican candidates than by mainstream Democratic candidates.
- Mainstream candidates relied on authoritarian claims more frequently than on exclusionary claims.
- Inclusive nationalist claims were employed more frequently by mainstream Democratic candidates than by mainstream Republican candidates.
- Low levels of national pride were evoked by both Democratic and Republican mainstream candidates throughout the time series.

THEORY – HYPOTHESES

Hypotheses:

- Low national pride and populism were positively correlated among mainstream campaigns.
- Low national pride and high national pride were negatively correlated among mainstream campaigns.
- Authoritarianism and inclusive nationalism were negatively correlated among mainstream campaigns.

THEORY – HYPOTHESES

Hypotheses:

- Low national pride and populism were used more frequently by challenger candidates than by nominees of the incumbent party.
- High national pride was used more frequently by incumbent party candidates than by challengers.
- Donald Trump's 2020 campaign, when he ran as an incumbent, featured fewer references to populism and low national pride than his 2016 campaign, when he ran as a challenger.

DATA AND METHOD

- 2,956 speeches of democrat and republican candidates, split into 71,808 paragraphs
- challenge: the concepts are hard to measure – basically impossible with BoW approaches because of irony, polysemy; also: rarely occurring
- Used advanced large model (RoBERTa) and active learning to do classification

Table 2. Analytical steps for labeling data, RoBERTa fine-tuning, active learning, and corpus classification.

1. Randomly sample 2,224 paragraphs from the corpus. Then, follow the steps below for each each of the six non-mutually-exclusive frames.
2. Annotate each paragraph (by two independent annotators, with disagreements adjudicated by a third annotator).
3. Identify suitable RoBERTa fine-tuning hyperparameters through a manual search.
4. Generate 25 random splits in the labeled data, with 80 percent of each split assigned to a training set and 20 percent to a test set.
5. Run separate RoBERTa models on the 25 splits and use each test set to evaluate model performance.
6. Retain the best model (using area under the precision-recall curve—i.e., PR-AUC) to predict classification probabilities for each paragraph in the unlabeled data.
7. Generate a new 200-paragraph sample from the unlabeled data with 20 percent of the paragraphs drawn randomly and 80 percent drawn based on high entropy (i.e., paragraphs with classification probability closest to 0.5).
8. Annotate the 200 paragraphs (by two independent annotators, with disagreements adjudicated by a third annotator).
9. Add the annotated paragraphs to the training sets of the 25 random splits described in step 4.
10. Repeat steps 5–9, evaluating on the same test sets. In our case, two rounds of hand-coding sufficed (resulting in three runs of the RoBERTa classifiers).
11. In the final round of models, use all 25 train-test splits to generate independent predictions for the whole corpus. Average these predictions for each paragraph and assign a label according to a .5 probability cutoff.

SIDE NOTE: ACTIVE LEARNING

- Problem: how to efficiently annotate instances of text
- Idea: let classifier come up with suggestions where it could “learn” the most
  - this is, where it is the most insecure about the decision

Approach:

1) Annotate sample
2) train model
3) predict on new sample, extract numeric predictions (probability of class being 1 or 0 when annotating 2 classes)
4) choose predictions that are as close as possible to 0.5
5) annotate them, add them to the training set

Table 3. Classifier performance and proportion of positive cases in all paragraphs.

	Inclusion	Exclusion	Authoritarianism	Populism	High pride	Low pride
Precision	0.71	0.85	0.66	0.68	0.63	0.64
Recall	0.76	0.78	0.72	0.61	0.72	0.55
Accuracy	0.97	1	0.98	0.98	0.95	0.96
Prevalence in training set	0.048	0.01	0.027	0.029	0.076	0.055
Predicted corpus prevalence	0.051	0.005	0.023	0.026	0.088	0.042

FIGURE B.1. Proportion of speeches containing one or more populist, nationalist, and authoritarian paragraphs by campaign, 1952-2020.

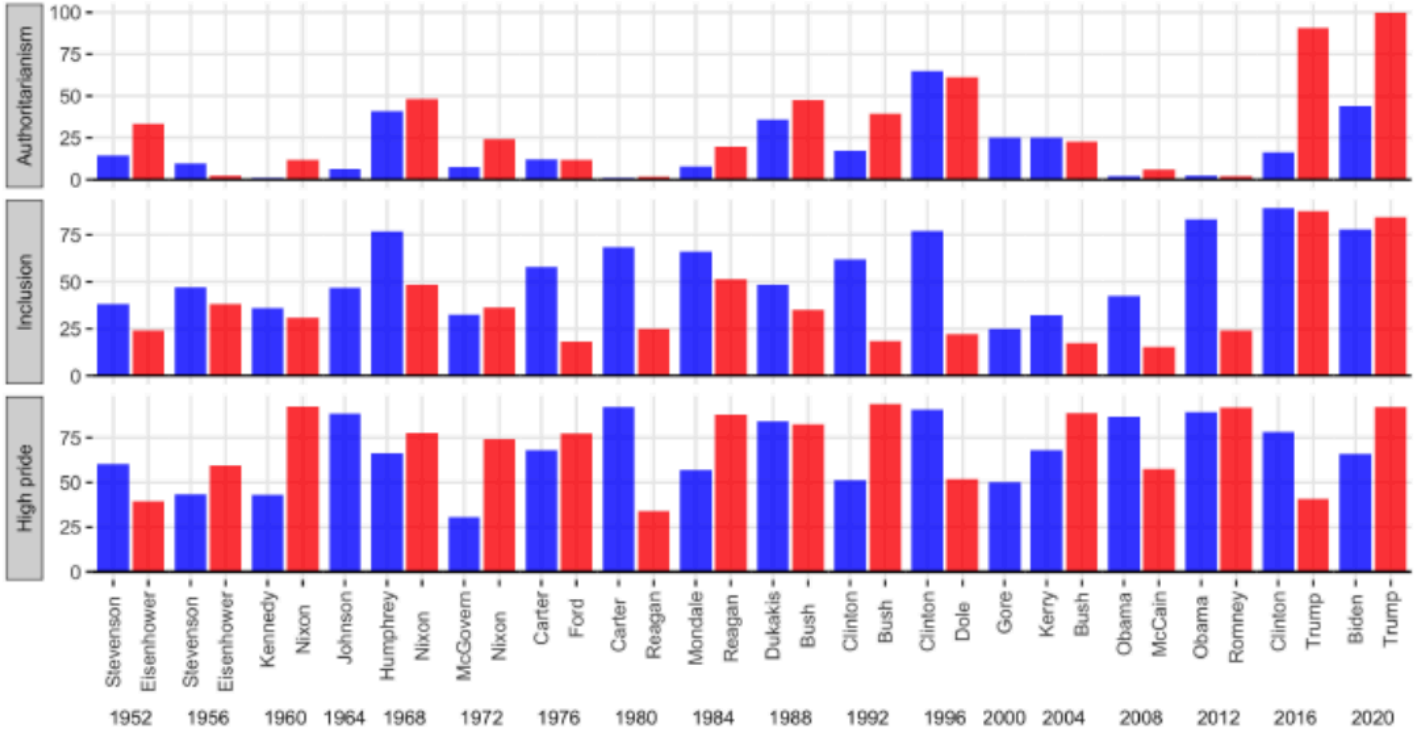
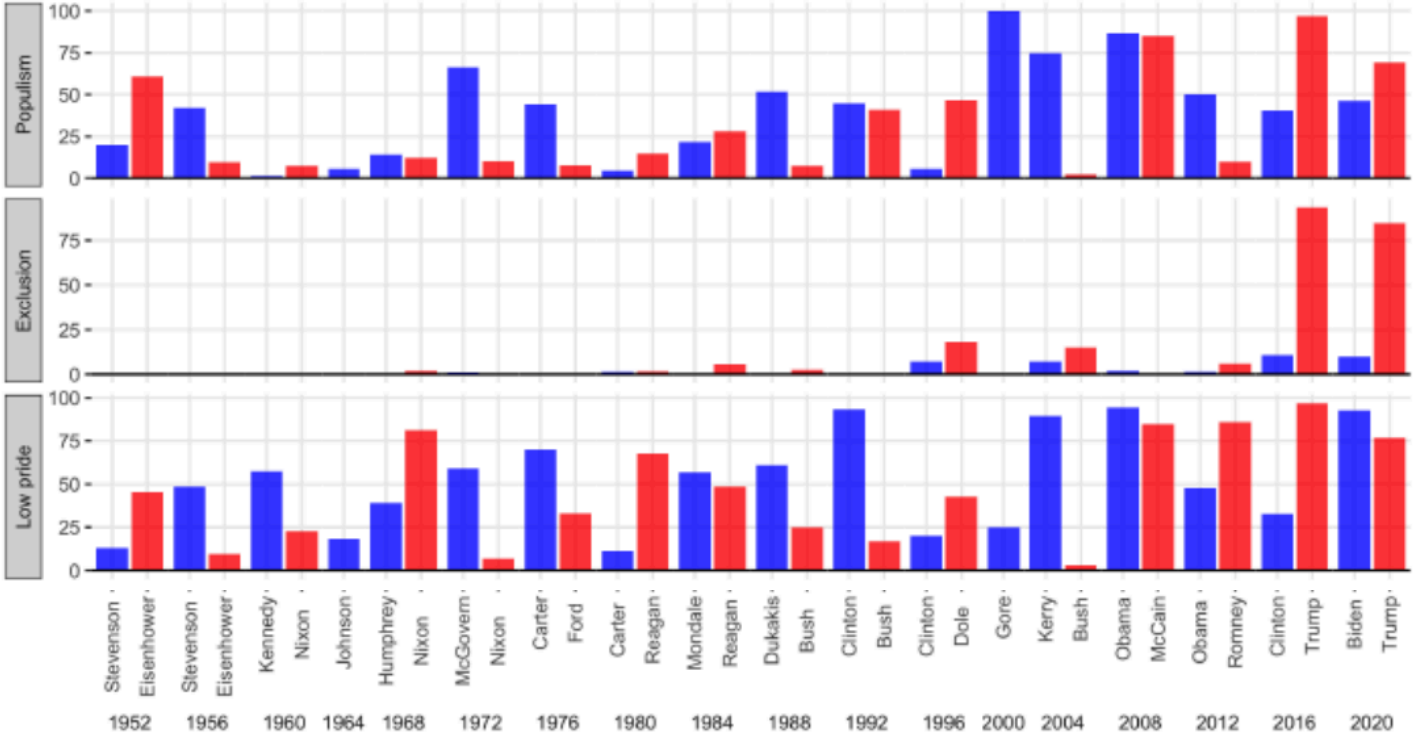


FIGURE B.1. Proportion of speeches containing one or more populist, nationalist, and authoritarian paragraphs by campaign, 1952-2020.



THEORY – HYPOTHESES

Hypotheses:

- More populism, exclusionary nationalism, low level pride, and authoritarianism in Trump's speeches – **YES, low level pride only in 2016 though**
- Populist claims are used by both parties – **YES**
- Exclusionary nationalist claims were employed by mainstream Republican candidates, but less frequently so than populism, authoritarianism, and low pride. – **NO, TRUMP ONLY**
- Authoritarian claims were used more frequently by mainstream Republican candidates than by mainstream Democratic candidates. – **YES**
- Mainstream candidates relied on authoritarian claims more frequently than on exclusionary claims. – **YES**
- Inclusive nationalist claims were employed more frequently by mainstream Democratic candidates than by mainstream Republican candidates. – **YES**
- Low levels of national pride were evoked by both Democratic and Republican mainstream candidates throughout the time series. – **YES**

THEORY – HYPOTHESES

Hypotheses:

- Low national pride and populism were positively correlated among mainstream campaigns. – **YES**
- Low national pride and high national pride were negatively correlated among mainstream campaigns. – **YES**
- Authoritarianism and inclusive nationalism were negatively correlated among mainstream campaigns. – **YES**

THEORY – HYPOTHESES

Hypotheses:

- Low national pride and populism were used more frequently by challenger candidates than by nominees of the incumbent party. – **YES**
- High national pride was used more frequently by incumbent party candidates than by challengers. – **YES**
- Donald Trump's 2020 campaign, when he ran as an incumbent, featured fewer references to populism and low national pride than his 2016 campaign, when he ran as a challenger. – **YES**

UNIVERSITÄT
LEIPZIG

MERCI

Felix Lennert
Institut für Soziologie

felix.lennert@uni-leipzig.de
www.uni-leipzig.de

REFERENCES

- Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis* 29(1):19–42.
- Denny, Matthew J. and Arthur Spirling. 2018. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." Political Analysis 26(2):168–89.
- DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41(6):570–606.
- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press.
- Hurtado Bodell, Miriam, Måns Magnusson, and Marc Keuschnigg. n.d. "Seeded Topic Models in Digital Archives: Analyzing the Swedish Understanding of Immigration, 1945–2019." OSF Preprint.
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. 2013. "The Structural Topic Model and Applied Social Science." in *NIPS 2013 Workshop on Topic Models*.