Bias in Text as Data

Proportional Classification, Evaluation and Correction

Snorre Ralund, Ph.D Fellow, University of Copenhagen, SoDaS

KØBENHAVNS UNIVERSITET

Bias in NLP

Bias in NLP

- "The assumptions of random sampling is violated" –
  - Training != Target ("In the wild")
  - Performance is only estimated in the Test Set.
- Control over the sampling and dataset construction to mitigate bias is lost using pre-trained language models.
- Model is greedy, picks up any association / pattern. !=Causal.
  - If a certain group is involved in many conflicts, model associates group with conflict.

Bias in NLP

Bias in NLP: Examples

- **People express biases that models learn**
  - Model learns correlations and association which are non-causal.
  - Bolukbasi et al. 2016: "Man is to woman as computer programmer is to homemaker"
  - Manzini et al. 2019: "Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings"
- **Different people express themselves differently.**
  - Model learns styles
  - Can be used for author attribution
    - geo located (privacy)
    - gender
    - ethnicity
  - Johansen, Hovy and Søgaard 2015: "Cross-lingual syntactic variation over age and gender"

Bias NLP - Biased Measurement

Using NLP systems for measurement introduce bias.

- Hovy and Søgaard 2015: "Tagging performance correlates with author age"
  - Parsers were trained on old newspaper data.
- Jørgensen, Hovy and Søgaard 2015: "Challenges of studying and processing dialects in social media"
  - Parsers were significantly worse in relation to african american dialect.
  - --> NLP technology systematically disadvantages groups of non-standard language users.
- Kiritchenko & Mohammad 2018: "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems"

Bias NLP - Biased Measurement

Feature	Positive	Negative	Total count
/r/ → /Ø/ or /ə/	brotha	brother	9528
	foreva	forever	3673
	hea	here	4352
	lova	lover	1273
	motha	mother	4668
	ova	over	3441
	sista	sister	5325
	wateva	whatever	2974
	wea	where	5153
	total		40,387
/str/ → /skr/	skreet	street	1226
	skrong	strong	1629
	skrip	strip	1101
	total		3956
/ð/ → /d/ or /v/	brova	brother	3715
	dat	that	2610
	deez	these	4477
	dem	them	3645
	dey	they	2434
	dis	this	2135
	mova	mother	2462
	total		21,478
/θ/ → /t/ or /f/	mouf	mouth	3861
	nuffin	nothing	2861
	souf	south	1102
	teef	teeth	1857
	trough	through	2804
	trow	throw	1090
	total		13,575
All tweets			79,396

Table 1: Word pairs and counts

Bias in NLP

http://onlinehub.stanford.edu/cs224/stanford-cs224n-nlp-with-deep-learning-winter-2019-lecture-19-bias-in-ai

- Conceptualization of Bias:
  - Cognitive (Annotator as Source), Sampling, Statistical and Algorithmic Bias.
- Ethical obligations and consequences in relation to open sourcing and applications.
- Bias Detection – Curated critical test dataset, Synthetic datasets (e.g. Kiritchenko & Mohammad 2018)
- Bias Mitigation Methods
  - Multi-task adversarial learning for bias mitigation https://www.aclweb.org/anthology/P17-1001/
  - Mitigation by adding neutral data to learn neutral representation of subgroup concepts.

Bias in Text as Data

"A Method of Automated Nonparametric Content Analysis for Social Science"

- *"computer scientists may be interested in finding the needle in the haystack [···], but social scientists are more commonly interested in characterizing the haystack"* (Hopkins & King 2010:230)

- *"Unfortunately, except at the extremes, there exists no necessary connection between low misclassification rates and low bias: it is easy to construct examples of learning methods that achieve a high percent of individual documents correctly predicted and large biases for estimating the aggregate document proportions, or other methods that have a low percent correctly predicted but nevertheless produce relatively unbiased estimates of the aggregate quantities."*(Hopkins & King 2010: 234)

Optimizing for a Different Goal

Individual Classification vs Proportional Classification

Goal: the estimation of category proportions, trends and time series analysis, , correlation of categories with text-external covariates.

Problem: Bias in the Proportional estimate → errors in conclusions and potentially artificial effects.

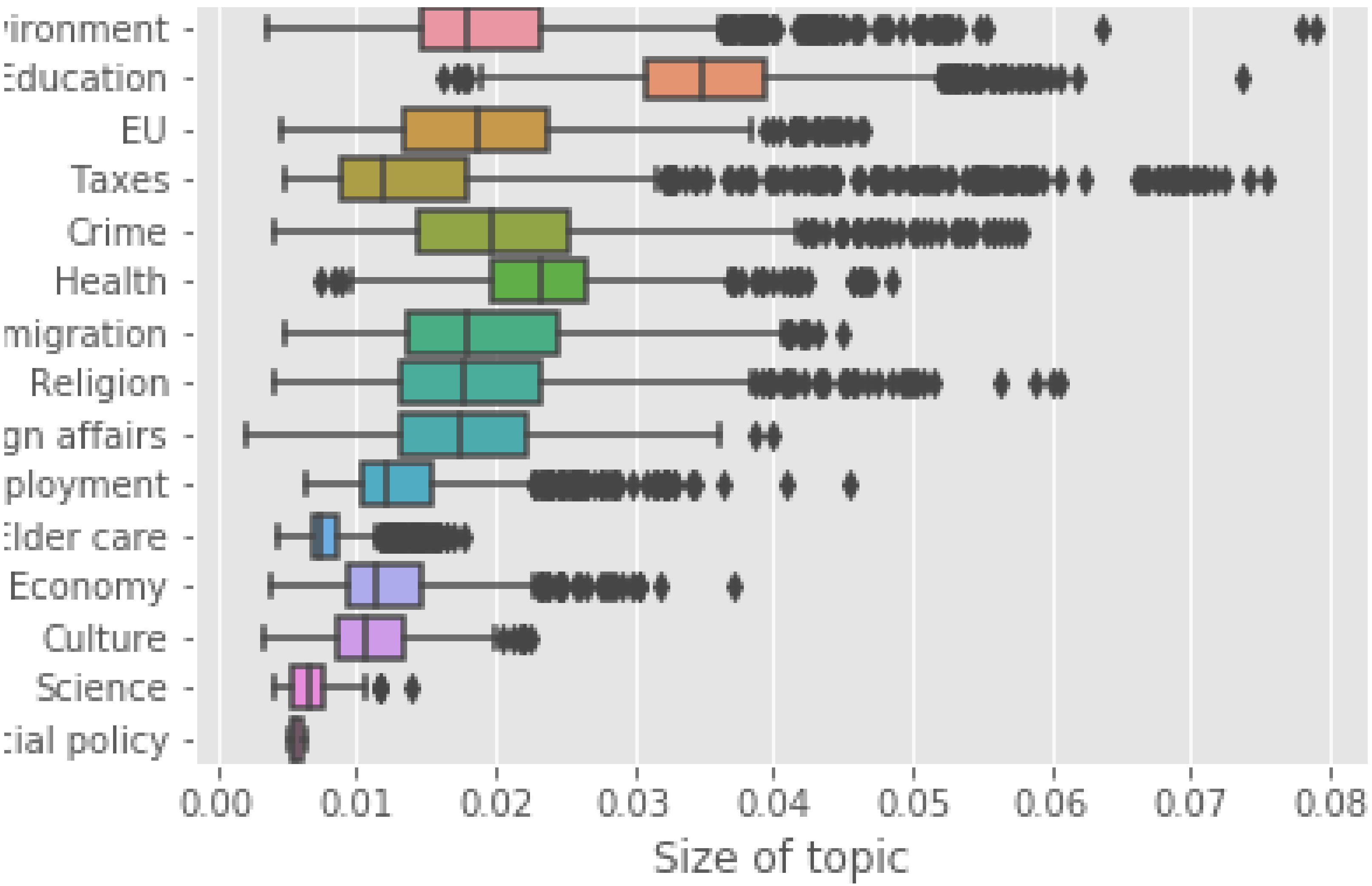
Case: Unvalidated Topic Models

***Automation Bias* in the Social Sciences**
Researchers use unvalidated unsupervised models for measurement.

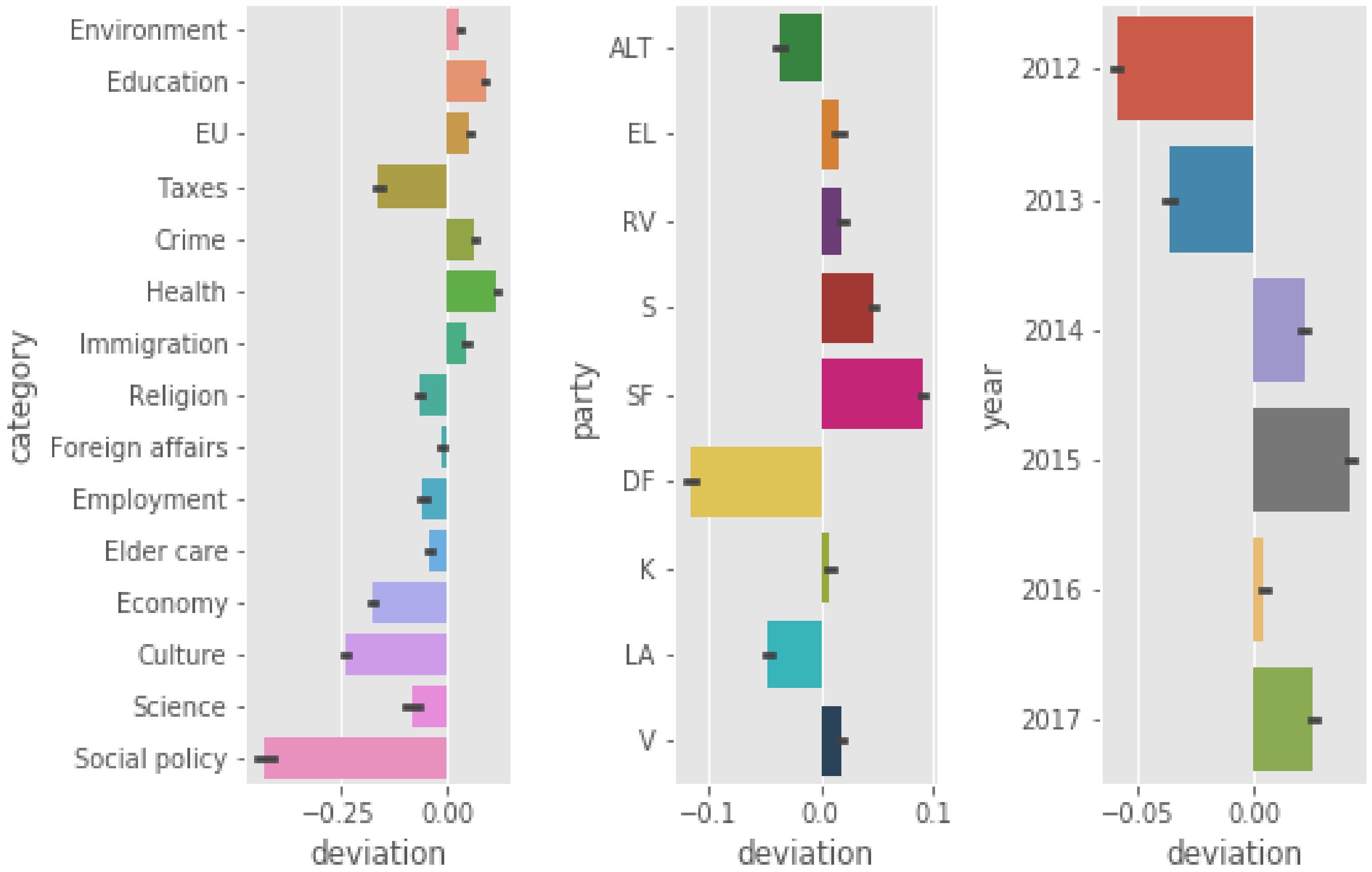
- Asume you discover “natural clusters”, and do simple quality checks.

··· However choice of model, including large number of hyperparamters can alter the result.

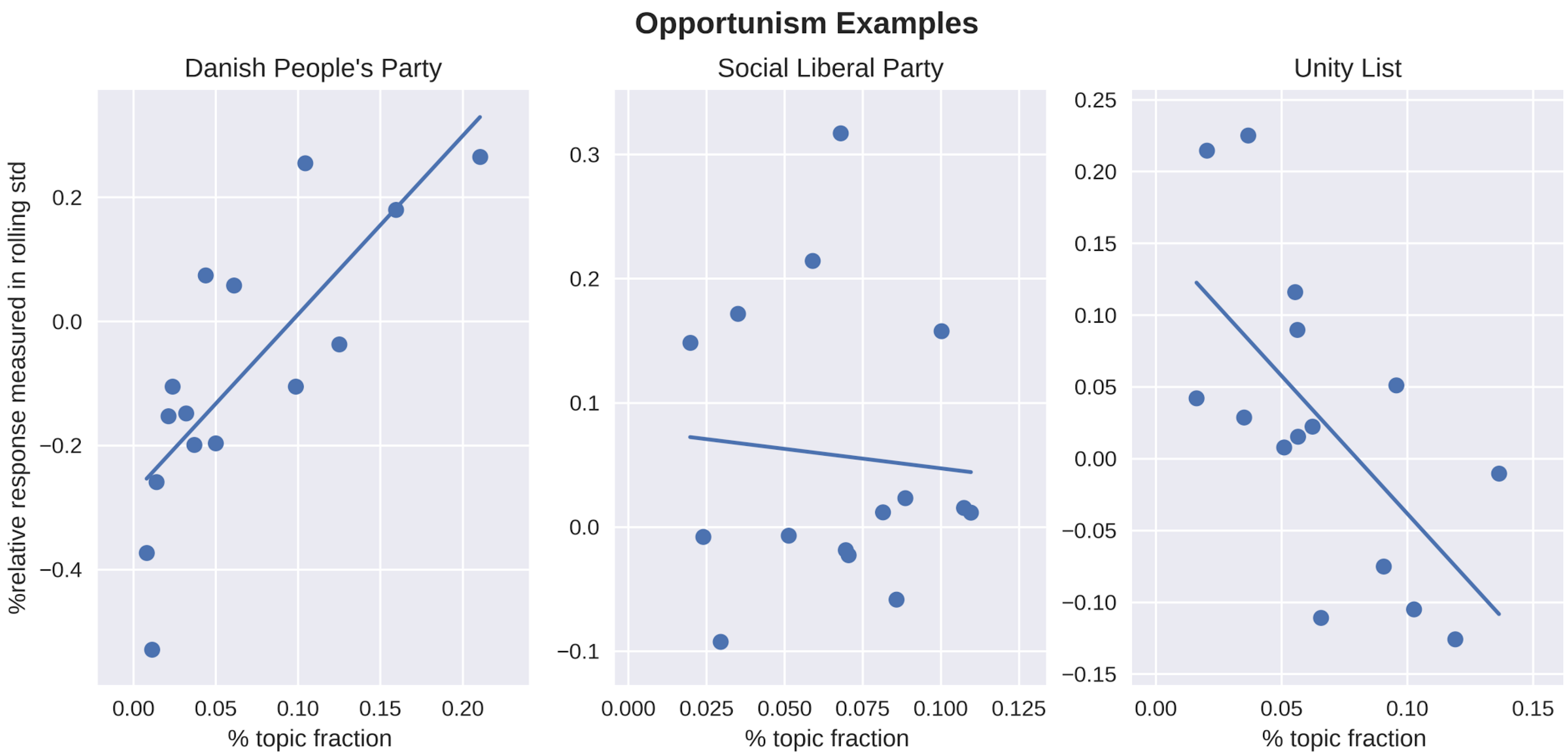
Case: Unvalidated Topic Models



Case: Unvalidated Topic Models
Differential Bias

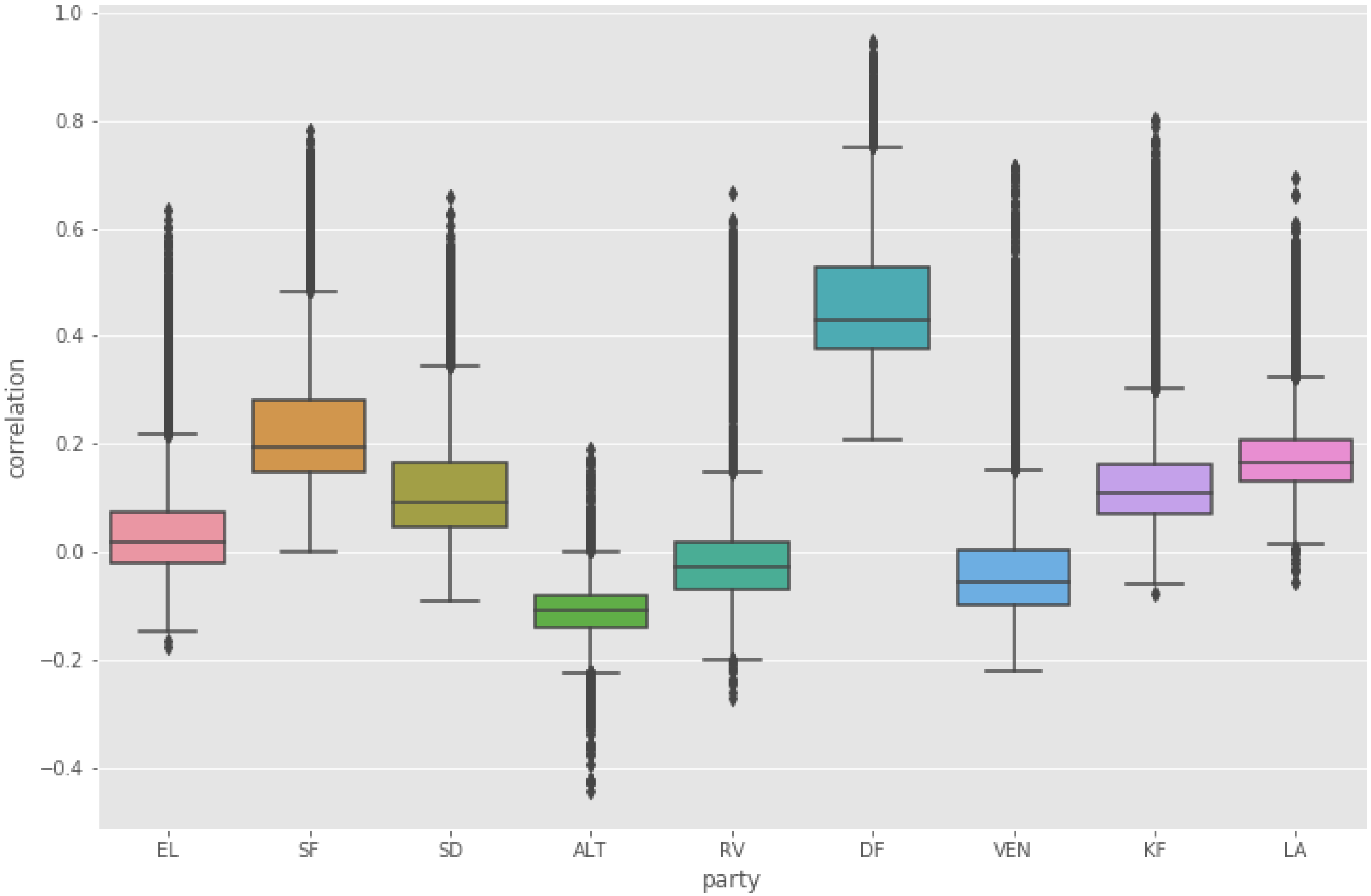


Case: Unvalidated Topic Models



Case: Unvalidated Topic Models

Distribution of Conclusions



Optimizing for a Different Goal

Individual Classification vs Proportional Classification

- "Classify-and-Count".

- "Direct estimation" of proportions without individual classification.

Hopkins & King 2010: Two settings

1. Ideal: Where Labeled and Unlabeled Data is drawn from the same distribution.

   - Here we can use "Classify-and-Count" plus simple correction using the Test Set for estimating the errors.

2. Continous flow of data + Re-use of Model.

   - E.g. relevant social media or news articles coming in continously.
   - "Count-and-classify" will not work (Hopkins and King 2010)
     - Use Direct Estimation / "regression" method.

Ideal Case: Estimation and Correction of Misclassification

- *[U]se the test set's labels to calculate the specific misclassification probabilities between each pair of actual classifications given each true value P(ˆDi=j|Di=j'). These misclassification probabilities do not tell us which documents are misclassified, but they can be used to correct the raw estimate ofthe document category proportions.* (Hopkins & King 2010:235)

Confusion Matrix for Bias Correction

		Predicted class	
		+	-
Actual class	+	**TP** True Positives	**FN** False Negatives Type II error
	-	**FP** False Positives Type I error	**TN** True Negatives

Confusion Matrix for Bias Correction

		Predicted class	
		+	-
Actual class	+	**TP** True Positives	**FN** False Negatives Type II error
	-	**FP** False Positives Type I error	**TN** True Negatives

$$\mathrm{P(D)} = \frac{TP-FP}{TP+FP} * \hat{P} + \frac{FN}{FN+TN} * \hat{N}$$

-- i.e. substract overestimation (FP), add underestimation / missed cases (FN).

Confusion Matrix for Bias Correction

		Predicted class	
		+	-
Actual class	+	**TP** True Positives	**FN** False Negatives Type II error
	-	**FP** False Positives Type I error	**TN** True Negatives

$$P(\hat{D}=1) = (\text{sens})\, P(D=1) + (1-\text{spec})\, P(D=2)$$

$$P(D=1) = \frac{P(\hat{D}=1) - (1-\text{spec})}{\text{sens} - (1-\text{spec})}.$$

Ideal Case

“We just need a large enough test set to measure cells in the confusion matrix with high statistical certainty.”

Direct estimation of Proportions

- The words chosen, *S*, is a function of Document Category, *D*, —since the words chosen are by definition a function of the document category—it is simplest to use it directly. Thus, we have:

$$P(S = s) = \sum_{j=1}^{J} P(S = s \mid D = j) P(D = j)$$

- We observe P(S=s) and P(S=s | D = j)
  - \- if we assume it is stable across labelled and unlabelled data
- and now we can estimate P(D=j) using standard regression calculation:
- Let P(S | D) be X, P(S) is Y, and β is P(D).

$$Y = X\beta \text{ (with no error term)}$$

$$\beta = (X'X)^{-1}X'y$$

What is P(S=s) : Document representation

- To allow for estimation you need uncorrelated feature representation of the document.

  → Discrete word stem profiles (unordered sets of word stems).

Problems

- 1. The combinatorial space (i.e. no. of unique word stem profiles) is very large: 2^K $where\ K\ is\ no.of\ words$
- 2. Sparseness problem since the number of observations available for estimating P(S) and P(S|D) is much smaller than the number of potential word profiles (n<<2K)

What is P(S=s) : Document representation

- To allow for estimation you need uncorrelated feature representation of the document.

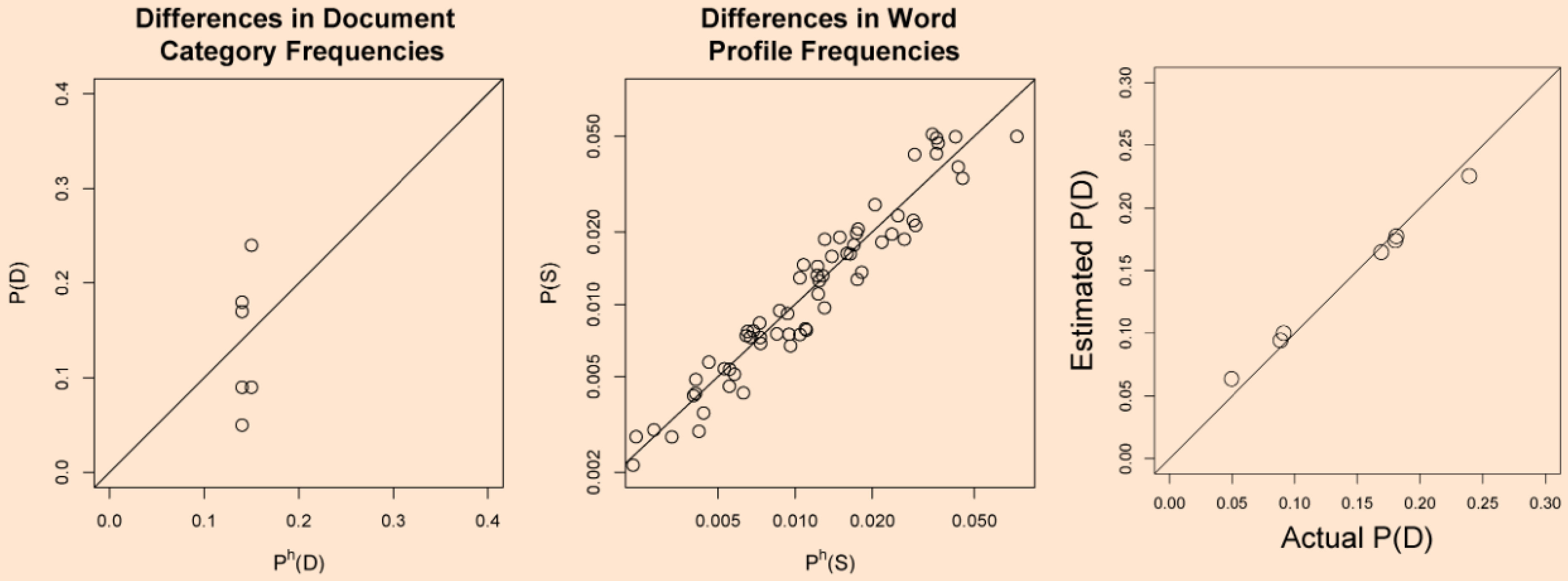
  → Discrete word stem profiles (unordered sets of word stems).

Solutions

- Subsample N words to form word stem profile distribution.
  - The combinatorial space (i.e. no. of unique word stem profiles) become: 2^N $e.g. N = 15;\ 2^N = 32768$
- Run the estimation *j* times and average over results.

It seems to work

FIGURE 2 **Accurate Estimates Despite Differences Between Labeled and Population Sets**



Notes: For both $P(D)$ on the left and $P(S)$ in the center, the distributions differ considerably. The direct sampling estimator, $P^h(D)$, is therefore highly biased. Yet, the right panel shows that our nonparametric estimator remains unbiased.

Jerzak, King and Strezhnev 2020: An Improved Method of Automated Nonparametric Content Analysis for Social Science

Following hurt performance

- "Emergent" and "Vanishing" discourse
  - i.e. P(S | D) not equal across labelled and unlabelled data
- Lack of *textual discrimination* between categories
- If discrimination is weak: Divergence between proportions in Labelled and Unlabelled set.

Improvements

- Change feature space to improve textual discrimination
  - Sentence representation based on GloVe X 3 i.e. (10^{th} ,50^{th} ,90^{th} percentile of each dimension)
- Preprocess feature space to avoid multicollinearity reducing statistical error arising from linear regression estimation.

Software: *Readme R-package (https://gking.harvard.edu/readme)*

Readme2: https://github.com/iqss-research/readme-software

Wiedemann 2018: Proportional Classification Revisited

- What about *differential bias* and *population and concept drift*?

  The assumption that P(S | D) is stable across labelled and unlabelled. .

  - *Time* changes the way language is used.
  - Different people use language differently (Party, Country)

  Simple Solution: Correct for each meta category:

    - Expensive because: N_categories*N_metacats*N_samples.

- Rare case → Classifier and "Direct estimator" cannot learn the differentiating features.
- Rare case → makes bias correction intractable. When cases are rare, N_samples needed for statistical certainty become large.

Proportional Classification Revisited

Readme / Direct estimation does not work for individual.

Table 4. Proportional Classification Performance (Hopkins & King, 2010).

	RMSD		Pearson's *r*
Code	Entire Test Set	Single Manifestos	Relative Proportions
504	.015	.179	.431
411	.015	.150	.546
501	.006	.160	.666
506	.017	.148	.596
605	.007	.124	.699
303	.011	.199	.511
706	.010	.108	.559
301	.004	.155	.564
107	.000	.094	.476
402	.004	.140	.502
Mean	.009	.146	.555

Note. RMSD = root mean square deviation.

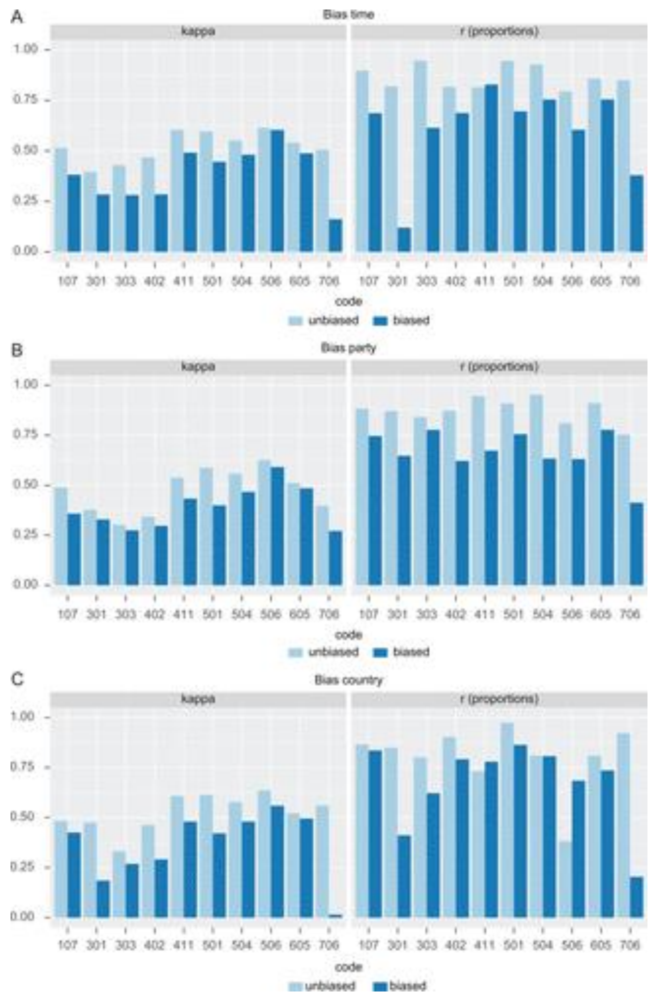
Table 5. Proportional Classification Performance Using Individual Classification (Logistic Regression).

	RMSD	RMSD	*r* (Frequencies)		*r* (Proportions)	
Code	Entire Training Set	Single Manifestos	Baseline	Absolute	Baseline	Relative
504	.0176	.0290	.807	.961	.0011	.865
411	.0022	.0220	.729	.970	−.0485	.926
501	.0135	.0326	.478	.956	.1105	.905
506	.0106	.0263	.796	.952	−.1775	.827
605	.0125	.0207	.709	.885	.4987	.854
303	.0079	.0253	.543	.933	.1750	.804
706	.0134	.0230	.581	.927	.2235	.808
301	.0010	.0205	.622	.907	−.0171	.821
107	.0045	.0142	.677	.883	.3356	.858
402	.0044	.0152	.522	.928	.0284	.869
Mean	.0088	.0229	.646	.930	.1130	.854

Note. RMSD = root mean square deviation.

Proportional Classification Revisited

Does not work when labelled population is different from unlabelled population

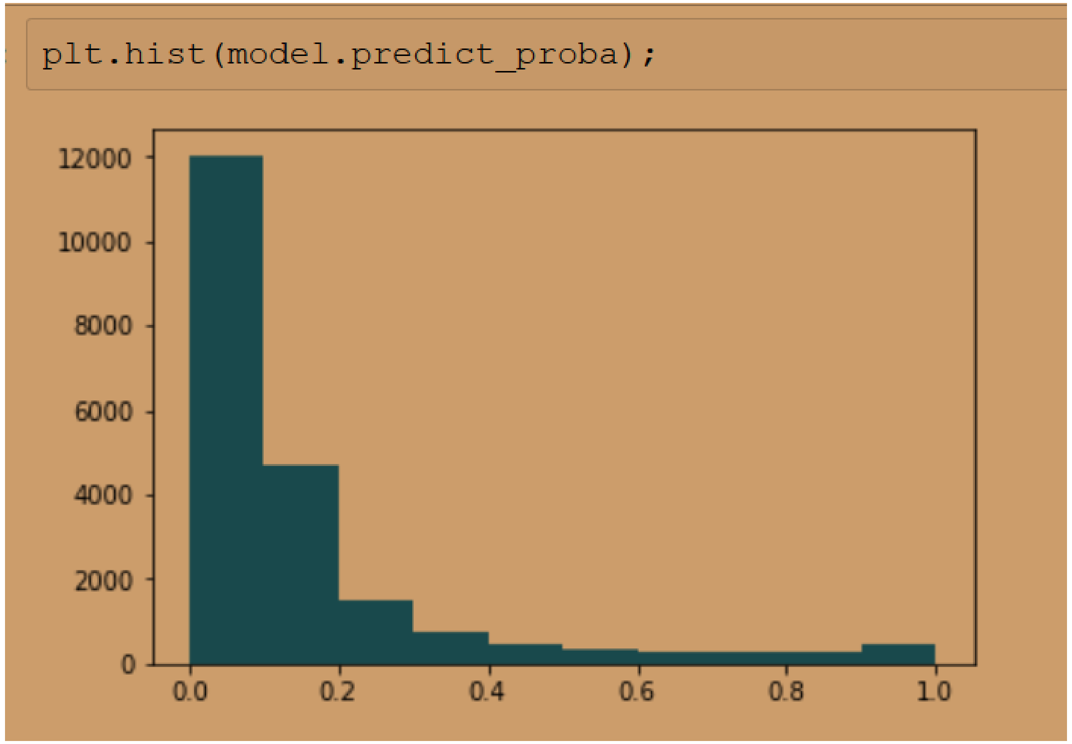


Wiedemann: Efficiently Train a Good Classifier

Active Learning for efficient training

"Let your classifier decide which cases it needs labelled"

- Uncertainty sampling: Pick samples that the classifier knows little about:
  - uncertainty = abs(model_probability-0.5)

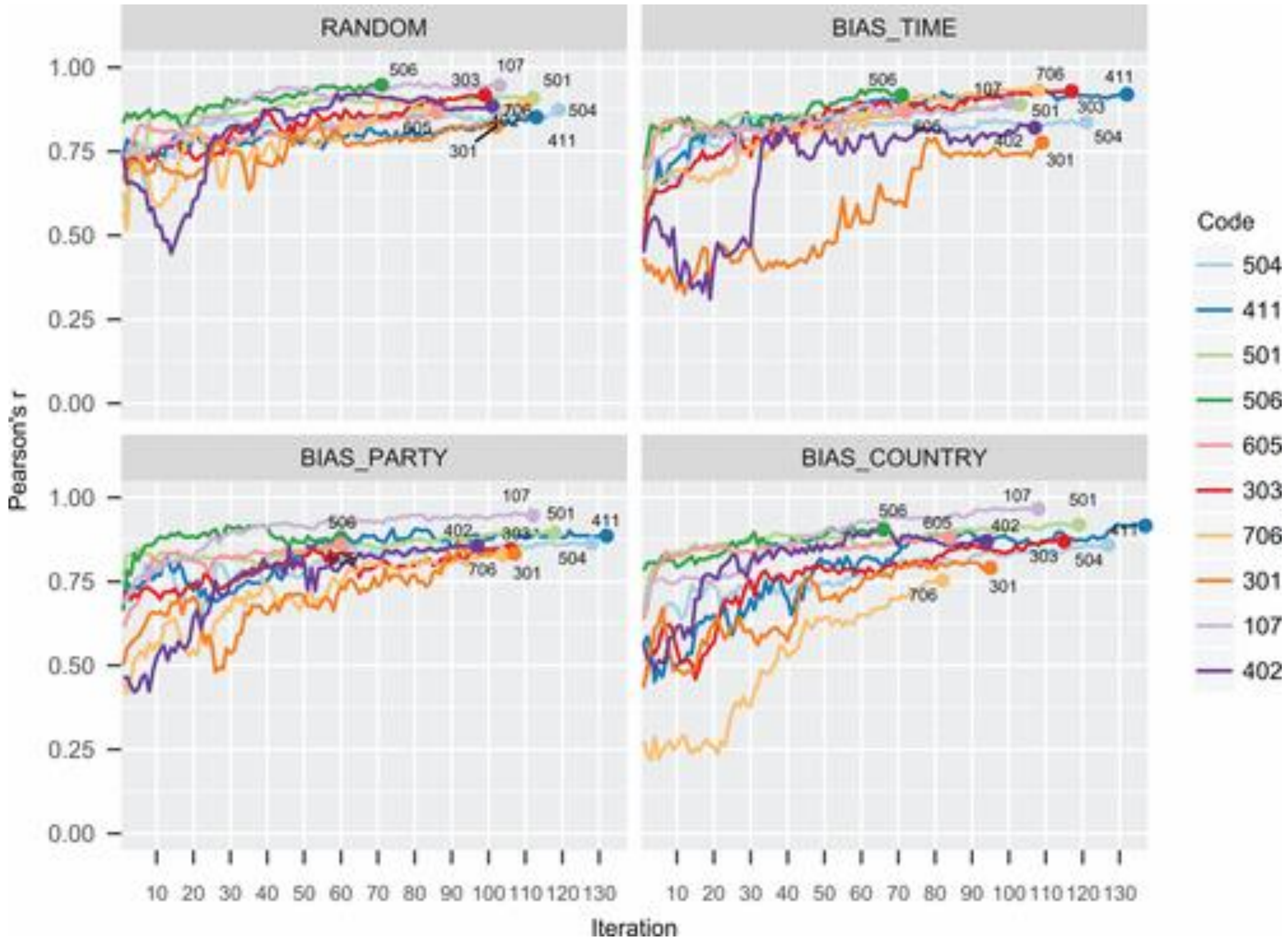


Active Learning for efficient training

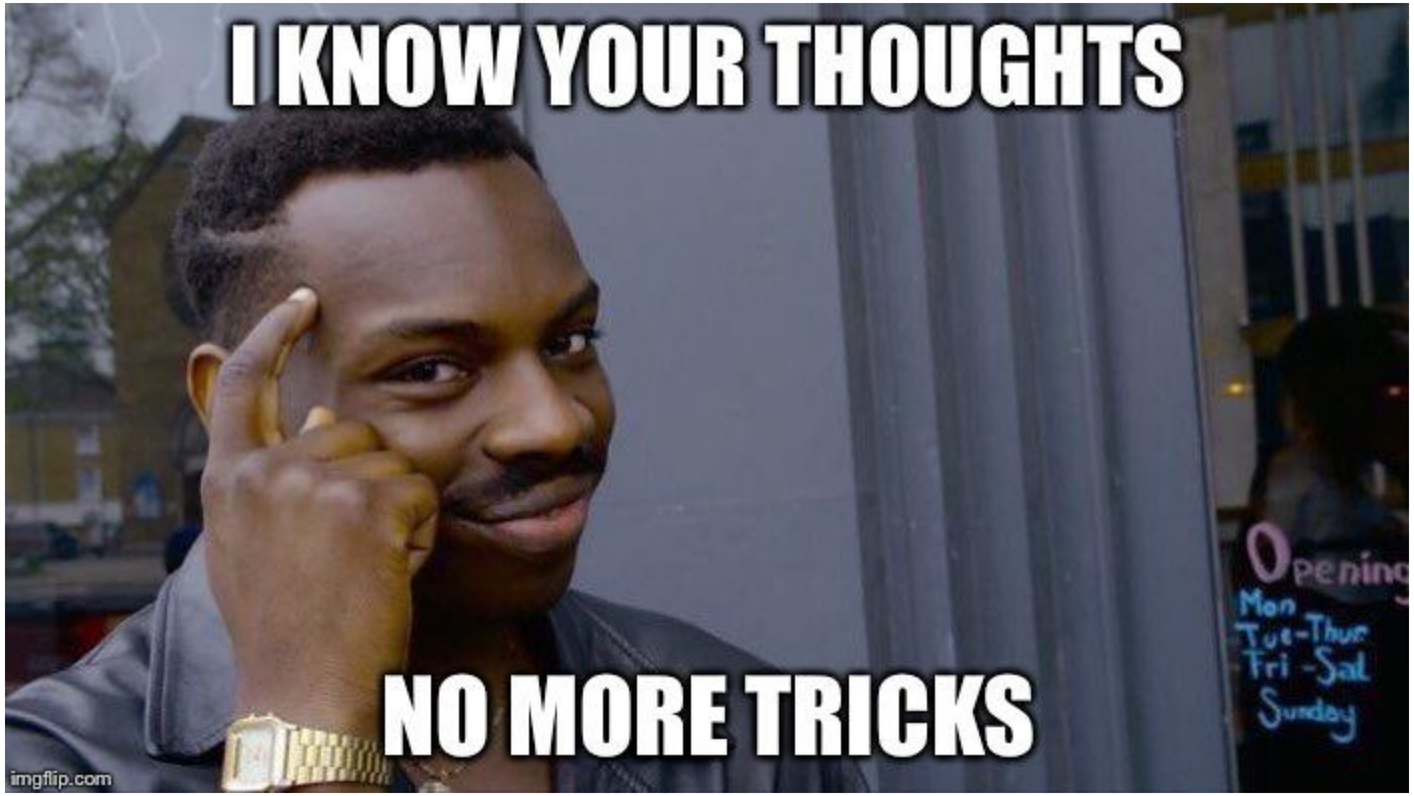
"Let your classifier decide which cases it needs labelled"

- Uncertainty sampling: Pick samples that the classifier knows little about:
  - uncertainty = abs(model_probability-0.5)
- Rare case mitigation: Weighted uncertainty sampling to increase no. of positive examples.
- Stop labelling when consecutive models produce almost identical predictions. Cohen Kappa = 0.99.

It also seems to Work?



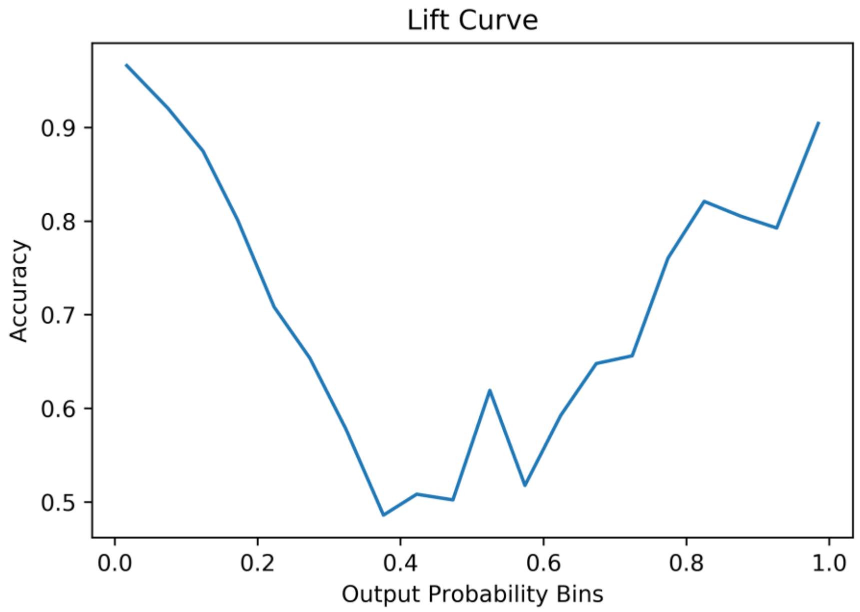
Digression: Semi-supervised learning for efficient training

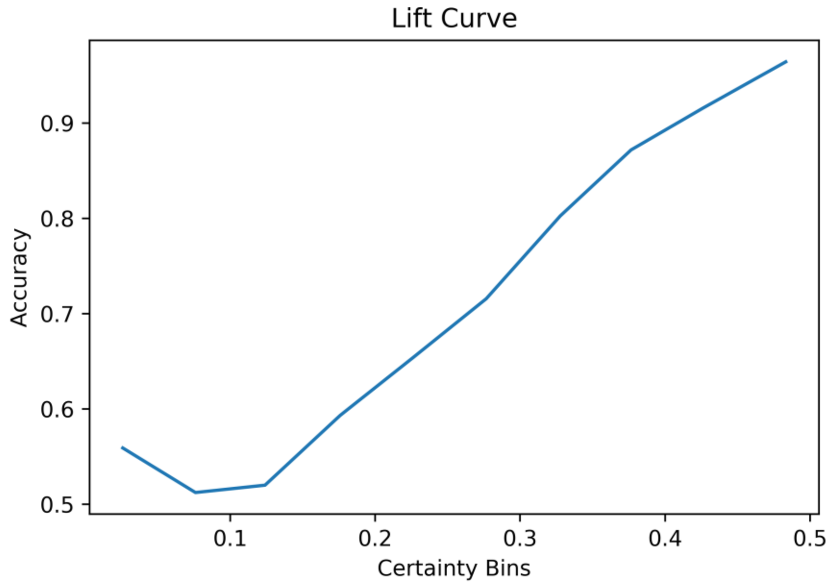


Digression: Semi-supervised learning for efficient training

"Let your classifier learn from itself."

- Semi-supervised learning as Weak-**self**-supervision

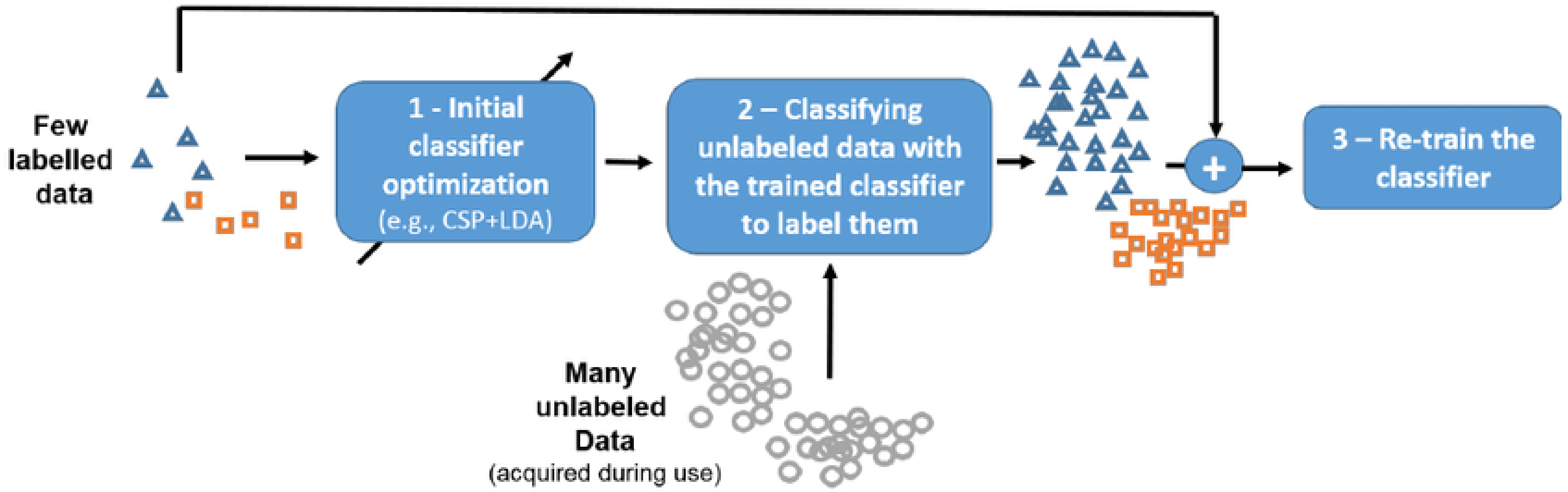




Digression: Semi-supervised learning for efficient training

"Let your classifier learn from itself."

- Semi-supervised learning as Weak-**self**-supervision
  - Instead of picking the most *uncertain* predictions for *manual* labelling, pick the most *certain* for *automatic* labelling.



Digression: Semi-supervised learning for efficient training

"Let your classifier learn from itself."

- Semi-supervised learning as Weak-**self**-supervision
  - Instead of picking the most *uncertain* predictions for *manual* labelling, pick the most *certain* for *automatic* labelling.
- *"you* shall know a *word* by *the company it keeps"*
  - Co-occurrence: Discover new variations by following patterns you already know.

Digression: Semi-supervised learning for efficient training

"Let your classifier learn from itself."

- Semi-supervised learning as Weak-**self**-supervision
  - Instead of picking the most *uncertain* predictions for *manual* labelling, pick the most *certain* for *automatic* labelling.
- *"you* shall know a *word* by *the company it keeps"*
  - Co-occurrence: Discover new variations by following patterns you already know.

Problems

- Degeneration and the incestuous model.
- *Stopping criterion: When performance drops on labelled (untouched! validation) sample.*

Criticisms

- Missing fair comparison: What is based on bad features and what is the method itself (cf. Jerzak et. al 2020)
- Wiedemann does not solve the problem of differential bias, just assumes that a saturated classifier is unbiased.
    - If the model used is essentially worse at modelling certain subgroups (e.g. sarcastic university students)

Future directions

- Wiedemann & Jerzak: Better discriminating models/ features means less error.

Future directions

Wiedemann & Jerzak: Better discriminating models/ features means less error.

- Resources released from applying tricks (active learning, semi-supervised learning, transfer learning, few shot learning) use for estimation of differential bias.

Future directions

Wiedemann & Jerzak: Better discriminating models/ features means less error.

- Resources released from applying tricks (active learning, semi-supervised learning, transfer learning, few shot learning) use for estimation of differential bias.
- Estimation of differential bias in rare case scenario still intractable.

----> We should figure out a way to estimate and correct for the differential bias efficiently.

Future directions

Concentration sampling

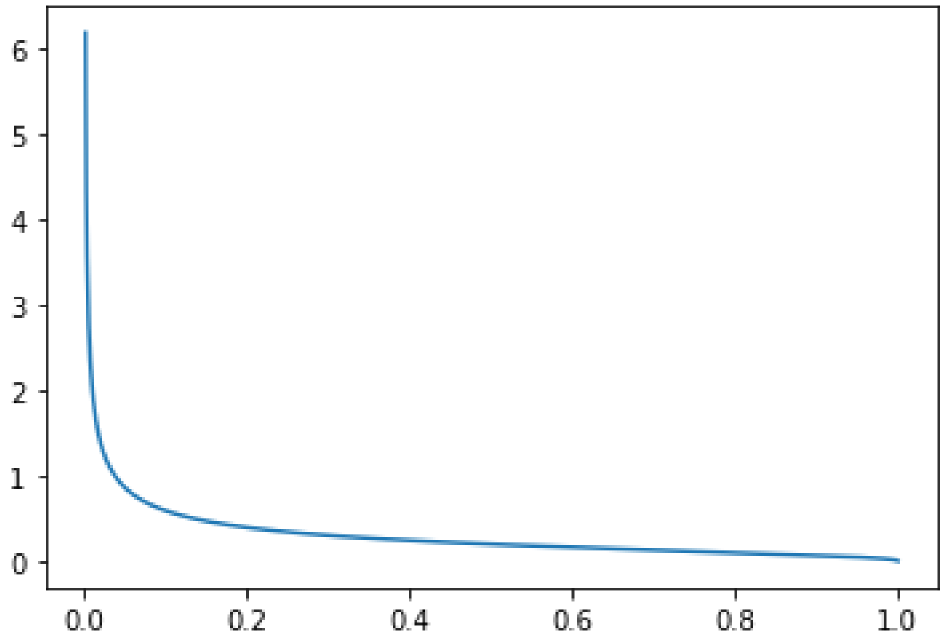
Concentration of probability will lower sampling cost.

$$X \sim B(p)$$

$$\mu = p$$

$$\sigma = p * (1 - p)$$

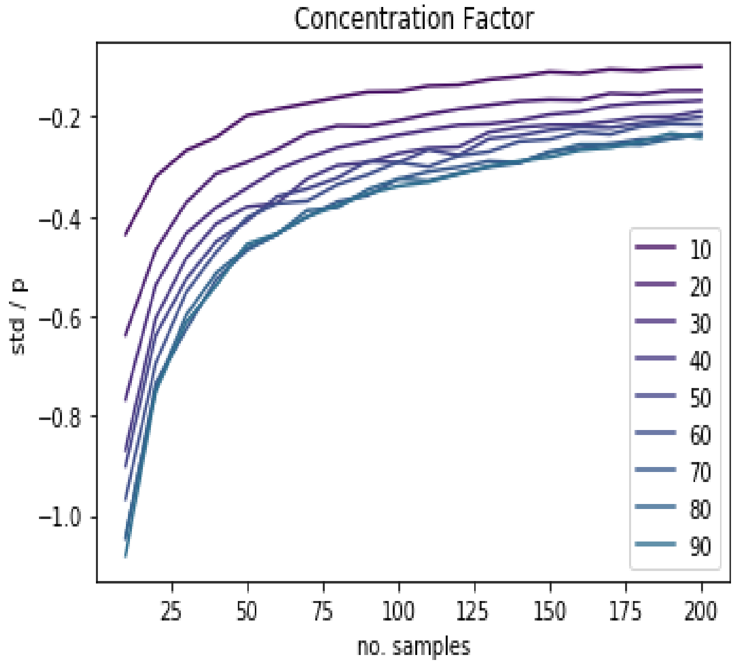
$$\textbf{Sampling distribution: } N(\mu, \frac{\sigma}{\sqrt{n}})$$



Future directions

Concentration sampling

Concentration of P(y=1) by partitioning into subgroups can lower sampling cost.

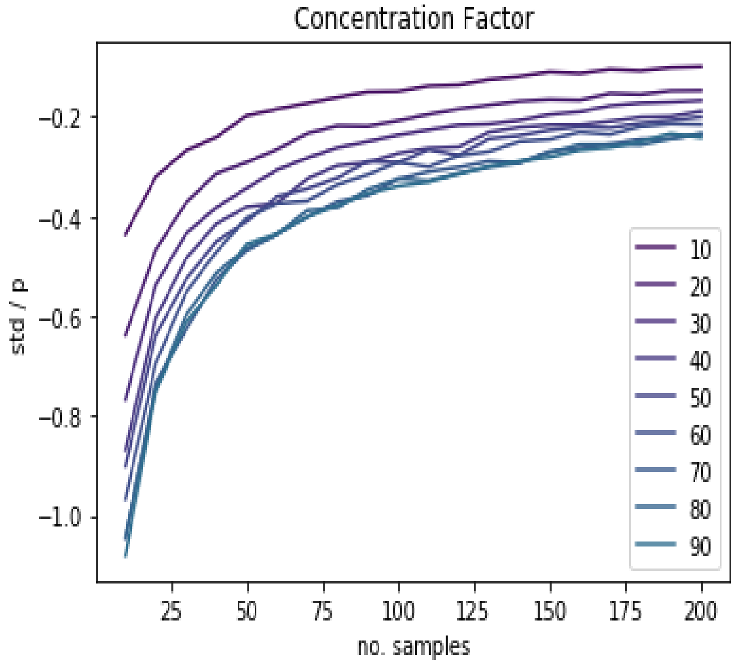


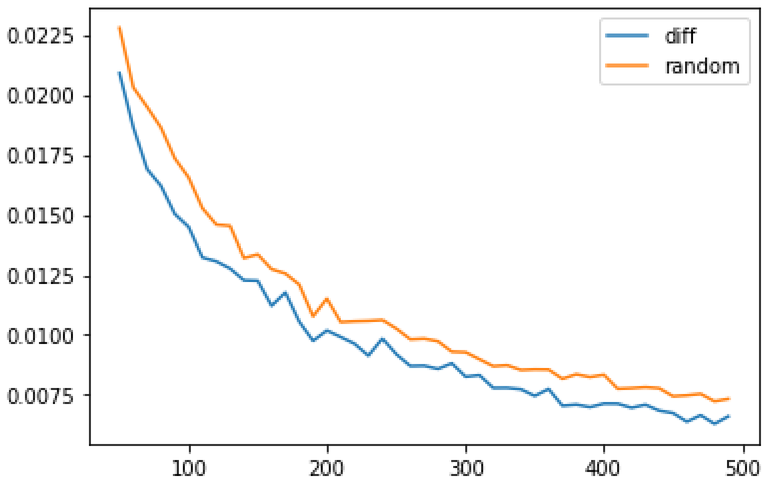
Future directions

Concentration sampling

Concentration of P(y=1) by partitioning into subgroups can lower sampling cost.

Concentration happens when sampling from the probability estimate of a discriminative





Thank you