Discussion: LLMs & Annotation
TADA 2023

Katherine A. Keith
Department of Computer Science
Williams College

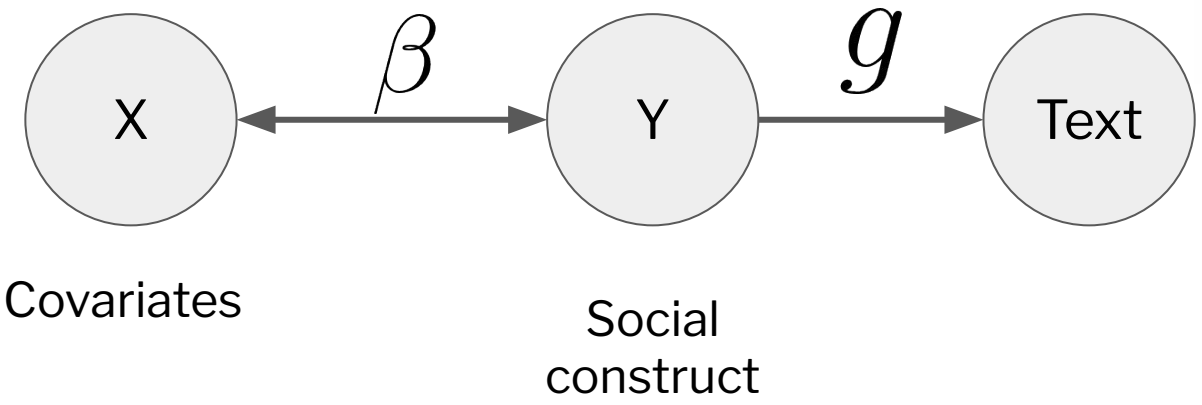
Surrogates - Egami et al.

Summary

Surrogates - Egami et al.

Summary

Surrogates - Egami et al.



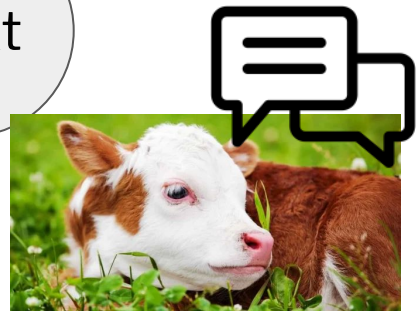
Summary

Surrogates - Egami et al.

X β Y g Text

Covariates

Social construct



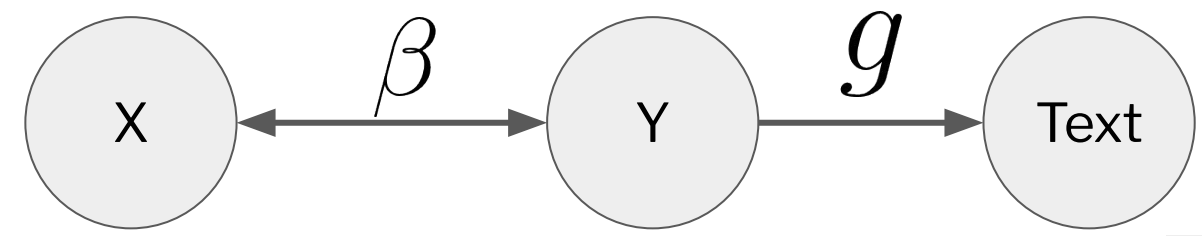
Cow Farmers:
Age, Gender, Income level

Positive/negative sentiment towards cows

Cow farmer online forums

Summary

Surrogates - Egami et al.



CI coverage, unbiased, efficiency

RQ: Given the goal is i*nferring* beta, how do we “best” estimate *g* from *both* gold-standard hand labels and predictions from a model (e.g, LLM)?

Summary

Surrogates - Egami et al.

Surrogates

$$\tilde{Y}_i := \underbrace{\hat{Y}_i}_{\text{Predicted Outcomes}} + \underbrace{\frac{R_i}{\pi_i}(Y_i - \hat{Y}_i)}_{\text{Bias-Correction Term}}$$

1 if i is gold
0 else

Gold-standard human-labels

Known probability of sampling for gold human labeling

Central to the **design-based** framework

$$\tilde{Y}_i = \begin{cases} \hat{Y}_i \text{ if not gold} \\ \frac{\pi_i - 1}{\pi_i}\hat{Y}_i - \frac{1}{\pi_i}Y_i \text{ if gold} \end{cases}$$

Interpolate between prediction and gold

Looks very similar to AIPTW from causal inference! (Robins et al., 1994)

Strengths

Surrogates - Egami et al.

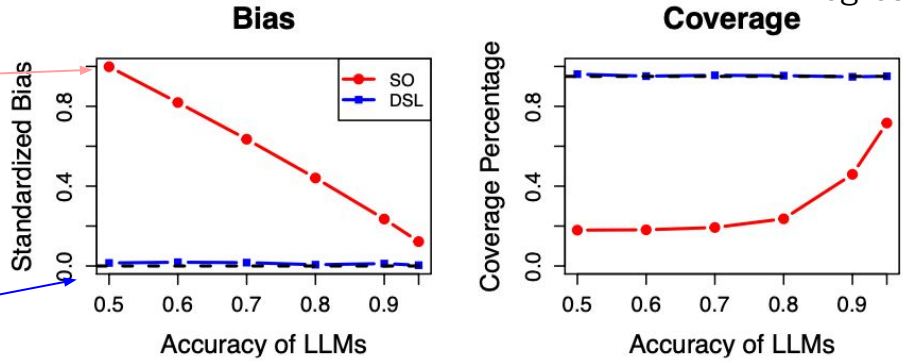
- Important problem! Combining measurement with inference.
- It works!

Bias in beta (downstream regression)

Coverage for CI of beta (downstream regression)

SO = surrogate only (LLM predictions)

DSL = design-based supervised learning (theirs)



Synthetic data

Discussion - Question 1

Surrogates - Egami et al.

$$\tilde{Y}_i = \begin{cases} \hat{Y}_i \text{ if not gold} \\ \frac{\pi_i - 1}{\pi_i}\hat{Y}_i - \frac{1}{\pi_i}Y_i \text{ if gold} \end{cases}$$

Why not use the “soft” probabilities (possibly after post-hoc calibration, e.g., Platt scaling)?

Discussion - Question 2

Surrogates - Egami et al.

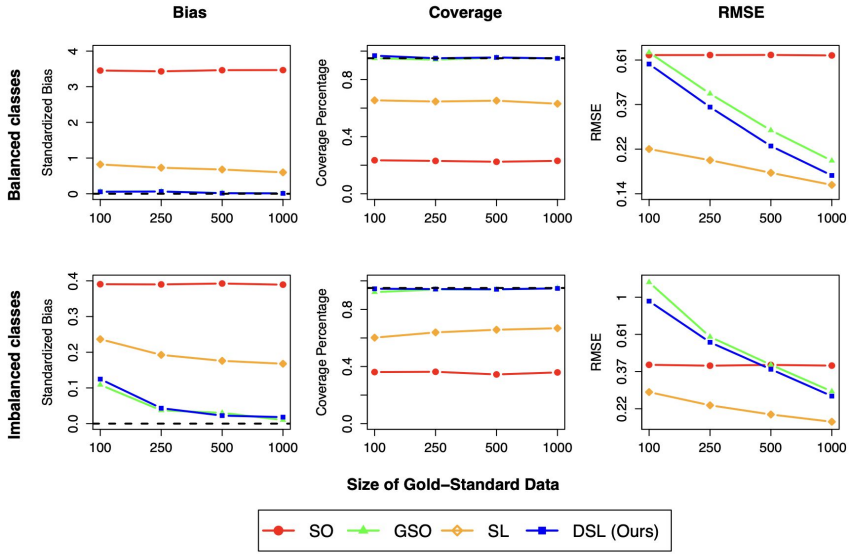


Figure 2: **Logistic regression estimation with Congressional Bills Project Data.** Results for

Here, only small gains in efficiency for DSL versus gold (all other metrics similar)

- Concrete recommendations for applied practitioners to use gold only?
- Other advantages of gold-only (interpretability)?

French-Language Subjectivity - Escouflaire et al.

Summary

French-Language Subjectivity - Escouflaire et al.

- **Data:** Web articles published by the RTBF (French language Belgian publication) 2008-2021
- **Categories (metadata):** Opinion versus news
- 36 student annotators
  - Subjectivity, 1-5
  - Confidence, 1-5
  - Token-level annotations

Summary

French-Language Subjectivity - Escouflaire et al.

Heatmap of human token-level annotations – indicators of “subjectivity”:

hardly surprising

amateurism

[1] Le gouvernement de Charles Michel est divisé avant un budget, rien d’étonnant en fait... Ce qui se passe est d’une banalité affligeante. On retrouve dans la séquence qui se déroule en ce moment une bonne partie des maux qui frappent la politique belge depuis au moins 15 ans, depuis le dernier gouvernement Dehaene. On retrouve des négociations marathons où telle taxe, telle coupe dans les soins de santé est décidée au bout de la nuit parce qu’il faut bien avoir quelque chose à livrer aux médias et au parlement. On retrouve le même empressement, le même amateurisme qui conduit des mesures importantes à s’écraser en plein vol faute de préparation.

distressingly banal

Strengths

French-Language Subjectivity - Escouflaire et al.

- Careful and rigorous annotation process (e.g., background of annotators)
- Non-English!
- **Paper**: Very strong interdisciplinary background section
- Found new indicators of subjectivity
  - Sequential discourse markers: *en fait* (“in fact”)
  - Adverbials: *hélas (“sadly”)*
  - Intensifiers and mitigators: presque (“almost”), peut-être (“maybe”)

Discussion - Question 1

French-Language Subjectivity - Escouflaire et al.

First step to select documents for annotating: Logistic regression classifier with 18 linguistic features

- Is the RQ about **discovering** new linguistic features of subjectivity?
- If so, why build in this linguistic inductive bias into the initially chosen docs?
- Related work in *lexicon induction*
  - Hamilton et al. “Inducing domain-specific sentiment lexicons from unlabeled corpora. “
  - Pryzant et al. “Deconfounded Lexicon Induction for Interpretable Social Science”

Discussion - Question 2

French-Language Subjectivity - Escouflaire et al.

- Exciting part = token-level annotations
- What was the inter-annotator agreement on the token-level?
- High-quality annotators: After, could you ask them *why* they chose these token-level indicators?

Annotation Validation - Pangakis et al.

Summary

Annotation Validation - Pangakis et al.

- Uses GPT-4 (zero-shot) to replicate 27 annotation tasks from 11 social science datasets
- Median accuracy 0.85

Author(s)	Title	Journal	Year
Gohdes	Repression Technology: Internet Accessibility and State Violence	American Journal of Political Science	2020
Hopkins, Lelkes, and Wolken	The Rise of and Demand for Identity-Oriented Media Coverage	American Journal of Political Science	2023
Schub	Informing the Leader: Bureaucracies and International Crises	American Political Science Review	2022
Busby, and Gubler, Hawkins	Framing and blame attribution in populist rhetoric	Journal of Politics	2019
Müller	The Temporal Focus of Campaign Communication	Journal of Politics	2021
Cusimano and Goodwin	People judge others to have more voluntary control over beliefs than they themselves do	Journal of Personality and Social Psychology	2020
Yu and Zhang	The Impact of Social Identity Conflict on Planning Horizons	Journal of Personality and Social Psychology	2022
Card et al.	Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration	PNAS	2022
Peng, Romero, and Horvat	Dynamics of cross-platform attention to retracted papers	PNAS	2022
Saha et al.	On the rise of fear speech in online social media	PNAS	2022
Wojcieszak et al.	Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity	Science Advances	2022

Table A1: Articles Replicated

Strengths

Annotation Validation - Pangakis et al.

- Authors took care in choosing datasets
  - “To avoid the potential for **contamination**, we rely exclusively on datasets stored in password-protected data archives …”
- Easy to find replication materials

Discussion - Question 1

Annotation Validation - Pangakis et al.

Authors introduce a **consistency score**: they vary the **temperature** of the LLM and calculate the proportion of classifications that match the mode.

$$\text{softmax}(z_i) = \frac{e^{z_i/T}}{\sum_{i'} e^{z_{i'}/T}}$$

Temperature (T)
Large T → more uniform distribution over tokens

Small T (T<1) → more softmax mass on token with the highest raw score

- Shouldn't temperature be set to almost 0 for classification?
- Instead, multiple runs with the same temperature with different random seeds?

Discussion - Question 2

Annotation Validation - Pangakis et al.

More details about the tasks (multi-label, multi-class, hierarchical classes)? Could you have multiple stages of prompts so that you could better diagnose (cascading) errors?

Prompt for Card et al.

'You are an expert in American immigration and classifying political speeches based on several categories. Return your classifications in a table with one column for text number (the number preceding each text sample) and a column for each category. Use a csv format. These are the categories to classify each text: cat_imm – Classify as 1 if the text makes a reference to immigrants, immigration, or immigration policy either explicitly or indirectly. If concepts related to immigration are mentioned (i.e., border, citizenship, homeland, foreign countries), they must be mentioned in the context of immigrants, immigration, or immigration policy; these words (i.e., border, citizenship, homeland, foreign countries) on their own are insufficient. References to another country, diversity, ethnic groups or nationalities (Hispanic, Asian, etc) without a clear connection to immigration do not cause cat_imm to be 1. A mention of a border state without a clear connection to immigration or immigration policy does not cause cat_imm to be 1. Classify as 0 otherwise. If you coded cat_imm as a 1, also classify the text's tone into one of three categories. If cat_imm is equal to 1, select just one of these three categories, scoring the other two categories as 0. If cat_imm is equal to 0, then set the other three categories to 0.cat_anti – If you coded cat_imm as a 1, classify cat_anti as 1 if the text argues for a significant increase in restrictions on immigration, or expresses a negative sentiment towards immigrants or immigration. Classify as 0 otherwise. cat_neutral - If you coded cat_imm as a 1, classify cat_neutral as 1 if the text is neutral, unclear, or if there is a mixture of positive and negative sentiments. Classify as 0 otherwise. cat_pro - If you coded cat_imm as a 1, classify cat_pro as 1 if the text is favorable towards immigrants or expresses preferences for continued or increased immigration, or expresses any type of positive sentiment to immigrants, immigration, or immigration policy. Unless the tone has a clear positive or negative attitude towards immigrants, it should be classified as cat_neutral. Classify as 0 otherwise. Classify the following text samples: "

https://dataverse.harvard.edu/file.xhtml?fileId=7452168&version=3.0

Q&A