Getting More out of Human Coders with Statistical Models

Matthew Tyler
Stanford University

May 22, 2020

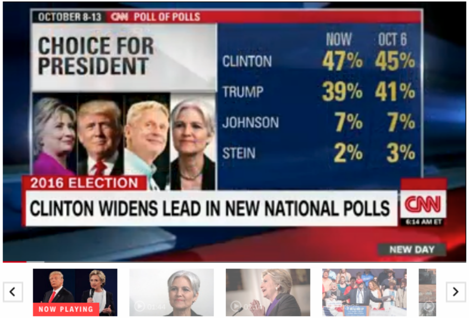
How different are the contents of pro-Democratic and pro-Republican news articles?

Labeling Election News Articles from 2016

- News articles from Peterson et al. (Forthcoming)
- 605 coders randomly assigned to articles
- Label slant: Pro-Democratic, Neutral, or Pro-Republican
- Estimate: relationship between article content and article slant
- 92% of articles: labeled by one coder

GOP sources: Trump still competing in Virginia

By Dana Bash
Updated 1:16 PM ET, Mon October 17, 2016



New Poll of Polls: Clinton widens lead
Jill Stein raises millions for recount
Scientists urge Clinton to call for recount
Donald Trump's wild ride
Trum the M

Washington (CNN) — After seeing reports last week that his campaign was pulling out of Virginia, Donald Trump picked up the phone and called his Virginia state director Mike Rubino to deliver a very clear message: he will not withdraw, and will give Rubino whatever resources needed to win the Old Dominion.

To back that up, CNN is told the Trump campaign plans to go on the air with television ads in Virginia starting Tuesday in what campaign officials say is a $2 million buy in "key markets" through election day. The campaign also announced a 19-member Virginia leadership team.

Standard Coding Practice

- Most objects: one label
- Some objects: multiple labels
- Compute coder agreement (intercoder reliability)
- Trust coders once agreement meets threshold ($\alpha \geq 0.7$)
- Analyze raw labels

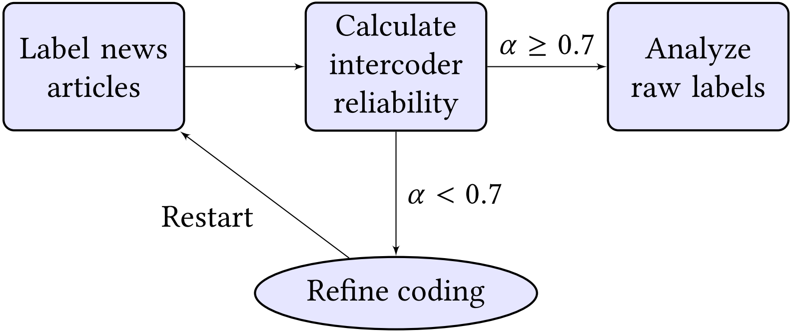
Problems:

- Coders err
- Measurement error (ME) on both sides of reliability threshold
- ME $\implies$ article proportions biased
- ME $\implies$ regressions biased

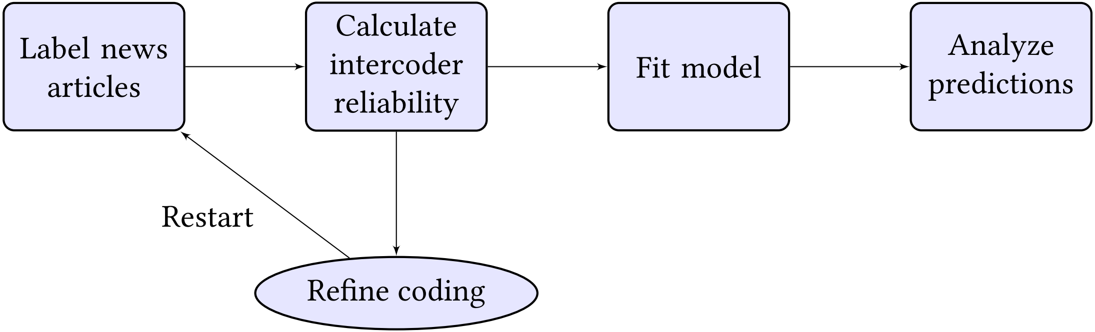
What To Do About It

- **Coder labels = noisy signal of the true label**
- Use model to identify more competent coders
- Model yields prediction E[True Label | Coder Labels]
- Predictions avoid ME problems; use as independent/dependent variable

Old Workflow: “Trust”



Proposed Workflow



Modeling Coder Labels: Core Idea

- Convert coder agreement $\longrightarrow$ coder accuracy
- When two coders agree, either both right or both wrong
- Two categories + two coders + same accuracy + independent decisions:

$$\text{Coder Agreement Rate} = (\text{Coder Accuracy})^2 + (1 - \text{Coder Accuracy})^2$$

- Agreement rate = 75% $\longrightarrow$ coder accuracy = 85%
- Relax assumptions, but always use agreement/accuracy relationship to identify model

Coder Competence Estimation (CCE)

- True label $z_i \overset{\text{ind.}}{\sim} \text{Categorical}(\boldsymbol{\lambda})$, proportion vector $\boldsymbol{\lambda} \in S^K$
- Coder competency parameter $c_j \in (0, 1)$
- Coder guessing parameter $\mathbf{g}_j \in S^K$ [c.f. MACE]
- Article i-coder j label is $w_{ij} \in \{1, \ldots, K\}$, drawn from:

$$s_{ij} \overset{\text{ind.}}{\sim} \text{Bernoulli}(c_j) \tag{1}$$

$$w_{ij} \mid s_{ij} = 1, z_i = k \overset{\text{ind.}}{\sim} \text{Categorical}(\boldsymbol{e}_k) \tag{2}$$

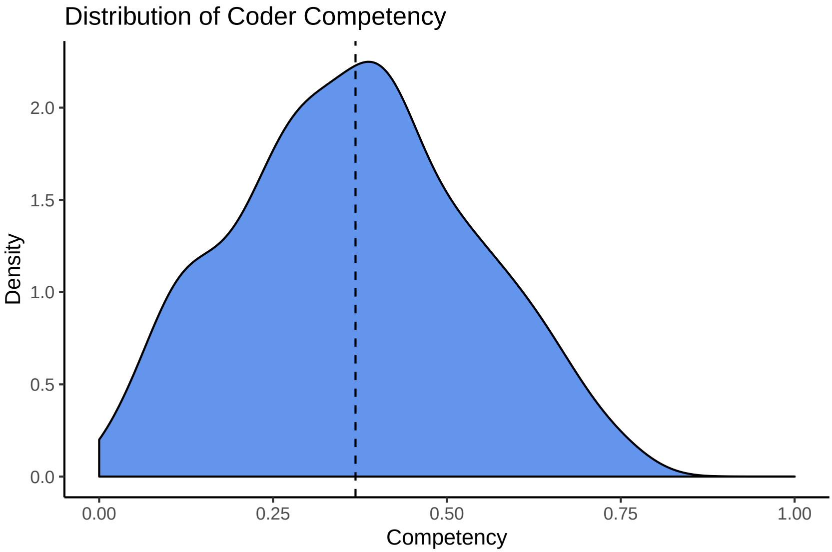
$$w_{ij} \mid s_{ij} = 0, z_i = k \overset{\text{ind.}}{\sim} \text{Categorical}(\mathbf{g}_j) \tag{3}$$

- If $s_{ij} = 1$, provides correct label; otherwise guesses
- $\therefore$ coders are correct $\implies$ coder agreement
- $\therefore$ competent coders agree with peers more frequently
- Upcoming paper: extend CCE further, provide identification results

Applying CCE to Election News Articles

- Recall: categorize articles as pro-Democratic, neutral, or pro-Republican
- Goal: estimate relationship between news content and slant
- Plan: fit CCE model with coder labels, regress content on predicted slant

Estimated Coder Competencies



Example Headlines

Straightforward examples:

- “Donald Trump’s Many Business Failures, Explained”
  P(Pro-Dem.) = 0.99, coders: (3 D)-(0 N)-(0 R)
- “Elected Democrat & Hillary Clinton Campaign Staffer SENT TO PRISON!”’
  P(Pro-Rep.) = 0.99, coders: (0 D)-(0 N)-(4 R)

Contentious example:

- “Donald Trump says soldiers with PTSD aren’t strong”
- Coders: (1 D)-(0 N)-(1 R), but model: P(Pro-Dem.) = 0.61, P(Neutral) = 0.36
- Pro-Democratic coder competency = 0.38
- Pro-Republican coder competency = 0.12
- Prediction captures uncertainty, favors more competent coder

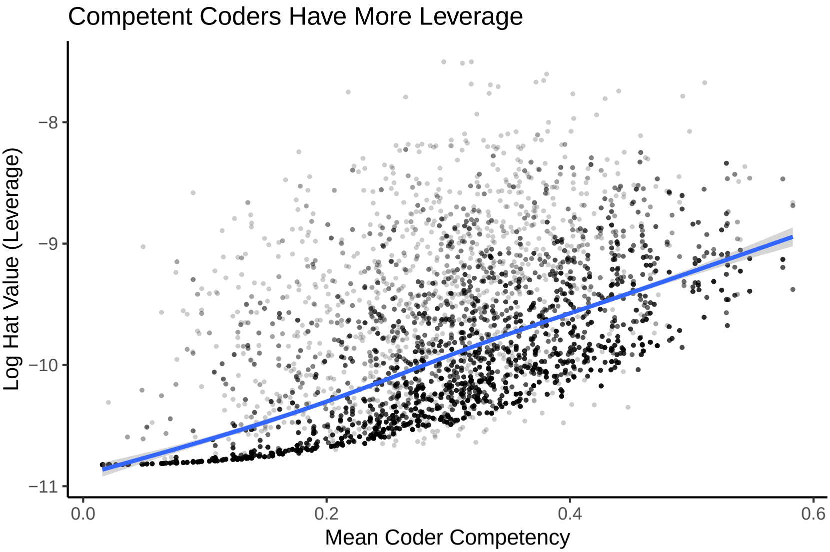
Downstream Regression: Topic Polarization

	Probability Headline Mentions			
	Trump	**Trump**	**Clinton**	**Clinton**
Pro-Rep.	-0.16^{**}	-0.58^{**}	0.16^{**}	0.36^{**}
	(0.03)	(0.02)	(0.03)	(0.02)
Neutral	-0.02	-0.42^{**}	-0.03^{**}	0.10^{**}
	(0.02)	(0.01)	(0.01)	(0.01)
(Intercept)	0.58^{**}	0.88^{**}	0.28^{**}	0.15^{**}
	(0.002)	(0.01)	(0.002)	(0.01)
Method	Trust	Prediction	Trust	Prediction
N	50,204	50,204	50,204	50,204

Downstream Regression: Topic Polarization

	Probability Headline Mentions			
	Trump	**Trump**	**Clinton**	**Clinton**
Pro-Rep.	-0.16^{**}	-0.58^{**}	0.16^{**}	0.36^{**}
	(0.03)	(0.02)	(0.03)	(0.02)
Neutral	-0.02	-0.42^{**}	-0.03^{**}	0.10^{**}
	(0.02)	(0.01)	(0.01)	(0.01)
(Intercept)	0.58^{**}	0.88^{**}	0.28^{**}	0.15^{**}
	(0.002)	(0.01)	(0.002)	(0.01)
Method	Trust	Prediction	Trust	Prediction
N	50,204	50,204	50,204	50,204

Why Does the Estimate Change so Much?



Conclusion

- Standard coding practice trusts the WRONG coders too much
- Reliability alone is insufficient
- Reliability measures average competency, but coders are heterogeneous
- Models elevate competent coders and adjust for uncertainty
- Replacing raw labels with predictions removes regression biases

Website: `https://stanford.edu/~mdtyler`
Twitter: `https://twitter.com/_mdtyler`