Using Machine Learning Models to Catch Mistakes in Coding of Open-ended Survey Questions

Gradon Nicholls, PhD Student, University of Waterloo
Advisor: Matthias Schonlau

CIPHER 2025
February 27, 2025

Outline

Closed- vs. Open-ended Survey Questions

People look for different things in a job. Which one of the following four things would you most prefer in a job?

- work that gives a feeling of accomplishment
- work where there is not too much supervision and you make most decisions yourself
- work that is pleasant and where the other people are nice to work with
- work that is steady with little chance of being laid off

Source: Schuman and Presser (1996, Ch. 3)

People look for different things in a job. What would you most prefer in a job?

[textbox]

Closed- vs. Open-ended Survey Questions

People look for different things in a job. Which one of the following four things would you most prefer in a job?

- work that gives a feeling of accomplishment
- work where there is not too much supervision and you make most decisions yourself
- work that is pleasant and where the other people are nice to work with
- work that is steady with little chance of being laid off

Source: Schuman and Presser (1996, Ch. 3)

People look for different things in a job. What would you most prefer in a job?

[textbox]

Coding Open-Ended Responses

We wish to assign a code (1,2,3,4) to each text response...

Text	Coder A		Coder B		Coder C	Final Code
response 1	1		1			
response 2	3		3			
response 3	2	$\neq$	1	$\rightarrow$	1	
response 4	2		2			
response 5	3	$\neq$	2	$\rightarrow$	3	
response 6	4		4			
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$

Coding Open-Ended Responses

...so we hire Coder A to manually code each text ("single-coding"):

Text	Coder A		Coder B		Coder C	Final Code
response 1	1		1			
response 2	3		3			
response 3	2	$\neq$	1	$\rightarrow$	1	
response 4	2		2			
response 5	3	$\neq$	2	$\rightarrow$	3	
response 6	4		4			
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$

Coding Open-Ended Responses

Would like to improve code quality by finding and revising possible mistakes.

Text	Coder A		Coder B		Coder C	Final Code
response 1	1		1			
response 2	3		3			
response 3	2	$\neq$	1	$\rightarrow$	1	
response 4	2		2			
response 5	3	$\neq$	2	$\rightarrow$	3	
response 6	4		4			
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$

Coding Open-Ended Responses

Hire Coder B to (independently) code the same texts ("double-coding"):

Text	Coder A		Coder B		Coder C	Final Code
response 1	1		1			
response 2	3		3			
response 3	2	$\neq$	1	$\rightarrow$	1	
response 4	2		2			
response 5	3	$\neq$	2	$\rightarrow$	3	
response 6	4		4			
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$

Coding Open-Ended Responses

Differences between coders indicate possible mistakes:

Text	Coder A		Coder B		Coder C	Final Code
response 1	1		1			
response 2	3		3			
response 3	2	$\neq$	1	$\rightarrow$	1	
response 4	2		2			
response 5	3	$\neq$	2	$\rightarrow$	3	
response 6	4		4			
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$

Coding Open-Ended Responses

Differences resolved by expert coder:

Text	Coder A		Coder B		Coder C	Final Code
response 1	1		1			
response 2	3		3			
response 3	2	$\neq$	1	$\rightarrow$	1	
response 4	2		2			
response 5	3	$\neq$	2	$\rightarrow$	3	
response 6	4		4			
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$

Coding Open-Ended Responses

Final "gold-standard" codes:

Text	Coder A		Coder B		Coder C	Final Code
response 1	1		1			1
response 2	3		3			3
response 3	2	$\neq$	1	$\rightarrow$	1	1
response 4	2		2			2
response 5	3	$\neq$	2	$\rightarrow$	3	3
response 6	4		4			4
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$

More on Double-Coding

Double-coding a **small subset** of texts is sometimes done to assess validity/reliability of coding (Spooren and Degand, 2010)

If single-coders are "good enough", more cost-effective to pay for additional single-coding rather than double-coding (He and Schonlau, 2020)

Conclusion: typically expect to have few or no double-coded texts.

$\Longrightarrow$ motivates us to use a model to "simulate" a second coder

More on Double-Coding

Double-coding a **small subset** of texts is sometimes done to assess validity/reliability of coding (Spooren and Degand, 2010)

If single-coders are "good enough", more cost-effective to pay for additional single-coding rather than double-coding (He and Schonlau, 2020)

Conclusion: typically expect to have few or no double-coded texts.

$\Longrightarrow$ motivates us to use a model to "simulate" a second coder

More on Double-Coding

Double-coding a **small subset** of texts is sometimes done to assess validity/reliability of coding (Spooren and Degand, 2010)

If single-coders are "good enough", more cost-effective to pay for additional single-coding rather than double-coding (He and Schonlau, 2020)

Conclusion: typically expect to have few or no double-coded texts.

$\implies$ motivates us to use a model to "simulate" a second coder

More on Double-Coding

Double-coding a **small subset** of texts is sometimes done to assess validity/reliability of coding (Spooren and Degand, 2010)

If single-coders are "good enough", more cost-effective to pay for additional single-coding rather than double-coding (He and Schonlau, 2020)

Conclusion: typically expect to have few or no double-coded texts.

$\implies$ motivates us to use a model to "simulate" a second coder

The Second Coder as a Mistake Classifier

With Coder A as baseline, Coder B is a classifier and Coder C is the "ground truth":

Text	Coder A	Coder B	Coder C
response 1	1	No mistake	No mistake
response 2	3	No mistake	No mistake
response 3	2	Mistake	Mistake
response 4	2	No mistake	No mistake
response 5	3	Mistake	No mistake
response 6	4	No mistake	No mistake
⋮	⋮	⋮	⋮

The Second Coder as a Mistake Classifier

With Coder A as baseline, Coder B is a classifier and Coder C is the "ground truth":

Text	Coder A	Coder B	Coder C
response 1	1	No mistake (TN)	No mistake
response 2	3	No mistake (TN)	No mistake
response 3	2	Mistake (TP)	Mistake
response 4	2	No mistake (TN)	No mistake
response 5	3	Mistake (FP)	No mistake
response 6	4	No mistake (TN)	No mistake
⋮	⋮	⋮	⋮

"Else-Trained" Approach (He and Schonlau, 2022)

Assume we have double-coded data:

1. use Coder B's codes to train model to predict $\mathbf{P}(Y_i^B = k \mid x_i)$
   - Y_i^B: Coder B's code (unobserved)
   - x_i: text
2. compute score $s_i \equiv \mathbf{P}(Y_i^B \neq y_i^A \mid x_i) = 1 - \mathbf{P}(Y_i^B = y_i^A \mid x_i)$
   - y_i^A: Coder A's code (observed)
3. classify as a mistake if $s_i \geq \tau$, $\tau \in (0, 1)$

"Else-Trained" Approach (He and Schonlau, 2022)

Assume we have double-coded data:

1. use Coder B's codes to train model to predict $\mathbf{P}\left(Y_i^B = k \mid x_i\right)$
   - Y_i^B: Coder B's code (unobserved)
   - x_i: text
2. compute score $s_i \equiv \mathbf{P}\left(Y_i^B \neq y_i^A \mid x_i\right) = 1 - \mathbf{P}\left(Y_i^B = y_i^A \mid x_i\right)$
   - y_i^A: Coder A's code (observed)
3. classify as a mistake if $s_i \geq \tau$, $\tau \in (0, 1)$

"Else-Trained" Approach (He and Schonlau, 2022)

Assume we have double-coded data:

1. use Coder B's codes to train model to predict $\mathbf{P}\left(Y_i^B = k \mid x_i\right)$
   - Y_i^B: Coder B's code (unobserved)
   - x_i: text
2. compute score $s_i \equiv \mathbf{P}\left(Y_i^B \neq y_i^A \mid x_i\right) = 1 - \mathbf{P}\left(Y_i^B = y_i^A \mid x_i\right)$
   - y_i^A: Coder A's code (observed)
3. classify as a mistake if $s_i \geq \tau$, $\tau \in (0, 1)$

"Else-Trained" Approach (He and Schonlau, 2022)

Assume we have double-coded data:

1. use Coder B's codes to train model to predict $\mathbf{P}\left(Y_i^B = k \mid x_i\right)$
   - Y_i^B: Coder B's code (unobserved)
   - x_i: text
2. compute score $s_i \equiv \mathbf{P}\left(Y_i^B \neq y_i^A \mid x_i\right) = 1 - \mathbf{P}\left(Y_i^B = y_i^A \mid x_i\right)$
   - y_i^A: Coder A's code (observed)
3. classify as a mistake if $s_i \geq \tau$, $\tau \in (0, 1)$

"Self-Trained" Approach

Assume we have **only** single-coded data:

1. use Coder A's codes to train model to predict $\mathbf{P}(Y_i^A = k \mid x_i)$
   - Y_i^A: Coder A's code (unobserved)
   - x_i: text
2. compute score $s_i \equiv \mathbf{P}(Y_i^A \neq y_i^A \mid x_i) = 1 - \mathbf{P}(Y_i^A = y_i^A \mid x_i)$
   - y_i^A: Coder A's code (observed)
3. classify as a mistake if $s_i \geq \tau$, $\tau \in (0, 1)$

"Self-Trained" Approach

Assume we have **only** single-coded data:

1. use Coder A's codes to train model to predict $\mathbf{P}\left(Y_i^A = k \mid x_i\right)$
   - Y_i^A: Coder A's code (unobserved)
   - x_i: text
2. compute score $s_i \equiv \mathbf{P}\left(Y_i^A \neq y_i^A \mid x_i\right) = 1 - \mathbf{P}\left(Y_i^A = y_i^A \mid x_i\right)$
   - y_i^A: Coder A's code (observed)
3. classify as a mistake if $s_i \geq \tau$, $\tau \in (0, 1)$

"Self-Trained" Approach

Assume we have **only** single-coded data:

1. use Coder A's codes to train model to predict $\mathbf{P}\left(Y_i^A = k \mid x_i\right)$
   - Y_i^A: Coder A's code (unobserved)
   - x_i: text
2. compute score $s_i \equiv \mathbf{P}\left(Y_i^A \neq y_i^A \mid x_i\right) = 1 - \mathbf{P}\left(Y_i^A = y_i^A \mid x_i\right)$
   - y_i^A: Coder A's code (observed)
3. classify as a mistake if $s_i \geq \tau$, $\tau \in (0, 1)$

"Self-Trained" Approach

Assume we have **only** single-coded data:

1. use Coder A's codes to train model to predict $\mathbf{P}\left(Y_i^A = k \mid x_i\right)$
   - Y_i^A: Coder A's code (unobserved)
   - x_i: text
2. compute score $s_i \equiv \mathbf{P}\left(Y_i^A \neq y_i^A \mid x_i\right) = 1 - \mathbf{P}\left(Y_i^A = y_i^A \mid x_i\right)$
   - y_i^A: Coder A's code (observed)
3. classify as a mistake if $s_i \geq \tau$, $\tau \in (0, 1)$

Notes on Else-Trained vs. Self-Trained

else-trained: use B's codes to train a model to catch mistakes in A's codes
self-trained: use A's codes to train a model to catch mistakes in A's codes

approaches differ only in choice of training data
$\implies$ if Coders A and B have identical behaviour, two approaches will be the same

can view self-trained as outlier detection (Hendrycks and Gimpel, 2016)
i.e. $\mathbf{P}\left(Y_i^A \neq y_i^A \mid x_i\right)$ is a measure of how poorly the model fits data point i

Notes on Else-Trained vs. Self-Trained

else-trained: use B's codes to train a model to catch mistakes in A's codes
self-trained: use A's codes to train a model to catch mistakes in A's codes

approaches differ only in choice of training data
$\implies$ if Coders A and B have identical behaviour, two approaches will be the same

can view self-trained as outlier detection (Hendrycks and Gimpel, 2016)
i.e. $\mathbf{P}\left(Y_i^A \neq y_i^A \mid x_i\right)$ is a measure of how poorly the model fits data point i

Notes on Else-Trained vs. Self-Trained

else-trained: use B's codes to train a model to catch mistakes in A's codes
self-trained: use A's codes to train a model to catch mistakes in A's codes

approaches differ only in choice of training data
$\implies$ if Coders A and B have identical behaviour, two approaches will be the same

can view self-trained as outlier detection (Hendrycks and Gimpel, 2016)
i.e. $\mathbf{P}\left(Y_i^A \neq y_i^A \mid x_i\right)$ is a measure of how poorly the model fits data point i

Notes on Else-Trained vs. Self-Trained

else-trained: use B's codes to train a model to catch mistakes in A's codes
self-trained: use A's codes to train a model to catch mistakes in A's codes

approaches differ only in choice of training data
$\implies$ if Coders A and B have identical behaviour, two approaches will be the same

can view self-trained as outlier detection (Hendrycks and Gimpel, 2016)
i.e. $\mathbf{P}\left(Y_i^A \neq y_i^A \mid x_i\right)$ is a measure of how poorly the model fits data point i

Notes on Else-Trained vs. Self-Trained

else-trained: use B's codes to train a model to catch mistakes in A's codes
self-trained: use A's codes to train a model to catch mistakes in A's codes

approaches differ only in choice of training data
$\implies$ if Coders A and B have identical behaviour, two approaches will be the same

can view self-trained as outlier detection (Hendrycks and Gimpel, 2016)
i.e. $\mathbf{P}\left(Y_i^A \neq y_i^A \mid x_i\right)$ is a measure of how poorly the model fits data point i

Notes on Else-Trained vs. Self-Trained

else-trained: use B's codes to train a model to catch mistakes in A's codes
self-trained: use A's codes to train a model to catch mistakes in A's codes

approaches differ only in choice of training data
$\implies$ if Coders A and B have identical behaviour, two approaches will be the same

can view self-trained as outlier detection (Hendrycks and Gimpel, 2016)
i.e. $\mathbf{P}\left(Y_i^A \neq y_i^A \mid x_i\right)$ is a measure of how poorly the model fits data point i

Text Classification

Two methods to estimate scores:

1. Support Vector Machines, as used in He and Schonlau (2022)
   - low computational cost
   - easy to code and understand
2. BERTje, dutch-language version of BERT
   - high computational cost
   - steep learning curve

Text Classification

Two methods to estimate scores:

1. Support Vector Machines, as used in He and Schonlau (2022)
   - low computational cost
   - easy to code and understand
2. BERTje, dutch-language version of BERT
   - high computational cost
   - steep learning curve

Text Classification

Two methods to estimate scores:

1. Support Vector Machines, as used in He and Schonlau (2022)
   - low computational cost
   - easy to code and understand
2. BERTje, dutch-language version of BERT
   - high computational cost
   - steep learning curve

Data: Patient Joe

Open-ended question fielded in Dutch as a web survey in the LISS panel in 2012 (Martin et al., 2011):

> *Joe's doctor told him that he would need to return in two weeks to find out whether or not his condition had improved. But when Joe asked the receptionist for an appointment, he was told that it would be over a month before the next available appointment. What should Joe do?*

$n = 1756$ texts double-coded into 4 categories:

1. Proactive
2. Somewhat proactive
3. Passive
4. Destructive

Data: Patient Joe

Open-ended question fielded in Dutch as a web survey in the LISS panel in 2012 (Martin et al., 2011):

> *Joe's doctor told him that he would need to return in two weeks to find out whether or not his condition had improved. But when Joe asked the receptionist for an appointment, he was told that it would be over a month before the next available appointment. What should Joe do?*

$n = 1756$ texts double-coded into 4 categories:

1. Proactive
2. Somewhat proactive
3. Passive
4. Destructive

Simulation Study

- split data into 1000 training, 756 test
- simulations $s = 1, 2, \ldots, 100$
- training sizes $m = 100, 200, \ldots 600$
- text classification method $\mathcal{M} = \text{SVM}, \text{BERTje}$
- for each s, m, $\mathcal{M}$
  - sample m texts from training set without replacement
    - B's codes $\rightarrow$ else-trained
    - A's codes $\rightarrow$ self-trained

Simulation Study

- split data into 1000 training, 756 test
- simulations $s = 1, 2, \ldots, 100$
- training sizes $m = 100, 200, \ldots 600$
- text classification method $\mathcal{M} = \text{SVM}, \text{BERTje}$
- for each s, m, $\mathcal{M}$
  - sample m texts from training set without replacement
    - B's codes $\rightarrow$ else-trained
    - A's codes $\rightarrow$ self-trained

Simulation Study

- split data into 1000 training, 756 test
- simulations $s = 1, 2, \ldots, 100$
- training sizes $m = 100, 200, \ldots 600$
- text classification method $\mathcal{M} = \text{SVM}, \text{BERTje}$
- for each s, m, $\mathcal{M}$
  - sample m texts from training set without replacement
    - B's codes $\rightarrow$ else-trained
    - A's codes $\rightarrow$ self-trained

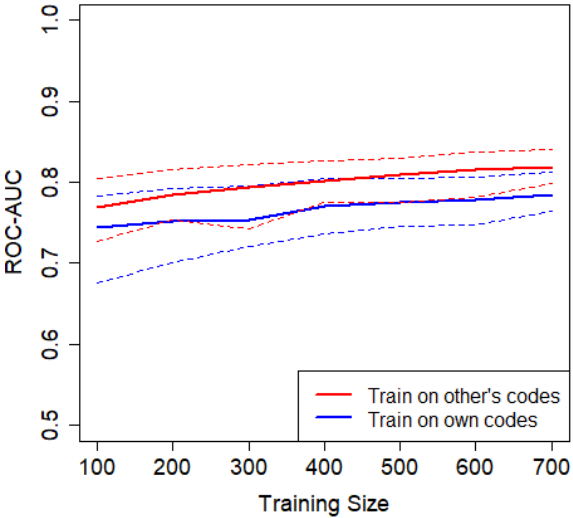
Simulation Study

- split data into 1000 training, 756 test
- simulations $s = 1, 2, \ldots, 100$
- training sizes $m = 100, 200, \ldots 600$
- text classification method $\mathcal{M} = \text{SVM}, \text{BERTje}$
- for each s, m, $\mathcal{M}$
  - sample m texts from training set without replacement
    - B's codes $\rightarrow$ else-trained
    - A's codes $\rightarrow$ self-trained

Simulation Study

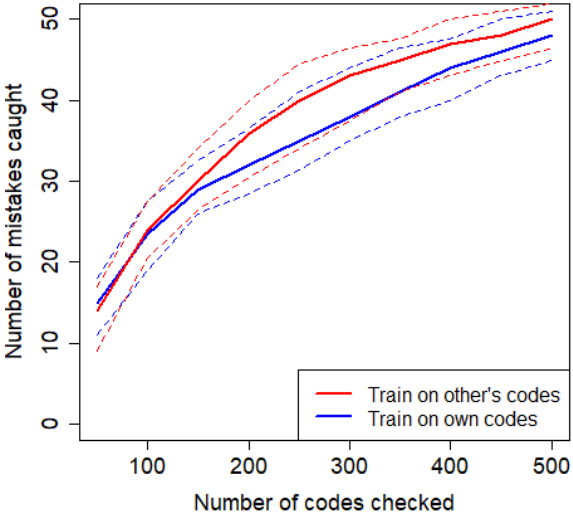
- split data into 1000 training, 756 test
- simulations $s = 1, 2, \ldots, 100$
- training sizes $m = 100, 200, \ldots 600$
- text classification method $\mathcal{M} = \text{SVM}, \text{BERTje}$
- for each s, m, $\mathcal{M}$
  - sample m texts from training set without replacement
    - B's codes $\rightarrow$ else-trained
    - A's codes $\rightarrow$ self-trained

Results: ROC-AUC

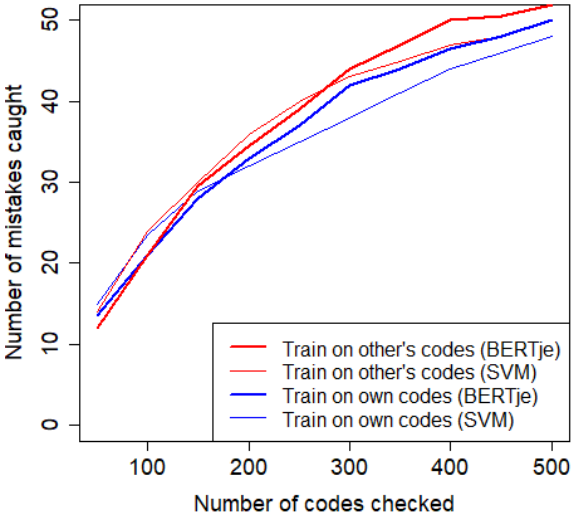


Results: ROC-AUC (SVM vs. BERTje)

Result: Number of Mistakes Caught



Result: Number of Mistakes Caught (SVM vs. BERTje)



Summary

- for a given training size, else-trained approach wins, but not by much!
- $\rightarrow$ shows that catching mistakes is feasible even with only single-coded data
- $\rightarrow$ especially so, considering we expect single-coded data to be much more prevalent
- large language models did not improve mistake-catching, at least for small training sizes

Summary

- for a given training size, else-trained approach wins, but not by much!
- $\rightarrow$ shows that catching mistakes is feasible even with only single-coded data
- $\rightarrow$ especially so, considering we expect single-coded data to be much more prevalent
- large language models did not improve mistake-catching, at least for small training sizes

Summary

- for a given training size, else-trained approach wins, but not by much!
- $\rightarrow$ shows that catching mistakes is feasible even with only single-coded data
- $\rightarrow$ especially so, considering we expect single-coded data to be much more prevalent
- large language models did not improve mistake-catching, at least for small training sizes

Summary

- for a given training size, else-trained approach wins, but not by much!
- $\rightarrow$ shows that catching mistakes is feasible even with only single-coded data
- $\rightarrow$ especially so, considering we expect single-coded data to be much more prevalent
- large language models did not improve mistake-catching, at least for small training sizes

Thank you!

Thank you!

References I

He, Zhoushanyue and Matthias Schonlau. 2020. Automatic coding of open-ended questions into multiple classes: Whether and how to use double coded data. In *Survey Research Methods*. Vol. 14 pp. 267–287.

He, Zhoushanyue and Matthias Schonlau. 2022. "A model-assisted approach for finding coding errors in manual coding of open-ended questions." *Journal of Survey Statistics and Methodology* 10(2):365–376.

Hendrycks, Dan and Kevin Gimpel. 2016. "A baseline for detecting misclassified and out-of-distribution examples in neural networks." *arXiv preprint arXiv:1610.02136* .

Martin, Laurie T, Matthias Schonlau, Ann Haas, Kathryn Pitkin Derose, Lindsay Rosenfeld, Stephen L Buka and Rima Rudd. 2011. "Patient activation and advocacy: which literacy skills matter most?" *Journal of health communication* 16(sup3):177–190.

Schuman, Howard and Stanley Presser. 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Spooren, Wilbert and Liesbeth Degand. 2010. "Coding coherence relations: Reliability and validity.".