Coding text answers to open-ended questions: human coders and statistical learning algorithms make similar mistakes

Zhoushanyue (Sophie) He, Ph.D., University of Waterloo, Canada *

Matthias Schonlau, Ph.D., University of Waterloo, Canada

* now: Roche Pharmaceutical Company

Introduction

- Open ended questions yield text answers
- They are often hand-coded into one of several classes/codes
- Recently, a trend to substitute human coders with automated coders
  - Automated coder = Coding by a statistical learning model
- Do automated coders and human coders find the same type of observations difficult to code?
  - i.e., do they make similar coding mistakes?
  - Or might one be able to compensate for the other's weaknesses?

Double coding

- In double coding two coders independently assign a code to a text answer
- When they disagree, the disagreement is resolved
  - Discuss until you agree
  - A third coder casts the deciding vote
  - Expert resolves
- We call the resolved code the gold-standard
- When two coders disagree, one of them made a coding error
  - (In the sense that code does not match the gold standard)

Data sets

Patient Joe

“Joe's doctor told him that he would need to return in two weeks to find out whether 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?”

4 classes: proactive, semi-proactive, passive, counterproductive

(Dutch)

Happiness

“What aspects of your life have you considered when assessing your happiness?”

10 classes such as social network & surrounding, health and job
(German)

Democracy

“What aspects did you think of when answering the question how satisfied you were with the way democracy works in Germany?”

7 classes such as “actors & groups”, “public policy areas” and “evaluation of behavior of politicians & parties”
(German)

Data sets vary in % disagreement / kappa

3 double-coded data sets

	Size of the (whole) dataset	Percentage of disagreement	Kappa
Patient Joe	1756	23.18%	0.61
Happiness	1438	5.77%	0.93
Democracy	1096	14.42%	0.82

Turn text into n-gram variables

N-gram variables are used as x-variables

	text	_challeng	_corona	_vaccin	n_token
1.	I say Corona, you say Covid	0	1	0	6
2.	Find a vaccine, please!	0	0	1	4
3.	No vaccine. All is challenging. CHALLENGE!	2	0	1	6
4.	Will Corona beer change its name?	0	1	0	6
5.	Home schooling is a challenge.	1	0	0	5

- stemming
- removing stop-words
- word needs to occur in at least 2 texts to be listed (threshold=2)
- Number of words (n_token)

Modeling and Prediction

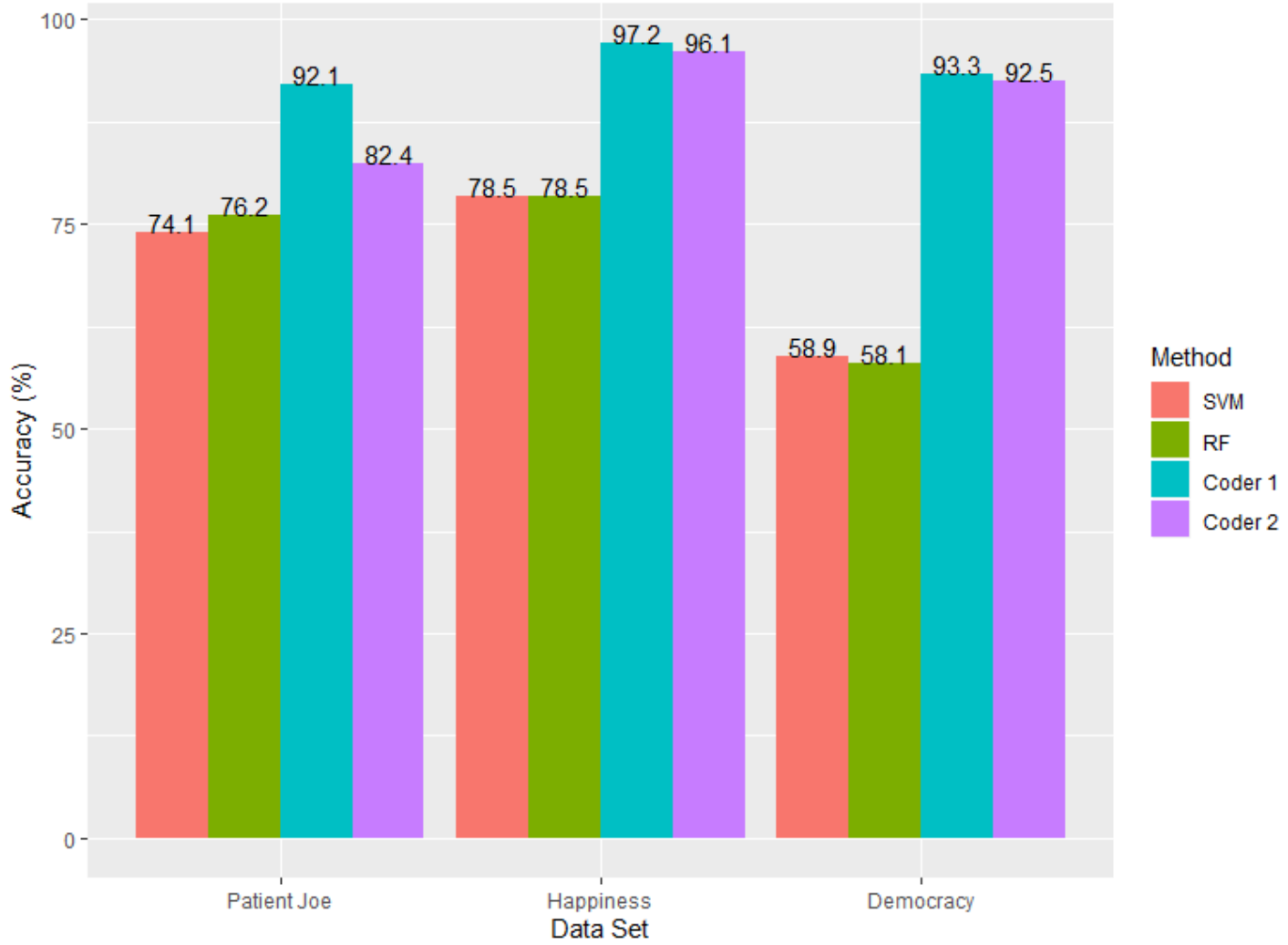
- Randomly split each of the datasets into a training/test datasets
- 2 statistical learning models: SVM and random forests
- Models are trained on the training data based on the gold-standard code
- Use the trained models to predict the codes of the test data

	Size of the (whole) dataset	Size of training dataset	Size of test dataset
Patient Joe	1756	1000	756
Happiness	1438	800	638
Democracy	1096	600	496

Note: training data includes validation data for tuning

Human accuracy is greater than that of statistical learning on ngram variables

- Accuracy is relative to the gold-standard code



Similar mistakes = correlated coding error

- For each coder, we compute the probability of a coding error (for a given answer text)
- If increased probability for human coding errors is associated with increased probability for automated coding error, we infer that they make similar mistakes

Estimated error probability from models

- For each text answer the statistical model predicts classification probabilities:
- E.g. for one text *i* and four classes:
  - p_{i1}=0.6 “proactive”
  - p_{i2}=0.2 “somewhat proactive”
  - p_{i3}=0.1 “passive”
  - p_{i4}=0.1 “counterproductive”
- We predict “proactive” for text *i*, because it has the largest probability
- The estimated probability of error for text *i* is 1-0.6=0.4.

Estimated error probability from humans

- Humans coders classify observations without giving probabilities
  - e.g. “proactive”
- We aggregate data into subsets
- The estimated error probability for a subset is the fraction of incorrect classifications in that subset

Estimated error probability from humans

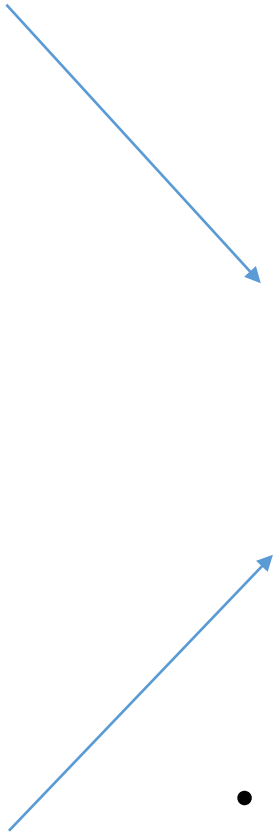
- Example with 3 subsets

How are observations sorted into subsets?

- sorted by average error probabilities of automated coders. However, it turns out, random ordering gives the same substantive results

Observation Index	Coder 1	Coder 2	SVM error prob.	RF error prob.	Average error probability of automated coders	Subset
7	incorrect	incorrect	0.6	0.4	0.5	A
10	incorrect	incorrect	0.5	0.4	0.45	A
15	incorrect	correct	0.4	0.5	0.45	A
4	correct	incorrect	0.5	0.3	0.4	A
5	incorrect	incorrect	0.2	0.4	0.3	A
8	correct	incorrect	0.2	0.4	0.3	B
9	correct	correct	0.3	0.3	0.3	B
3	incorrect	correct	0.3	0.2	0.25	B
13	correct	correct	0.3	0.2	0.25	B
14	correct	correct	0.2	0.3	0.25	B
12	incorrect	correct	0.2	0.2	0.2	C
1	correct	correct	0.1	0.2	0.15	C
11	correct	correct	0.2	0.1	0.15	C
2	correct	correct	0.1	0.1	0.1	C
6	correct	correct	0.1	0.0	0.05	C

Observation Index	Coder 1	Coder 2	SVM error prob.	RF error prob.	Average error probability of automated coders	Subset
7	incorrect	incorrect	0.6	0.4	0.5	A
10	incorrect	incorrect	0.5	0.4	0.45	A
15	incorrect	correct	0.4	0.5	0.45	A
4	correct	incorrect	0.5	0.3	0.4	A
5	incorrect	incorrect	0.2	0.4	0.3	A
8	correct	incorrect	0.2	0.4	0.3	B
9	correct	correct	0.3	0.3	0.3	B
3	incorrect	correct	0.3	0.2	0.25	B
13	correct	correct	0.3	0.2	0.25	B
14	correct	correct	0.2	0.3	0.25	B
12	incorrect	correct	0.2	0.2	0.2	C
1	correct	correct	0.1	0.2	0.15	C
11	correct	correct	0.2	0.1	0.15	C
2	correct	correct	0.1	0.1	0.1	C
6	correct	correct	0.1	0.0	0.05	C



Subset	Error probability of Coder 1	Error probability of Coder 2	Av. error probability of SVM	Average error probability of RF
A	0.8	0.8	0.44	0.4
B	0.2	0.2	0.26	0.28
C	0.2	0	0.14	0.12

- All subsequent analyses are on the aggregate level
- Each data set had ~30 subsets
- For the automated coders, we also compute average errors at the aggregate level

Correlations

- We now look at pairwise correlations between the 4 coders
  - (2 automated, 2 humans)

Correlation matrix of estimated error probabilities for each dataset

- All correlations are positive
- Very high correlations between automated coders
- Correlations (automated coder, human) similar to correlations (human,human)

Patient Joe				
	SVM	RF	Coder 1	Coder 2
SVM	1.00	0.95	0.44	0.88
RF		1.00	0.44	0.89
Coder 1			1.00	0.29
Coder 2				1.00
Happiness				
	SVM	RF	Coder 1	Coder 2
SVM	1.00	1.00	0.70	0.69
RF		1.00	0.71	0.69
Coder 1			1.00	0.65
Coder 2				1.00
Democracy				
	SVM	RF	Coder 1	Coder 2
SVM	1.00	1.00	0.53	0.31
RF		1.00	0.51	0.31
Coder 1			1.00	0.40
Coder 2				1.00

Beyond pairwise similarities

- Correlations reveal pairwise similarities
- We now look at all 4 coders (2 automated, 2 humans) simultaneously using principal component analysis

Principal components analysis

- Coders have far more in common (first component) than there are differences (other components)
- First component represents a weighted average of the 4 coders and explains most variation (65%-80%)
- Second component represents a contrast between automated and human coders and explains less variation (10%-20%)

Patient Joe

	Dim.1	Dim.2	Dim.3	Dim.4
SVM	0.97	0.10	0.18	0.15
RF	0.97	0.11	0.11	-0.17
Coder 1	0.55	-0.83	-0.05	0.00
Coder 2	0.92	0.28	-0.27	0.03
Variation explained	76.0%	19.7%	2.9%	1.3%

Happiness

	Dim.1	Dim.2	Dim.3	Dim.4
SVM	0.95	0.30	0.05	0.04
RF	0.95	0.29	0.04	-0.04
Coder 1	0.85	-0.27	-0.46	0.00
Coder 2	0.84	-0.41	0.37	-0.00
Variation explained	80.7%	10.4%	8.8%	0.1%

Democracy

	Dim.1	Dim.2	Dim.3	Dim.4
SVM	0.94	0.32	0.14	0.03
RF	0.93	0.33	0.16	-0.03
Coder 1	0.75	-0.25	-0.62	-0.00
Coder 2	0.55	-0.77	0.33	0.00
Variation explained	65.0%	21.7%	13.3%	0.1%

Correlations between principal components and the original estimated error probabilities

Conclusion

- Statistical models using n-gram variables have higher error rates than human coders
- Automated coders (models) and human coders tend to make similar coding mistakes
- Two very different statistical models give highly correlated estimated coding errors. The choice of statistical model does not appear to matter.

Contact: schonlau@uwaterloo.ca
www.schonlau.net

References

This talk:

- He Z, Schonlau M. Coding text answers to open-ended questions: human coders and statistical learning algorithms make similar mistakes. *Methods*, *Data*, *Analyses*. 15(1), 2021, pp. 103-120. https://mda.gesis.org/index.php/mda/article/view/2020.10
- He, Zhoushanyue (2021). On the Automatic Coding of Text Answers to Open-ended Questions in Surveys, Ph.D. Thesis, University of Waterloo, 2021-01-13. https://uwspace.uwaterloo.ca/handle/10012/16643

Other Open-ended questions:

- Schonlau, M., Couper M. Semi-automated categorization of open-ended questions. Survey Research Methods. Aug 2016, 10(2), 143-152. https://doi.org/10.18148/srm/2016.v10i2.6213

Stata Software:

- Schonlau, M., Guenther, N. Sucholutsky, I. Text mining using ngram variables. The Stata Journal. Dec 2017, 17(4), 866-881.