Summer School JGU Mainz—Advanced Methods in Behavioral Economics

Dr. Carina I. Hausladen

ETH Zurich

September 30, 2021

Human: What do we want?!
Computer: Natural Language Processing!
Human: When do we want it!?
Computer: When do we want what?

Posted by sachintripathi007 on Reddit

3

Predicting (Dis-)Honesty: Leveraging Text Classification for Behavioral Experimental Research

Carina I. Hausladen, Martin Fochmann, Peter Mohr

Literature

Explanations for the *dishonesty shift* in groups (Baeker et al. 2015; Chytilova et al. 2014; Conrads et al. 2013; Fochmann et al. 2018; Kocher et al. 2018; Weisel et al. 2015, e.g.):

- Groups act more strategically.
- The reduced observability of individual actions lead to less accountability (Conrads et al. 2013; Mazar and Aggarwal 2011).
- Communication leads to a change in norm perception (Chytilova et al. 2014; Gino et al. 2009; Kocher et al. 2018).
- Other people can benefit from dishonest behavior (Gino et al. 2013; Schweitzer et al. 2002; Weisel et al. 2015; Wiltermuth 2011).

Literature

- *Ethics* can be made salient by
  - signing at the beginning rather than at the end of a self-report. (Shu et al. 2012)
  - religious reminders (Mazar, Amir, et al. 2008).
  - the watching eyes effect (Bateson et al. 2013).
- Treatments can, however, lead to a *crowding-out* effect of intrinsic motivation (e.g. Gneezy et al. 2000).

Research Question

How can we assign interventions without risking a crowding-out effect?

Typical Behavioral Experimental Research

(a) Group Interaction (b) Decision Phase

Data can be interpreted as gold-standard labeled data.

Text Data as Process Data

Behavioral research uses text as process data:

- Capra (2019), Elten et al. (2020), Fochmann et al. (2018), and Kocher et al. (2018) manually assign labels.
- Andres et al. (2019) and Capra (2019) construct word clouds.
- Arad et al. (2018), Burchardi et al. (2014), Georgalos et al. (2019), and Penczynski (2019) assign labels to text in a (semi-)supervised way.

Research Goals

This research combines decision with process data in a novel way:

- We use supervised learning to predict honesty based on group chats.
- We test the classifier‘s generalizability.

Data, provided by Fochmann et al. (2018)

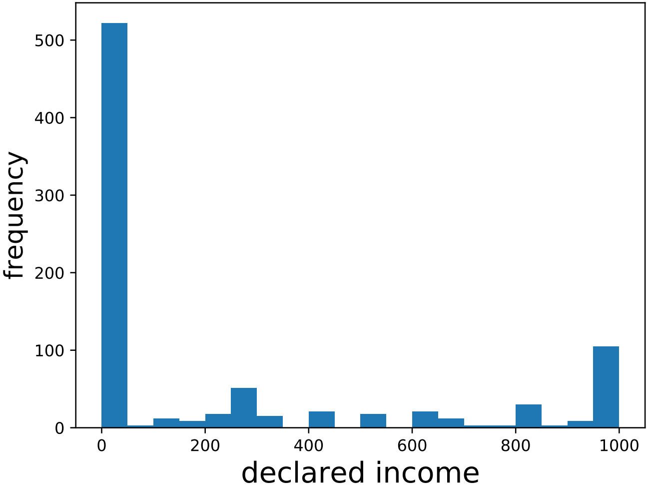


Figure: Distribution of the dependent variable

Threshold

	F1	prec	rec	AUC	acc
$<$mean	0.362	0.227	0.963	0.558	0.286
$<$500	0.295	0.185	0.848	0.529	0.427
$<$1000	0.209	0.126	0.857	0.544	0.449

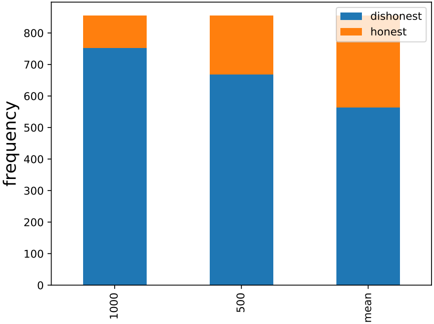


Figure: Distribution of the *binarized* dependent variable

Comparison of Feature Representations' Performance

	F1	pr	re	AUC	acc
bag of words	**0.490**	0.335	0.953	0.563	0.373
Word2Vec (pre, tf-idf)	0.489	0.336	0.982	0.509	0.372
Word2Vec (pre, smpl)	0.486	0.329	0.971	0.482	0.353
bag of words (tf-idf)	0.482	0.332	0.922	0.565	0.376
Word2Vec (tf-idf)	0.481	0.321	1.000	0.554	0.320
Doc2Vec	0.361	0.232	0.909	0.500	0.317
GloVe (pre, tf-idf)	0.356	0.223	0.959	0.494	0.272
GloVe (pre)	0.355	0.222	0.967	0.469	0.262
fastText	0.354	0.239	0.740	0.569	0.436
Word2Vec	0.350	0.275	0.659	0.547	0.489

Note: The classifier used was a logistic regression. F1-score, precision, and recall are reported for the minority label (= 1) "honest".

Comparison of the Classifiers' Performance

	F1	prec	rec	AUC	acc
Stacking	0.411	0.292	0.741	**0.597**	0.556
KNN	0.390	0.268	0.778	0.581	0.492
SVM	0.364	0.227	1.000	0.526	0.266
RF	0.364	0.236	0.894	0.550	0.343
NN	0.356	0.225	0.926	0.559	0.298
Bagging	0.356	0.220	1.000	0.543	0.238
LLR	0.355	0.229	0.893	0.522	0.319
XGBoost	0.349	0.215	1.000	0.474	0.214

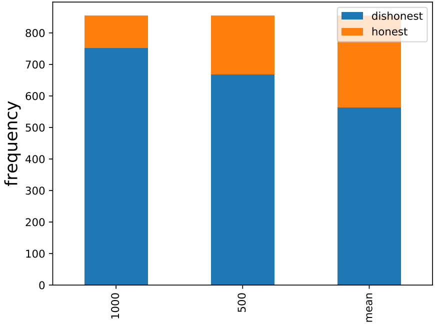
(a) Performance Metrics

	Model Weights
KNN	0.931
NN	0.805
LLR	0.618
RF	0.060
SVM	0.000
XGB	0.000

(b) Model Weights for Stacking

Note: Pretrained, tf-idf weighted Word2Vec embeddings were used. F1 score, precision, and recall are reported for the minority label (= 1) "honest".

Possible Explanations

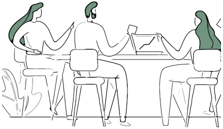


- train: 603 decisions (237 honest)
- test: 252 decisions (51 honest)

Testing Generalizability

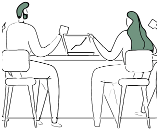
Experiment I

- tax evasion game
- 3 group members
- lying *upwards* earns more money

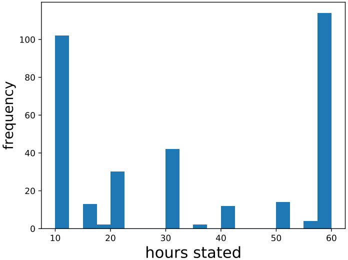


Experiment II

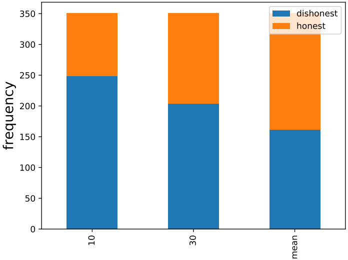
- report surplus hours
- 2 group members
- lying *downwards* earns more money



Data, Experiment II



(a) Distribution of Surplus Hours Stated



(b) Categories Based on Specific Thresholds

Figure: The Dependent Variable

Comparing Models *Across* Datasets

- As the F1 score is solely interested in the performance of the positive class, it is *sensitive to different class distributions*.
- The AUC is prevalence independent because it is built from a separate evaluation of the two classes (for an excellent overview see Straube et al. 2014).

Out-of-Context Performance of the Pretrained Classifier

y	F1	prec	rec	AUC	acc
$>$ mean	0.704	0.548	0.995	**0.529**	0.553
$>$ 30	0.591	0.425	0.986	0.510	0.433
$>$ 10	0.455	0.305	0.922	0.506	0.365

The classification was based on pre-trained Word2Vec embeddings, averaged over texts by tf-idf weighting and a stacking classifier. F1 score, precision, and recall are reported for the minority label ($= 1$) "honest".

Behavioral Analysis

Is text an independent predictor for decisions?

Yes, only 4% of participants (overall 15) stated to have changed their behavior:

- 5 participants thought about the chat being read but did not change their behavior.
- 4 participants put more effort into proper grammar and spelling.
- 4 participants wrote less text.
- 2 participants reported more honestly.

Behavioral Analysis

Lying (proxied by hours stated) and correlated concepts.

Table: Pearson‘s Correlation Coefficient

	estimates	p-value
Belief	0.610	0.000
Joy	0.276	0.000
Risk Attitude	0.266	0.000
Lying Attitude	0.191	0.001
Number of Words	0.232	0.000

Note: hours stated $\in$ [10..60], 10$\equiv$full compliance; joy experienced $\in$ [1..10], 1$\equiv$experienced no joy; beliefs $\in$ [0..100]: 0$\equiv$0 people state more than the true amount; risk attitude $\in$ [1..11], 1$\equiv$not risk-prone; lying attitude $\in$ [1..10], 1$\equiv$one should never lie.

Proxy for Joy

Rauh (2018)'s German Sentiment Dictionary includes $17,330$ terms indicating positive sentiment.

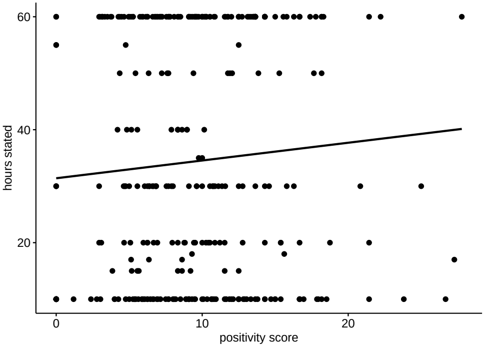


Figure: Relation of the Positivity Score ($\in [0, 100]$) and Surplus Hours Stated; Pearson's Correlation Coefficient: .104 (p-value$=$ .061)

Conclusion

- Text can be interpreted as an independent predictor of the decision.
- The predictive performance is better than a random guess (AUC$=.597$), despite a tiny and heavily skewed dataset.
- The classifier does not generalize to another context (AUC$=.523$).
- Risk attitudes, joy experienced and the number of words can approximate (dis-)honesty.

Let‘s Discuss!

- Are you aware of *experimental data*, combining text and decision data?
- Are you aware of *field data*, combining text and decision data?
- How would you improve predictive performance on a tiny and heavily skewed dataset?
- How would you proxy risk and lying attitudes?

References I

Andres, Maximilian et al. (2019). "The Effect of Leniency Rule on Cartel Formation and Stability: Experiments with Open Communication".

Arad, Ayala et al. (2018). "Multi-Dimensional Reasoning in Competitive Resource Allocation Games : Evidence from Intra-Team Communication".

Baeker, Agnes et al. (2015). *Peer Settings Induce Cheating on Task Performance*. Tech. rep. IAAEU Discussion Paper Series in Economics.

Bateson, Melissa et al. (2013). "Do Images of 'Watching Eyes' Induce Behaviour that is More Pro-Social or More Normative? A Field Experiment on Littering". In: *PloS one* 8.12, e82055.

Burchardi, Konrad B. et al. (Mar. 2014). "Out of Your Mind: Eliciting Individual Reasoning in One Shot Games". In: *Games and Economic Behavior* 84, pp. 39–57.

Capra, C. Mónica (2019). "Understanding Decision Processes in Guessing Games: A Protocol Analysis Approach". In: *Journal of the Economic Science Association* 5.1, pp. 123–135.

Chytilova, Julie et al. (2014). *Individual and Group Cheating Behavior: A Field Experiment with Adolescents*. Tech. rep. IES Working Paper.

Conrads, Julian et al. (2013). "Lying and Team Incentives". In: *Journal of Economic Psychology* 34, pp. 1–7.

Elten, Jonas van et al. (Jan. 2020). "Coordination Games with Asymmetric Payoffs: An Experimental Study with Intra-Group Communication". In: *Journal of Economic Behavior and Organization* 169, pp. 158–188.

Fochmann, Martin et al. (2018). "Dishonesty and Risk-Taking : Compliance Decisions of Individuals and Groups".

References II

Georgalos, Konstantinos et al. (2019). “Testing for the Emergence of Spontaneous Order”. In: *Experimental Economics*.

Gino, Francesca et al. (2009). “Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel”. In: *Psychological Science* 20.3, pp. 393–398.

— (2013). “Self-Serving Altruism? The Lure of Unethical Actions that Benefit Others”. In: *Journal of Economic Behavior & organization* 93, pp. 285–292.

∗ Gneezy, Uri et al. (Aug. 2000). “Pay Enough or Don’t Pay at All”. In: *Quarterly Journal of Economics* 115.3, pp. 791–810.

Kocher, Martin G. et al. (2018). “I lie? We lie! Why? Experimental Evidence on a Dishonesty Shift in Groups”. In: *Management Science* 64.9, pp. 3995–4008.

Mazar, Nina and Pankaj Aggarwal (2011). “Greasing the Palm: Can Collectivism Promote Bribery?” In: *Psychological Science* 22.7, pp. 843–848.

Mazar, Nina, On Amir, et al. (2008). “The Dishonesty of Honest People: A Theory of Self-Concept Maintenance”. In: *Journal of Marketing Research* 45.6, pp. 633–644.

Penczynski, Stefan P. (Feb. 2019). “Using Machine Learning for Communication Classification”. In: *Experimental Economics*, pp. 1–28.

Rauh, Christian (2018). “Validating a Sentiment Dictionary for German Political Language – A Workbench Note”. In: *Journal of Information Technology & Politics* 15.4, pp. 319–343.

Schweitzer, Maurice E. et al. (2002). “Stretching the Truth: Elastic Justification and Motivated Communication of Uncertain Information”. In: *Journal of Risk and Uncertainty* 25.2, pp. 185–201.

References III

* Shu, Lisa L. et al. (2012). "Signing at the Beginning Makes Ethics Salient and Decreases Dishonest Self-Reports in Comparison to Signing at the End". In: 108.38, pp. 15197–15200.

Straube, Sirko et al. (2014). "How to Evaluate an Agent's Behavior to Infrequent Events? – Reliable Performance Estimation Insensitive to Class Distribution". In: *Frontiers in Computational Neuroscience* 8, p. 43.

Weisel, Ori et al. (2015). "The Collaborative Roots of Corruption". In: *Proceedings of the National Academy of Sciences* 112.34, pp. 10651–10656.

Wiltermuth, Scott S. (2011). "Cheating More when the Spoils are Split". In: *Organizational Behavior and Human Decision Processes* 115.2, pp. 157–168.