PREDICTING HIGHER ORDER VALUES OF YOUTH FROM DIGITAL TRACES ON SOCIAL MEDIA

Mikhail Bogdanov
PhD student in Sociology
Junior Research Fellow at Institute of Education
National Research University «Higher School of Economics»
Moscow, Russia

mbogdanov@hse.ru
www.hse.ru/en/staff/bogdanov

MOTIVATION

As meta-analysis has demonstrated - human traits and attributes could be predicted from digital traces on social media (Settanni et al., 2018)

Probably the most studied phenomenon is personality traits (see Azucar et al., 2018 for meta-analysis and references)

However, there are only a few studies that attempt to predict human values from the digital traces (Boyd et al., 2015; Chen et al., 2014; Kalimeri et al., 2019; Mukta et al., 2019, 2016)

All these studies examined the general population and collected survey data through convenience sampling and digital traces mostly from Facebook (except for Chen et al., 2014; Kalimeri et al., 2019)

RESEARCH QUESTION

To what extent can we predict Human Values solely from digital traces of age homogeneous cohort which is broadly present on social media?

DATA

Trajectory in Education and Careers (TrEC)* – longitudinal cohort study based on participants of Trends in Mathematics and Science Study (TIMSS-2011)

4893

TIMSS-2011

Base wave
Cohort of Russian eighth graders in 2010-2011 academic year who participated in TIMSS-2011
Nationally representative sample
210 schools
42/83 regions

3732

8 wave - 2019

Included 21-item Portrait Values Questionnaire (PVQ) that measures Schwartz' Human Values
Modal age of cohort = 23 years

2083

Analyzed subsample

Participated in 8th wave &
Have reliable answers on PVQ (didn't have the same answer > 16 items) &
Gave informed consent to collect their data from VK profiles
N of subscriptions > 5 & N of friends > 0

*More info on https://trec.hse.ru/en/

DIGITAL TRACES

VK is the largest social network site in Russia and post-soviet countries with over 97 million monthly active users in 2018-2019*
Over 90% of youth in age 18-24 regularly used VK in 2016**

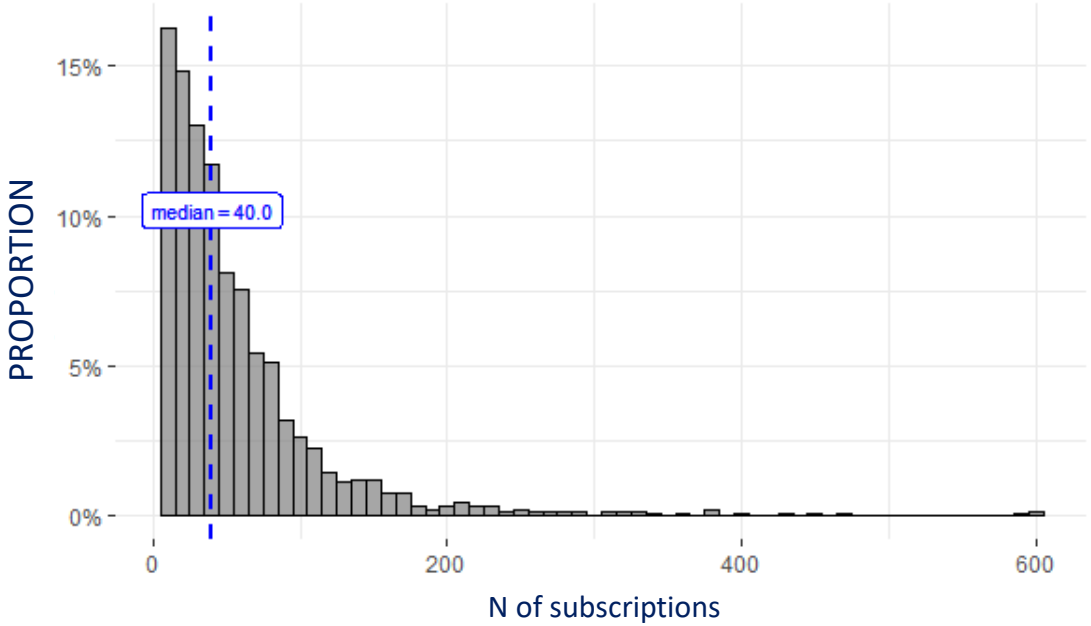
Collected features

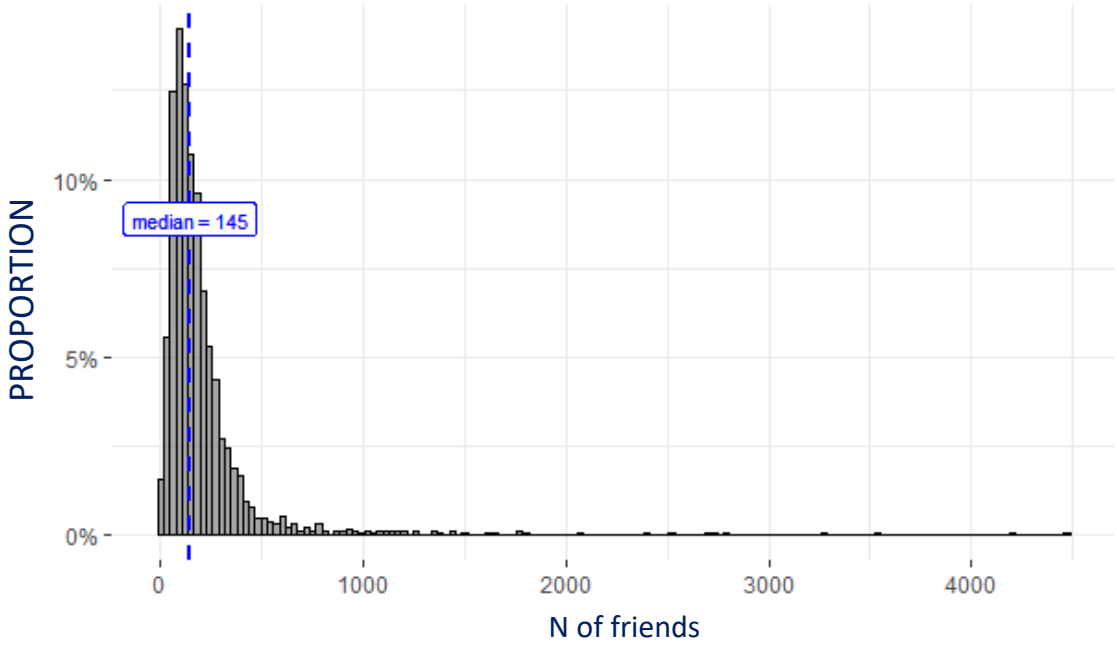
- Subscriptions on public pages/groups/communities
- Number of friends on social network

* https://vk.com/about
** FOM. Online practices of Russians: social networks. URL: https://fom.ru/SMI-i-internet/12495

PREPROCESSING DIGITAL TRACES

We removed respondents with less than 5 subscriptions and zero friends
Public pages/groups/communities with less than 10 subscribers among respondents were also removed
Similar threshold approach was used in Kosinski, 2016; Smirnov, 2018
After filtering there were 5122 public pages/groups/communities





DIFFERENCE BETWEEN TIMSS-2011 & ANALYZED SAMPLE

There are no substantial differences between base wave (TIMSS-2011) which is a representative sample of the cohort and subsample which were used for prediction

	TIMSS-2011	Not used in prediction	Used in prediction	p	SMD
N	4893	2810	2083		
Sex = boy (%)	50.7	52.6	48.0	0.002	0.093
Parents education (%)				0.014	0.110
university or higher	47.6	46.4	49.3		
post-secondary but not university	25.9	25.1	26.9		
upper secondary	10.1	11.1	8.8		
lower secondary	5.6	5.9	5.1		
some primary, lower secondary or no school	0.2	0.2	0.1		
omitted or invalid	10.6	11.3	9.8		
School area (%)				<0.001	0.140
urban	21.6	22.7	20.1		
suburban	11.5	11.9	11.0		
medium size city	45.7	42.9	49.4		
small town	6.7	6.7	6.7		
remote rural	14.6	15.9	12.8		

DIFFERENCE BETWEEN TIMSS-2011 & ANALYZED SAMPLE

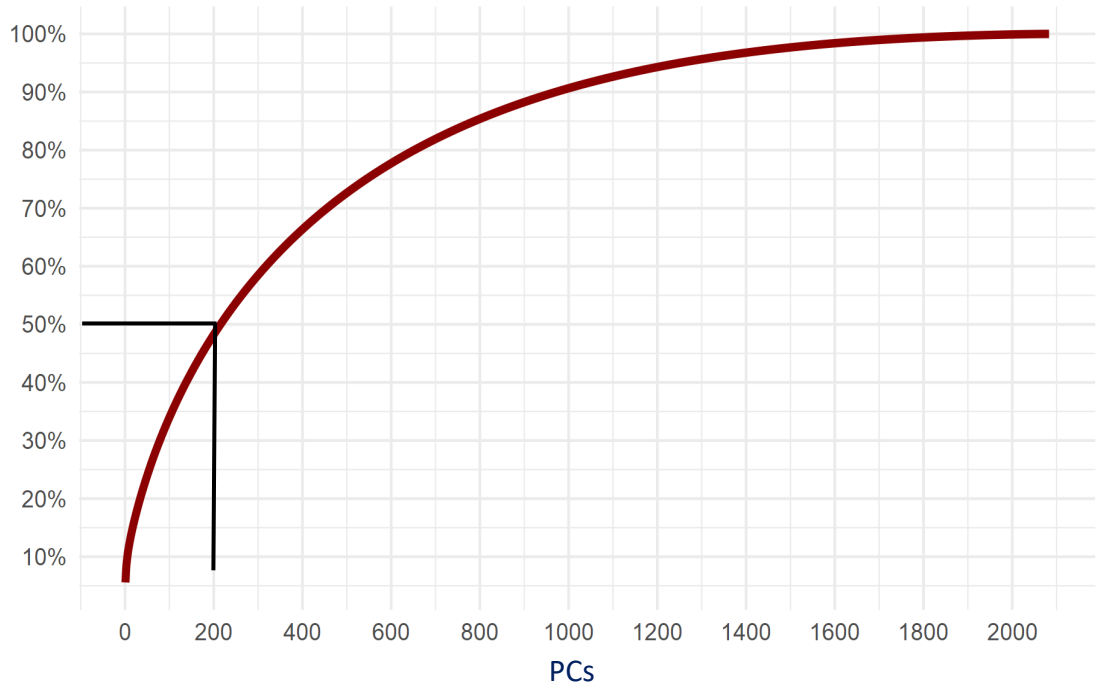
However, there are some difference in educational attainment

	TIMSS-2011	Not used in prediction	Used in prediction	p	SMD
Computer at home (%)	91.1	89.2	93.6	<0.001	0.157
Internet at home (%)	86.0	85.2	87.1	0.085	0.065
Computer usage home (%)				<0.001	0.140
every day or almost every day	79.7	77.5	82.6		
once or twice a week	10.8	11.6	9.8		
once or twice a month	1.5	1.5	1.3		
never or almost never	6.1	6.9	5.1		
TIMSS math (mean (SD))	542.63 (77.84)	535.42 (78.46)	552.36 (75.95)	<0.001	0.219
TIMSS science (mean (SD))	544.99 (72.26)	538.89 (73.66)	553.22 (69.50)	<0.001	0.200

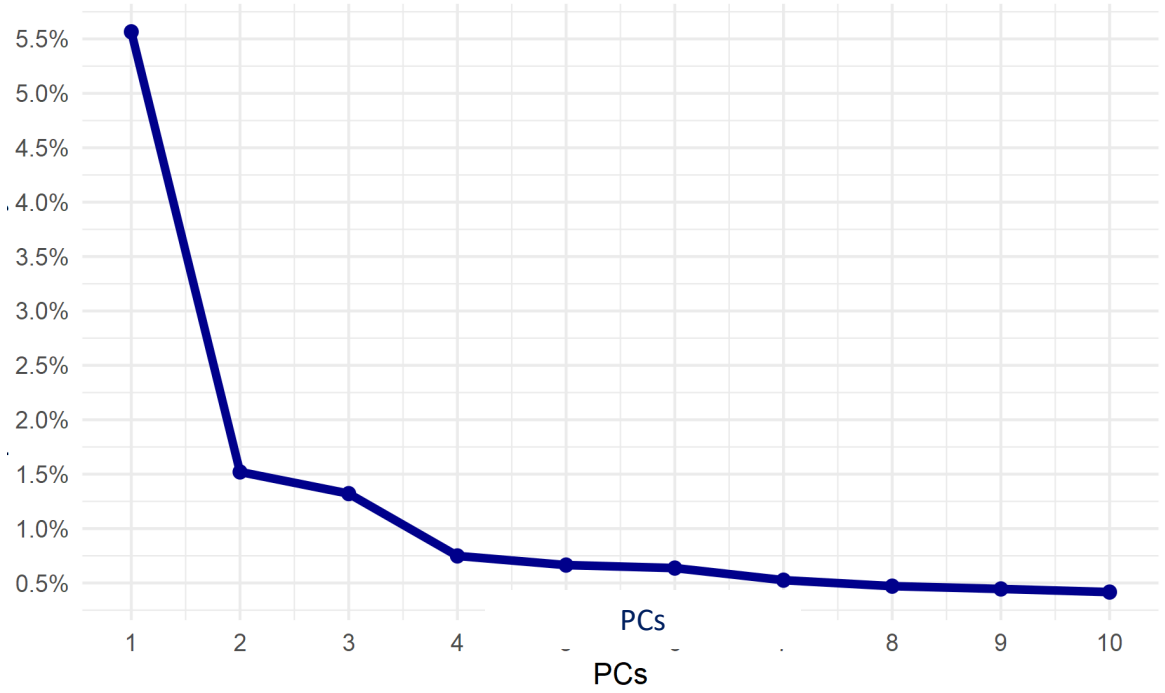
FEATURE ENGINEERING

Log(N of subscriptions)
Log(N of friends)
PCA on subscriptions: varying 30, 50, 75, 100 PCs for modelling

Cumulative explained variance by PCs



Explained variance by PCs



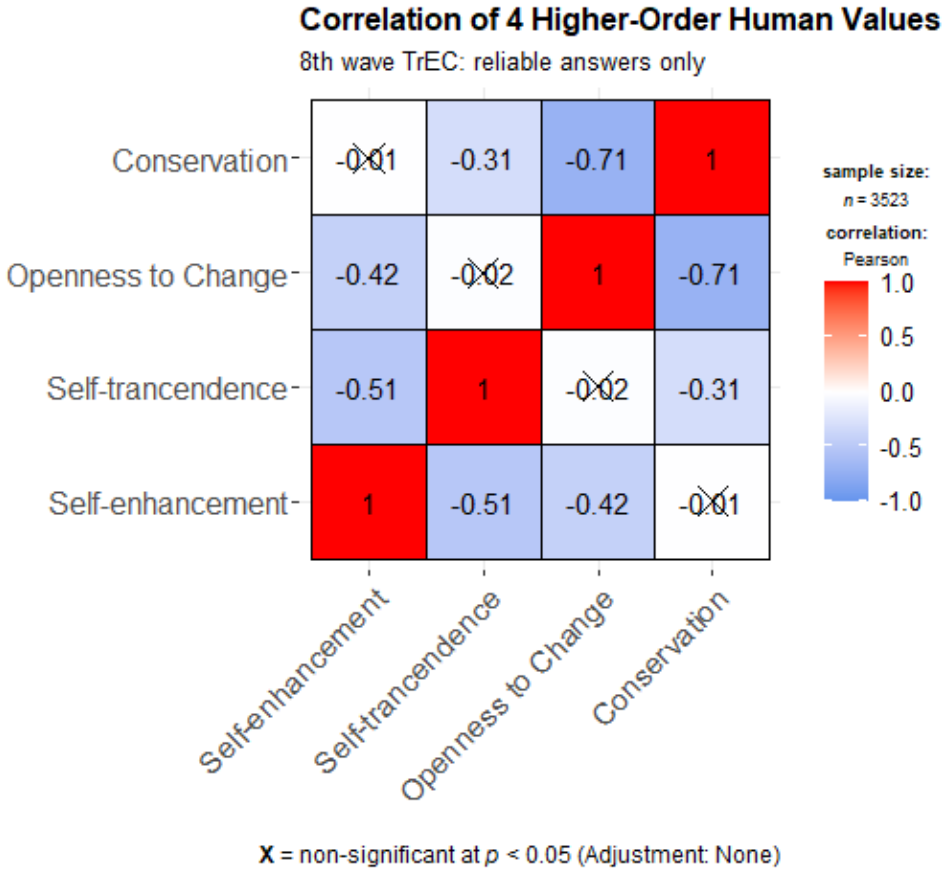
OUTCOME

All Higher Order Values were adjusted for response tendencies - the mean answer for all 21-items was subtracted from values' index

Correlation structure is in line with the observed relationship of HOV across the world and different surveys (see Rudnev, Magun, Schwartz 2018)

Hedonism (one of the 10 Basic Human Values) was treated as part of the Openness to Change (Schwartz insists that it is closer to Openness to Change rather than Self-enhancement in ~75% studies - see Schwartz 2003)

Higher Order Values	Cronbach's Alpha
Self-enhancement	0,72
Self-transcendence	0,633
Conservation	0,607
Openness to Change	0,694



MODELLING STRATEGY

TIDYMODELS PACKAGE FRAMEWORK IN R
5-FOLD-CROSS-VALIDATION

Outcome preprocessing for regression task: adjustment for response tendency by subtracting mean answer for 21 PVQ items of the respondent
Outcome preprocessing for classification task: dichotomization by median high and low (for comparability with previous studies)

Metrics for regression task: correlation of predicted and true values, R-squared
Metrics for classification task: ROC-AUC

	Package in R	Hyperparameters tuning
Linear/Logistic Regression	lm/glm	None
Elastic Net	glmnet	penalty, mixture
Random Forest	randomForest	N of trees, min N in node, N of randomly selected predictors
SVM RBF	kernlab	RBF sigma, margin

RESULTS

Subscriptions and N of friends plays best at the prediction of Self-Enhancement and Conservation

Correlation	Self-enhancement	Self-transcendence	Conservation	Openness to Change
Linear Regression	0,20	0,11	0,20	0,10
Elastic Net	0,20	0,11	0,20	0,11
Random Forest	0,18	0,07	0,21	0,12
SVM RBF	0,21	0,10	0,23	0,13
R-squared				
Linear Regression	4,0%	1,3%	4,0%	1,1%
Elastic Net	4,0%	1,3%	4,0%	1,1%
Random Forest	3,3%	0,5%	4,6%	1,5%
SVM RBF	4,2%	1,0%	5,4%	1,6%
ROC-AUC				
Logistic Regression	0,62	0,57	0,60	0,55
Elastic Net	0,62	0,57	0,60	0,55
Random Forest	0,60	0,53	0,61	0,55
SVM RBF	0,62	0,56	0,60	0,55

Tuning machine learning algorithms have not led to a substantial increase in the performance

Previous studies' R-squared/ROC-AUC					Metric	Type of digital footprints	N	Validation
Chen, 2014*	13,8%	17%	15,4%	18,1%	R^2	Posts on Reddit	799	10-fold-CV
	0,56	0,60	0,59	0,61	ROC-AUC			
Mukta, 2019*	10,2%	19,1%	21,3%	27,6%	R^2	Statuses on FB	726	10-fold-CV
Mukta, 2016*	15,4%	17,9%	20,9%	16,1%	R^2	Page-likes on FB	567	10-fold-CV
	0,62	0,71	0,74	0,78	ROC-AUC	Statuses, page-likes, share-links on FB		
Youyou, 2015	10 basic values was predicted, correlation ranges from 0,03 to 0,23				r	Page-likes on FB	70,520	10-fold-CV

*Hedonism was treated as separate higher order value, which don't belong neither to Openness to Change or Self-enhancement

REFERENCES

1. Schwartz, Shalom. (2003). A proposal for measuring value orientations across nations. Questionnaire Package of ESS. 259-290.
2. Smirnov, I. (2018). Predicting PISA Scores from Students' Digital Traces. Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018), 3, 360–365.
3. Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. Psychological Methods, 21(4), 493–506. https://doi.org/10.1037/met0000105
4. Rudnev, M., Magun, V., & Schwartz, S. (2018). Relations Among Higher Order Values Around the World: Journal of Cross-Cultural Psychology. https://doi.org/10.1177/0022022118782644
5. Mukta, M. S. H., Ali, M. E., & Mahmud, J. (2019). Temporal modeling of basic human values from social network usage. Journal of the Association for Information Science and Technology, 70(2), 151–163. https://doi.org/10.1002/asi.24099
6. Mukta, Md. S. H., Ali, M. E., & Mahmud, J. (2016). User Generated vs. Supported Contents: Which One Can Better Predict Basic Human Values? In E. Spiro & Y.-Y. Ahn (Eds.), Social Informatics (pp. 454–470). Springer International Publishing. https://doi.org/10.1007/978-3-319-47874-6_31
7. Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159. https://doi.org/10.1016/j.paid.2017.12.018
8. Kalimeri, K., Beiró, M. G., Delfino, M., Raleigh, R., & Cattuto, C. (2019). Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior*, *92*, 428–445. https://doi.org/10.1016/j.chb.2018.11.024
9. Boyd, R. L., Wilson, S. R., Pennebaker, J., Kosinski, M., Stillwell, D., & Mihalcea, R. (2015). Values in Words: Using Language to Evaluate and Understand Personal Values. *ICWSM*.

Mikhail Bogdanov

mbogdanov@hse.ru

www.hse.ru/en/staff/bogdanov