Items from Psychometric Tests as Training Data for Personality Profiling Models of Twitter Users

Anne Kreuter[1]**, Kai Sassenberg**[2,3]**,** and **Roman Klinger**[1]

[1]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
[2]Leibniz-Institut für Wissensmedien, Tübingen, Germany
[3]University of Tübingen, Germany
`{anne.kreuter,roman.klinger}@ims.uni-stuttgart.de`
`k.sassenberg@iwm-tuebingen.de`

Abstract

Machine-learned models for author profiling in social media often rely on data acquired via self-reporting-based psychometric tests (questionnaires) filled out by social media users. This is an expensive but accurate data collection strategy. Another, less costly alternative, which leads to potentially more noisy and biased data, is to rely on labels inferred from publicly available information in the profiles of the users, for instance self-reported diagnoses or test results. In this paper, we explore a third strategy, namely to directly use a corpus of items from validated psychometric tests as training data. Items from psychometric tests often consist of sentences from an I-perspective (*e.g.*, "I make friends easily."). Such corpora of test items constitute 'small data', but their availability for many concepts is a rich resource. We investigate this approach for personality profiling, and evaluate BERT classifiers fine-tuned on such psychometric test items for the big five personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism) and analyze various augmentation strategies regarding their potential to address the challenges coming with such a small corpus. Our evaluation on a publicly available Twitter corpus shows a comparable performance to in-domain training for 4/5 personality traits with T5-based data augmentation.

1 Introduction

The field of author profiling originally emerged from the study of stylometry (Lutoslawski, 1898) and, with the rise of social media (Bilan and Zhekova, 2016), now considers a variety of attributes, including demographic data such as age, sex, gender, nationality (Schwartz et al., 2013), personality traits (Golbeck et al., 2011), or psychological states such as emotions, or medical conditions like mental disorders (De Choudhury et al., 2013). Such automatic methods enable large-scale social media data analyses even for (combinations

of) variables for which results from surveys are not available. Therefore, personality profiling in social media helps to paint a more comprehensive, complete, and timely picture for parts of a society.

State-of-the-art models reconstruct personality traits or mental health states from posts of social media users by relying on ground-truth data that links such posts to the correct annotation (Guntuku et al., 2017). The ground-truth data is typically obtained by either (1) asking participants to complete a validated survey that measures the desired variable and asking the participants to share their social media profiles, (2), by relying on self-reports of users, e.g., disclosure of a condition in the user's profile description, or (3), by having experts annotate profiles for particular properties. The quality of data obtained might therefore suffer from social-desirability bias, from being a non-representative subsample, from a lack of validated diagnoses, or from noise stemming from the challenge that annotators do not have access to the actual characteristics of users (Ernala et al., 2019).

We explore another route for which we hypothesize that it addresses these issues, but at the cost of only having access to very small data sets: We propose to leverage the existing set of high-quality, validated, and reliable psychometric instruments to measure psychological traits *directly*. Psychometric tests often come in the form of questionnaires which contain items, allowing a person to report

Variable	Cor.	Item Text
Openness	+	"Am interested in many things."
Openness	−	"Do not like art."
Extraversion	+	"Warm up quickly to others."
Extraversion	−	"Am hard to get to know."

Table 1: Example items from a psychometric test to assess personality traits (Lee and Ashton, 2018). 'Cor.' indicates if the item has been shown to correlate positively or negatively to the respective concept.

about themselves. These items are sentences formulated as descriptions of the self (Table 1 shows some examples). This structure motivates our hypothesis that such psychometric tests can be used directly to induce classifiers that profile individuals in social media without the existence of designated, manually annotated in-domain training data. If indeed possible, this would lead to a straightforward route to develop a myriad of classifiers for all those concepts for which psychometric tests exist. To dampen the issue of these sets of items being comparably small, we make use of pre-trained language models (Howard and Ruder, 2018; Devlin et al., 2019; Brown et al., 2020) to transfer knowledge acquired through pretraining rich semantic representations. Some subtypes of such models can be considered few-shot learners (Brown et al., 2020; Ruder et al., 2019), however, the transfer might not be successful to data outside of the pre-training domain. Therefore, we evaluate if various data augmentation methods can further leverage the challenges coming with such small corpora.

Thus, our contributions in this paper are that we (1) assemble a corpus from publicly available psychometric tests for the 'Big Five' variables of openness, conscientiousness, extraversion, agreeableness, and neuroticism (Costa and McCrae, 1992), which have been shown to be principled factors of personality (Cattell, 1945). Based on these data, we (2) fine-tune BERT (Devlin et al., 2019) and evaluate it on an existing personality trait corpus (Rangel et al., 2015). Furthermore, (3) we evaluate three data augmentation methods, namely paraphrasing with T5 (Raffel et al., 2020), and item generation with GPT-2 (Radford et al., 2019) and synonym replacements with Easy Data Augmentation (Wei and Zou, 2019). Our results, (4), show that the models perform en par with in-domain training for 4/5 personality trait variables.

2 Related Work

Psychometric Personality Tests. A psychometric test is a standardized instrument used to measure the cognitive, behavioral, or emotional characteristics of a person. One possible form are questionnaires, which can be designed for self-reporting. For each item the information is available if it is correlated positively or negatively with the concept to be measured. Publicly available psychometric tests can be found in various online repositories.[1]

An established test for personality traits following the so-called 'big five' variables is the *International Personality Item Pool Representation* of the NEO PI-R with 300 items[2] (IPIP-NEO-300, Goldberg et al., 1999). This test is a proxy of the *Revised NEO Personality Inventory* (NEO PI-R) by Costa and McCrae (1992), which is copyrighted and can only be ordered by professionals and used with permission. We use all items of the IPIP-NEO-300 as the source of our training corpus.

Another test of personality traits would be the *HEXACO Personality Inventory-Revised* (Lee and Ashton, 2008). It measures six factors of personality (Ashton et al., 2004) with 200 items, namely Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to experience.

Data. Psychometric tests found application in the analysis of social media user's personality in the past. An influential study has been the work by Schwartz et al. (2014), who collected Facebook data with a dedicated application (Stillwell and Kosinski, 2004) in which users completed the 100-item IPIP-NEO-100 questionnaire (Goldberg et al., 1999). The users further shared access to their status updates. This data is not available any longer.

The data for the PAN-author-profiling shared task in 2015 has been collected in a similar way (Rangel et al., 2015).[3] It consists of Tweets of 294 English Twitter profiles (besides Spanish, Italian and Dutch Twitter profiles), which are annotated with gender, age, and the 'Big Five' personality traits. The personality traits were self-assessed by the Twitter users with the BFI-10 (Rammstedt and John, 2007), which is an economic psychometric test that allows the personality to be recorded with only 10 items. We use this corpus for evaluation.

Combining Tests and Social Media Data. An interesting combination of psychometric tests with social media posts, which is likely the one most similar to our paper, is the work by Vu et al. (2020). The authors make use of social media data of users to automatically fill the IPIP-NEO (Goldberg et al., 1999) psychometric test to predict the social media user's 'Big Five' personality traits. They do so by embedding sentences and items with BERT into the same distributional space, followed by a k-nearest-

---

[1] https://psychology-tools.com/, https://ipip.ori.org/, https://www.psychometrictest.org.uk/

[2] http://www.personal.psu.edu/~j5j/IPIP/

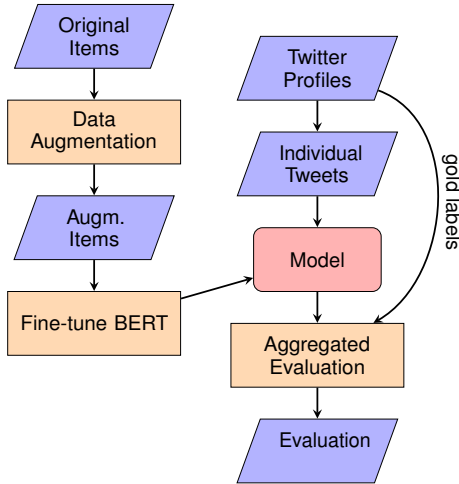[3] https://zenodo.org/record/3745945,https://pan.webis.de/

Figure 1: Workflow of our approach.

neighbor classification. This approach constitutes the opposing approach that we chose in our paper – Vu et al. (2020) use social media data to fill a psychometric test. We use psychometric tests to classify social media data.

We refer the reader to Stajner and Yenikent (2020) for a more comprehensive overview of related work.

3 Methods

3.1 Workflow

We depict the general workflow in Figure 1. The original items from the questionnaire are first augmented. The resulting augmented items inherit the labels from the respective original items. We fine-tune BERT with these items which leads to a model to make predictions for comparably short instances, like Tweets. From the labeled corpus of Twitter profiles, we obtain labels for each individual tweet with the BERT-based model and then aggregate the individual labels to obtain a label for the whole profile. In the evaluation, this predicted profile label is compared to the annotated gold label.

3.2 Corpora

We use all items of the psychometric test IPIP-NEO-300 (Goldberg et al., 1999) as training data and label each item following the evaluation guidelines accompanying the IPIP-NEO-300 (see also Table 1). These guidelines provide the information if a confirmative answer to the item indicates a positive correlation or negative correlation with the target variable, which leads to a binary label.

For evaluation, we use the English subset of the PAN-author-profiling-2015 data (Rangel et al.,

	IPIP-NEO		Profiles		Tweets	
Class.	+	−	+	−	+	−
Open.	28	32	288	3	26,743	236
Consc.	31	29	229	15	21,391	1,428
Extra.	36	24	235	21	21,686	2,000
Agree.	24	36	223	29	20,441	2,831
Neurot.	33	27	76	197	18,076	7,168

Table 2: Corpus Statistics regarding the Twitter evaluation data and the IPIP-NEO-300-based training corpus. We extracted all profiles with positive or negative scores and excluded profiles with neutral scores.

2015) with annotated Twitter profiles. Table 2 summarizes the corpus statistics. Note that the distribution of the items from the test data is skewed towards positive instances – this might be a direct consequence of people with particular personality traits being more likely to share particular information on social media.

3.3 Classification Model

As our source domain, we consider a set of items $Q_C = \{(q_i, y_i)\}_{i=1}^n$ from a reliable psychometric test. Each of these items corresponds to one psychological concept C and consists of the item text q_i and the label $y_i \in \{\text{pos}, \text{neg}\}$ which stems from the evaluation guidelines for this test.

The task is to find a parameterized function $f_{C,\lambda}(u)$ which takes as input all posts of a user u and predicts a label for each concept C. The important aspect in our setup is that the parameters λ are only optimized on the psychometric data Q_C. This is a mismatch – we train a classifier to label short texts but need as output a prediction for a set of tweets which represents the user. Hence, to obtain a label for each user, we aggregate the labels for all their posts by accepting the majority class, for each concept separately.

To obtain the text classifier, we fine-tune BERT (Devlin et al., 2019) to approximate each function $f_{C,\lambda}$, based on *bert-base-uncased*. The sequence classification head is randomly initialized on top of the encoder.[4] For each concept C, we fine-tune a separate BERT model (no multi-task learning).

3.4 Data Augmentation

With 60 items per personality trait, our training corpora are small. To address this issue, we perform data augmentation with three different methods.

---

[4] https://huggingface.co/transformers/model_doc/bert.html#bertforsequenceclassification
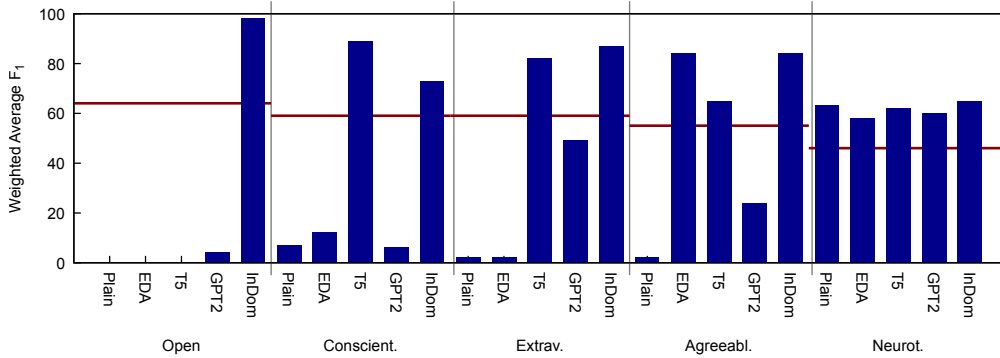
Figure 2: F_1 scores for all models and classes. Horizontal lines depict the baseline.

For every n instances $(q_i, y_i) \in Q_C$, we perform each data augmentation m times (obtaining $n \cdot m$ instances). Thus, we generate m augmented items q_i^a for each q_i. Each newly generated instance inherits the label y_i of its original instance q_i. We show examples for automatically generated items in the Appendix A.3.

Easy Data Augmentation. Easy Data Augmentation (EDA, Wei and Zou, 2019) consists of four operations on the sentence level: synonym replacement, random insertion, random deletion, and random swap. We use the default parameter of 10% of words in the sentence being changed (30% for random deletion) to perform each operation of EDA on each sentence (item) 5 times, hence generate 20 instances out of each original instance. This leads to 1,160 items for openness, 1,130 for conscientiousness, 1,080 for extraversion, 1,160 for agreeableness, and 1080 for neuroticism.

T5 item paraphrasing. We use T5 (Raffel et al., 2020) to paraphrase each item, based on the *T5ForConditionalGeneration* model provided by HuggingFace[5]. We do not perform fine-tuning to our domain, but rely on the original pre-trained parameters. For each item, we generate up to 50 paraphrases which leads to 2,285 items for openness, 2,383 for conscientiousness, 2,149 for extraversion, 2,126 for agreeableness, 2,130 for neuroticism.

GPT-2 item generation. We fine-tune GPT-2 (Radford et al., 2019) for each personality trait separately in 150 epochs, based on gpt-2-simple[6]. We generate 3000 items for each class label with a sentence length of 100 tokens and a temperature of 1.5. This leads to 6,279 items for openness, 6,177 for conscientiousness, 6,204 for extraversion, 6,271 for agreeableness, 6,242 for neuroticism.

4 Experiments

4.1 Experimental Settings

We split the psychometric test data to 80 % for training and use 20 % for hyperparameter optimization, while we ensure that augmented items stay in one set with their original item.[7] To avoid overfitting, we apply early stopping via observing the loss on the validation data. The maximum number of epochs is set to 200.

For a comparison to an "upper-bound" of in-domain training on Twitter, we split the corpora of social media profiles such that 50% of the Twitter profiles are in the test set. The remaining 50 % are used for training and further split into 90 % for training and 10 % for hyperparameter optimization of the in-domain model. The settings for fine-tuning the in-domain models are identical to the settings of the psychometric models.

4.2 Results

We show our main results as weighted F_1 values in Figure 2 (complete results in Table 4 in the Appendix). We compare the "plain" models without data augmentation to the augmented methods (as bar plots) and a random baseline (as horizontal line). We further show the performance of the in-domain model.

All "plain", non-data augmented models get outperformed by the random baseline, except for the personality trait neuroticism (F_1 =.63 versus F_1 =.46 random baseline). The plain psychometric models are inferior to the in-domain models for all concepts, but to various extends: Neuroticism is the only trait where the plain model shows a performance *en par* with the in-domain model.

Regarding the augmentation methods, T5 shows considerable improvements for conscientiousness, extraversion, and agreeableness (F_1 =.89, F_1 =.82, F_1 =.65, respectively, vs. .73, .87, .84 for in-domain models). This is also the best-performing augmentation method for conscientiousness and extraversion, however, EDA shows a further improvement for agreeableness (.84). T5 does not harm the performance for neuroticism in comparison to the plain model. Therefore, we conclude that T5 augmentation is a promising choice for 4/5 traits, while the other augmentation methods appear less stable in their contribution.

In summary, we obtain a substantial model performance without the use of in-domain training data for Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The transfer to or the difficulty of these concepts appears not to be the same, the performance for Conscientiousness is substantially higher than for Neuroticism. These results can only be partially compared to previous work due to the differences in the evaluation setup. However, it should be noted that the concepts that appear to be more challenging in our setup show also lower evaluation measures in related work (see for instance Table 3 in Rangel et al., 2015, note that their evaluation measure is an RSME, lower is therefore better).

4.3 Model Introspection

To provide some insights on the decision process by the classification models, we provide one example for each personality trait from the Tweet corpus with LIME explanations (Ribeiro et al., 2016) in Table 3. In the example for openness, the classifier relies on the word "love" as a positive indicator. This word can indeed be found in items from the test, namely in "Love to daydream", "Love flowers", and "Love to read challenging material". It is also a term that appears frequently in augmented data, such as in "Love problem solving" or in "Love flowers. Is it not hard to tell if you like something that's especially beautiful?". A positive indicator for conscientiousness is the word "August" and "year". This is interesting, given that these words appear not to be directly related to conscientiousness, and they do not appear in the original items of the test. However, the augmented data contains items that refer to "year", such as in "I truly love Excel and have used it for years.".

T	Tweet
O	@username What my love life will hold instore for me in the future. I'd never ask when I'm gonna die...????????
C	"@username: @username I like your profile photo. Very nice!!! You look very pretty. :)" THANK YOU! Took this photo in August this year.
E	@username Slade!!! Cool memories of my grammar school days!!
A	@username I rocked so much to their music!
N	"@username: Karma has no menu. You get served what you deserve."

Table 3: Examples of LIME explanations. Green indicates a positive contribution of the word, and red a negative contribution. The augmentation approach used in each example is the best-performing method for the respective concept. All examples are true positives.

5 Conclusion & Future Work

We outlined a novel methodology for automatic author profiling in social media users without a costly collection of annotated social media data. Instead, we directly train on items from validated psychometric tests. This data selection procedure has some advantages: items of psychometric tests are carefully validated textual instances. Such corpora of such items constitute "small data", but are available for a large number of concepts. Therefore, developing a method to induce classifiers directly from psychometric tests is also a promising avenue for future research.

For the tasks of developing models measuring the big five personality traits, we tested on Twitter data that has been collected by asking users to fill out a (different) test. The transfer appears to be achievable, we obtain results for four out of five personality traits which are *en par* with in-domain models, using T5 data augmentation (except Openness, which has very few test instances).

An important remaining research question is how models can be obtained that show consistently good results across concepts. In a real-world setup, test data from the target domain would not be available to make model selection decisions. One way to go might be to combine various augmentation methods. Another approach would be to use items as prompts in a zero-shot learning setup.

Acknowledgements

Ethical considerations

The fact that the current research deals with the sensitive topic of personality warrants for some ethical considerations. First, the study has been conducted with anonymized publicly available data. We did not collect data ourselves and importantly the data did not allow to identify subjects. Therefore, it is neither required nor possible to request IRB approval for the current research, given that IRB is concerned with the protection of human subjects. We had no reasons to doubt that the parties, who originally collected the data got IRB approval and informed consent form the participants who provided their data.

However, we acknowledge that automatic systems for personality trait analysis can be misused. Further, the application of our proposed model creation strategy can also be used for other more sensible concepts, for instance regarding mental health. We propose that such systems are only made available in such a manner that no personalized results can be retrieved.

References

Michael C Ashton, Kibeom Lee, Marco Perugini, Piotr Szarota, Reinout E De Vries, Lisa Di Blas, Kathleen Boies, and Boele De Raad. 2004. A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages. *Journal of personality and social psychology*, 86(2):356–366.

Ivan Bilan and Desislava Zhekova. 2016. Caps: A cross-genre author profiling system. In *CLEF (Working Notes)*, pages 824–835.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Raymond B Cattell. 1945. The description of personality: Principles and findings in a factor analysis. *The American journal of psychology*, 58(1):69–90.

Paul T Costa and Robert R McCrae. 1992. *Neo personality inventory-revised (NEO PI-R)*. Psychological Assessment Resources Odessa, FL.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *International AAAI Conference on Web and Social Media (ICWSM)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16.

Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems*, pages 253–262.

Lewis R Goldberg et al. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28.

Sharath C Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49. Big data in the behavioural sciences.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kibeom Lee and Michael C Ashton. 2008. The hexaco personality inventory: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, 39(2):340–345.

Kibeom Lee and Michael C Ashton. 2018. Psychometric properties of the hexaco-100. *Assessment*, 25(5):543–556.

Wincenty Lutoslawski. 1898. Principes de stylométrie appliqués à la chronologie des œuvres de platon. *Revue des études grecques*, 11(41):61–81.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212.

Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at PAN 2015. In *CLEF*, page 2015. sn.

Marco T Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 118–125.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Sanja Stajner and Seren Yenikent. 2020. A survey of automatic personality detection from texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6284–6295, Barcelona, Spain (Online). International Committee on Computational Linguistics.

David J Stillwell and Michal Kosinski. 2004. mypersonality project: Example of successful utilization of online social networks for large-scale social research. *American Psychologist*, 59(2):93–104.

Huy Vu, Suhaib Abdurahman, Sudeep Bhatia, and Lyle Ungar. 2020. Predicting responses to psychological questionnaires from participants' social media posts and question text embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1512–1524.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

A Appendix

A.1 Detailed Results per Class

		Psychometric Models												in-domain			Baseline		
		Plain			EDA			T5			GPT-2								
	Class	P	R	F₁	P	R	F₁	P	R	F₁	P	R	F₁	P	R	F₁	P	R	F₁
Open.	−	.01	1.00	.01	.02	1.00	.04	.02	1.00	.04	**.02**	**1.00**	**.04**	.0	.0	.0	.01	.33	.03
	+	.0	.0	.0	.0	.0	.0	.0	.0	.0	**1.00**	**.02**	**.04**	.99	1.00	.99	.97	.50	.66
	AVG	.0	.50	.01	.01	.50	.02	.01	.50	.02	**.51**	**.51**	**.04**	.49	.50	.50	.49	.41	.34
	w-avg	.0	.01	.00	.0	.02	.0	.0	.02	.0	**.98**	**.04**	**.04**	.97	.99	.98	.95	.49	.64
Consc.	−	.07	.89	.13	.08	1.00	.15	**.0**	**.0**	**.0**	.08	1.00	.14	1.00	.17	.29	.05	.33	.09
	+	.80	.04	.07	1.00	.06	.12	**.93**	**.99**	**.96**	1.00	.03	.05	.96	1.00	.98	.90	.48	.63
	AVG	.43	.46	.10	.54	.53	.13	**.46**	**.50**	**.48**	.54	.51	.10	.96	.96	.94	.47	.41	.36
	w-avg	.75	.10	.07	.93	.13	.12	**.86**	**.92**	**.89**	.93	.10	.06	.78	.80	.73	.84	.47	.59
Extrav.	−	.09	1.00	.17	.09	1.00	.17	**.0**	**.0**	**.0**	.06	.42	.11	1.00	.14	.25	.08	.42	.13
	+	.0	.0	.0	.0	.0	.0	**.90**	**.92**	**.91**	.87	.38	.53	.90	1.00	.95	.89	.49	.63
	AVG	.05	.50	.08	.05	.50	.08	**.45**	**.46**	**.45**	.46	.40	.32	.95	.57	.60	.48	.45	.38
	w-avg	.01	.09	.02	.01	.09	.02	**.82**	**.83**	**.82**	.79	.38	.49	.91	.91	.87	.82	.48	.59
Agree.	−	.10	1.00	.19	**.18**	**.15**	**.17**	.08	.31	.12	.08	.62	.14	.0	.0	.0	.05	.23	.08
	+	.0	.0	.0	**.91**	**.92**	**.91**	.88	.59	.71	.77	.15	.25	.89	1.00	.94	.84	.46	.60
	AVG	.05	.50	.09	**.54**	**.54**	**.54**	.48	.45	.42	.42	.38	.19	.44	.50	.47	.44	.35	.34
	w-avg	.01	.10	.02	**.83**	**.84**	**.84**	.80	.56	.65	.70	.20	.24	.79	.89	.84	.76	.44	.55
Neur.	−	**.73**	**.93**	**.81**	.71	.78	.73	.73	.90	.80	.72	.91	.80	.73	1.00	.84	.66	.45	.53
	+	**.25**	**.06**	**.09**	.16	.10	.12	.23	.08	.12	10	.03	.04	1.00	.12	.22	.19	.36	.25
	AVG	**.49**	**.50**	**.46**	.43	.45	.44	.47	.49	.46	.41	.47	.42	.86	.56	.53	.43	.41	.39
	w-avg	**.60**	**.70**	**.63**	.56	.60	.58	.58	.68	.62	.55	.66	.60	.81	.74	.65	.53	.43	.46

Table 4: Detailed results for *Psychometric Models* vs. *in-domain Models* vs. *Random Baseline* for psychological traits in Twitter users. The random baseline generates predictions by respecting the training sets' class distribution. The weighted average values for P, R, F₁ correspond to the average across all labels considering the proportion for each label in the data set. The bold typo highlights our best performing model w.r.t. the highest w-avg. −: scored negative, +: scored positive.

A.2 Implementation Details

We performed the experiments on 4 NVIDIA GeForce GTX 1080 Ti GPUs with Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GH. The number of parameters is defined by the base model that we used, namely BERT base, with 110 M parameters.

We show the run-time of models (training + testing) in Table 5. The numbers do not include startup/loading times. Note that the test data is (sometimes dramatically) larger than the training data.

	Model			
Concept	Plain	EDA	T5	GPT-2
Depression	2460+3900	360+3900	900+3960	900+3960
Anxiety	450+2850	407+2905	921+2615	374+2584
ADHD	550+2650	380+2856	720+2100	875+2175
Openness	180+120	120+120	420+120	780+120
Conscientiousness	480+120	180+120	780+120	360+120
Extraversion	1020+120	180+120	420+120	720+120
Agreeableness	60+120	84+120	855+120	450+120
Neuroticism	474+120	85+120	400+120	650+120

Table 5: Runtime of models in seconds (train+test).

A.3 Examples for Augmentation Methods

- As an example for the EDA augmentation method, synonym replacement lead to "Love thinking about things" based on the IPIP-NEO-300 item "Enjoy thinking about things" for the trait of openness.

- As an example for the T5 augmentation method, T5-paraphrasing lead to "Have fun and be wildly inspired by wild fantasy dreams" based on the IPIP-NEO-300 item "Enjoy wild flights of fantasy" for the trait of openness.

- An example for a GPT-2 generated item measuring agreeableness is "I am an average person".