Automated personality assessment

Hakan Akyürek
Technical University of Munich
hakan.akyuerek@tum.de

Stefan Frisch
Technical University of Munich
stefan.frisch@tum.de

Abstract

Personality assessment is an important topic for researchers and companies, as our personality profoundly influences many of our life choices. In this work, we analyze the task of automated personality assessment by employing only written text. We explore two personality prediction datasets and introduce a new Twitter-based Myers-Briggs personality type dataset, where Twitter users are assigned to one class depending on their self-reported test results. To predict personalities from written text, we tried classification and siamese-based approaches. However, with both approaches, the results are not very promising and are hardly above random guessing for all but one instance. In this case, our models could reliably predict if a person is more into Thinking vs. Feeling and Judging vs. Perceiving.

1 Introduction

Our personality influences almost all life choices that we take. Therefore it is immensely valuable for businesses and science to understand in detail how the different personality traits affect our decision-making and adequately assess an individual's personality. Several frameworks for personality trait categorization have been proposed, like the Big Five personality factors (Raad, 2000) and Myers-Briggs type indicator (Myers, 1962). Both personality assessment types have been extensively reviewed (Cobb-Clark and Schurer, 2012; Pittenger, 1993) and are widely used in practice. Furthermore, it is well known that job performance is linked to personality, where different positions require unique personal characteristics (Ashton, 1998; Tett and Burnett, 2003). Thus, it is unsurprising that companies take a great interest in the assessment process. Additionally, training and work environments tailored explicitly to personality types can lead to higher performance in general and a more satisfactory experience for the individual.

Personality assessment is usually done manually by professionals and is a costly task. Even if a questionnaire is used to learn a person's personality type, it needs to be evaluated by a trained professional, as the questionnaires often have open-ended questions. However, those questionnaires are prone to manipulation if the tested individual aims for a specific result. Additionally, it is possible that the questionnaires do not represent the personality correctly, making the assessment challenging. Therefore, we aimed to explore an automated personality trait assessment based on written text on social media or free-form text.

2 Related work

Automated personality assessment in a similar way has been explored before by Mehta et al. (Mehta et al., 2020). There is a flaw in their methodology that we will illustrate in the following and address in our study. In particular, they did not specify a fixed number of training epochs and reported results only on a validation set. For this validation set, they evaluated the performance of their models after every epoch. Finally, they reported the accuracy of their models by cherry-picking the best accuracy among the validation set accuracies over all the epochs. This clearly leads to an artificially increased accuracy that is most likely due to chance alone. This is especially a problem since the datasets are rather small. We illustrated the shortcoming of this approach in table 3.

3 Dataset

There are only a very limited number of quality datasets for the task of personality assessment available. One big challenge is that labeling the data is difficult since either a trained professional has to do the personality assessment, which is expensive, or a questionnaire can be employed, which can be subject to manipulation by the participants. We tested our methods on the following publicly available datasets:

Essays Big Five personality factors (BFPF): This dataset consists of 2468 stream-of-consciousness essays written by students and with binary labels for the Big Five personality traits that were found by a self-report questionnaire (Pennebaker, 1999). The labels stand for EXT: extraversion, NEU: neuroticism, AGR: agreeableness, CON: conscientiousness, and OPN: openness to experience.

PersonalityCafe Myers-Briggs type indicator (MBTI): In this dataset consists of the last 50 tweets of an individual on an online forum called PersonalityCafe together with their Myers-Briggs personality types (Čerkez et al., 2021).

Twitter MBTI: Furthermore, we also created a dataset based on self-reported Myers-Briggs personality tweets that were written in English. We only included tweets that are not replies or retweets and do not have any links or media. To find out the twitter users that reported their own MBTI personality type we used the following query on the Twitter API (Twitter, 2022):

```
lang:en -has:media -is:retweet -
   has:links -is:reply -"’s mbti"
    -"’s personality"
```

On the results of the query we run the following regex pattern to filter the results in a more controllable manner:

```
’^.*((my.*personality|me being|my
    mbti|am\s)|([iI]['\s](am|m\s|
   got|ve|have|went|used)))+(\s
   |.\w+.){0,4}#*%s+.*$’ %
   personality_type
```

With this query we extracted 1074 tweet histories for twitter users in total on the 10.08.2022. The user distribution on the personality types are shown in Figure 1. After further pre-processing we analyzed the tweet histories of 941 twitter users. We disregarded non-english tweets and users who mentioned that they did a personality test several times and got different traits.

The labels of the Myers-Briggs type indicator stand for: Extraversion vs. Introversion (E/I), Sensing vs. Intuition (S/N), Thinking vs. Feeling (T/F), and Judging vs. Perceiving (J/P).

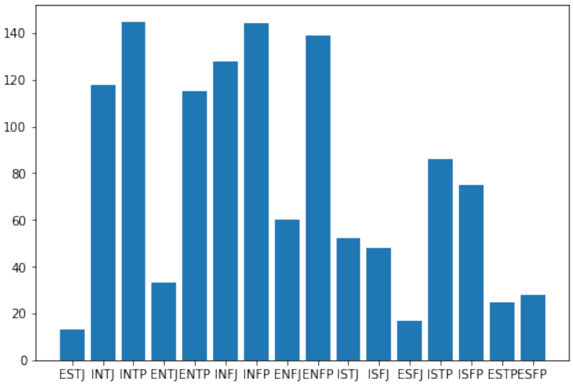

Figure 1: User counts per personality in Twitter Data

3.1 Essays Dataset Exploration

Over the course of the research, we also conducted a dept exploration of the Essays dataset. Here we are not going to share all of the explorations, but only the most interesting results.

The exploration is done on three levels:

- Text level: We did exploration directly on the essays themselves. These explorations include usually word counts and similar measures.
- Topic level: Here we explore the topics of the essays with the help of topic modeling.
- LIWC level: Lastly, we explore the LIWC (Pennebaker et al., 2015) values of the essays.

These explorations are also done in three categories each:

- Encoding-wise: Here we care about only the classes of each essay.
- Label-wise: Here we care about the classes themselves, so the essays that are part of each class.
- One-hot encoding-wise: Here we only do exploration on one class essays.

As it can be seen in Table 1 the 5 top topics for each personality have considerable similarities. We believe this is because the essays were written on the same roof topic. This implies that the models may not differentiate the personalities only on the topics. We also checked topics for one-label essays. Also, as it can be seen in Table 1 the topics for those essays have also similarities and also not in the same order. However, we believe that there isn’t a distinctive difference.

Personality	label-wise	one-hot encoding-wise
EXT	28, 2, 9, 39, 6	28, 38, 8, 9, 3
NEU	28, 2, 9, 39, 20	28, 20, 36, 2, 3
AGR	28, 2, 9, 8, 36	28, 8, 2, 20, 7
CON	28, 2, 9, 8, 36	28, 39, 8, 9, 20
OPN	28, 2, 9, 20, 3	2, 28, 20, 31, 3

Table 1: Top topics for one-hot encoding-wise and label-wise, presented in numbers

LIWC Values	EXT	REST
Social	8.74	8.63
Analytic	13.8	14.31
BigWords	11.95	12.04
WPS	24.28	23.47
shehe	1.62	1.58
negate*	2.27	2.31
verb*	21.94	21.94
i*	10.75	10.70

Table 2: Some LIWC values between extroverted essays and non-extroverted essays

Additionally, we saw interesting results in some of the LIWC values when we checked them for each label. For instance in Table 2 we can see that the difference of some LIWC values between extroverted essays and non-extroverted essays. Even though a lot of values follow the extroverted pattern there are values that do not. For instance, following (Mairesse et al., 2007) we can see that extroverted people often talk less analytical and usually not self focused. In a stream-of-consciousness setting, the essays were written in that setting, the extroverted people use third person pronouns more than the rest and construct less analytical sentences with less complex words. However, we see that they use less verbs than the rest, which is not characteristic.

Overall, we can see from the exploration there isn't a big separation between the personalities in the dataset, which indicates that the task is difficult to learn.

4 Method

Two approaches have been explored in this study: classification with large language models and a siamese-based approach also with large language models.

4.1 Classification with large language models

A logical approach is to formulate the problem of personality assessment as a classification task. Our models take as input the text or tweets of a person and output the personality traits. We have explored two approaches predicting all 5 personalities at once or having a separate model for each personality. Predicting all personalities at once led to a performance loss for the Big Five personality types in the essay dataset, which can be seen in figure 3. A detailed analysis can be found in the results. We, therefore, explored the prediction of all personalities separately in greater detail. The basic structure of our architecture consists of a language encoding module and a classification module on top. We experimented with BERT (Devlin et al., 2019), and Longformer (Beltagy et al., 2020) as a language encoder since BERT was used by other researchers for this task before and Longformer because the texts in the Essays dataset are much longer than the 512 token limit from BERT. For the classification module, we took the pooler output from Bert and the last hidden state of the CLS token for the Longformer. A Multi-Layer Perceptron (MLP) reduces this output to a single number. Then, by applying a sigmoid, the number can be interpreted as the probability that the predicted personality trait is present.

4.2 Siamese-based approach

One other approach we tried was checking document similarities for each personality and assigning them to the respective personality if the document is similar enough. For this purpose, we worked on Siamese Networks, more specifically with S-Bert (Reimers and Gurevych, 2019).

Like the previous approach, we experimented with predicting personalities separately and combined. Additionally, we tried a few different approaches to generate text pairs:

- Summaries of n number of documents as a representative for each personality. The summaries were extracted using (Xiao et al., 2021).
- Choosing n documents for each personality and generating n pairs for each personality

It should be noted that these n documents for pair generation were chosen randomly from one-label essays.

Additionally, we worked on training separate networks for each personality rather than having a single network to calculate the distance between documents with the second approach of pair generation. The basic structure uses only one network for each distance measurement. Instead, we have a separate network for each personality specifically.

5 Results

The task of predicting personalities from textual data is difficult. To put our results into relation, we compare our work Mehta et al. (Mehta et al., 2020). We have previously already mentioned that there are shortcomings in their methodology. In table 3 we illustrate that reporting the results on the best epoch for each cross-validation set leads to artificially boosted results on the Essays BFPF dataset. We also found that fine-tuning Bert for classification is especially prone to instability, particularly for the smaller datasets (Essays BFPF and Twitter MBIT) we have tested. This follows the existing research on the instability of Bert depending on the random seed (Mosbach et al., 2020) (Dodge et al., 2020). We combat this instability by training for more epochs than standard fine-tuning and selecting random seeds where the training loss improves.

Our own results for the classification with large language models can be observed in the table 4 for the Essays BFPF dataset, in the table 7 for the Twitter MBTI dataset, and in the table 6 for the PersonalityCafe MBTI dataset. We found that fine-tuning Bert helps the prediction accuracy in almost all cases. However, the improvement over random guessing and the baseline (Mehta et al., 2020) are rather small. For 4 of 5 personalities, the improvement over random guessing is less than 5% for the Essays dataset. The results for our gathered Twitter dataset are even worse. Our models achieved less than 3% improvement over random guessing for all types. On the two types Thinking vs. Feeling (T/F) and Judging vs. Perceiving (J/P), we can report better results for the PersonalityCafe dataset. Our model achieved 29% and 17% over random guessing, respectively. This shows that automated personality classification from text is still a challenging task. It would be an interesting study if this task could be reliably solved by a trained professional. The performance spike for the two classes in the PersonalityCafe dataset is probably due to 2 reasons. On the one hand, users who sign up on the PersonalityCafe platform are, in all likelihood, very personality conscious, and this will be reflected in their writing style. On the other hand, are these two types probably easier to differentiate than the others. We also found evidence in our gathered Twitter dataset for this hypothesis.

Additionally, we explored how much textual data Bert and other large language models need to form reasonable predictions. The results can be observed in table 5. With a token limit of 512, one can see that Bert performs best on 3 out of the 5 personalities. It also becomes evident that the performance improves the more textual data we give the Bert model. However, the performance decreased when we tried to include even more tokens in the classification by introducing the Longformer to our task. The Longformer-based architecture could only improve on one category over the Bert language encoder, even though we give the model at least twice as much textual data. The partial attention mechanism of the Longformer architecture seems to hurt performance more than the gains of additional text.

Our results with the S-Bert also did not show any promising results. In Table 8, we present the results of the different approaches to generating pairs for the Siamese network. We used a single network for both experiments, and the training results did not differ and were negligibly better than random guessing. However, we can also see that in OPN trait, we achieve higher results, similar to other approaches we have tried. Also, as it can be seen, having a single network works worse than having separate networks for each personality, regardless of the number of documents used for pairing. The results show that using more pairs (in the case of the second pair generation approach) provides better results. However, they also are not a significant improvement over random guessing. Additionally, we tried training the network with summaries of these documents, shown in the table. The results are similar when a singular model is used.

6 Conclusion

Automated personality assessment remains a challenging task. Even with the recent advances in natural language processing techniques like transformers and unsupervised training of language models, it is still impossible to reliably assess an individual's personality from their social media posts or free written text.

The patterns in the data for personality assess-

Model	EXT	NEU	AGR	CON	OPN
Random guessing	51.68	50.02	53.06	50.83	51.52
Mehta et al. variable training	**60.00**	**60.50**	**58.80**	**59.20**	**64.60**
Mehta et al. 7 epochs training	54.56	55.78	56.10	56.79	60.03

Table 3: The results of the paper Mehta et al. with the variable training length, that was reported in their paper are compared to a fixed training length of 7 epochs for the Essays dataset. 7 epochs amount to the maximum performance.

Model	EXT	NEU	AGR	CON	OPN
Random guessing	51.68	50.02	53.06	50.83	51.52
Mehta et al. 7 epochs training	54.56	**55.78**	56.10	**56.79**	60.03
Ours 5 epochs training Bert 512 tokens	**55.56**	54.85	**58.26**	55.86	**63.57**

Table 4: The results of the paper Mehta et al. with the variable training length, that was reported in their paper are compared to a fixed training length of 7 epochs for the Essays dataset. 7 epochs amount to the maximum performance.

Model	EXT	NEU	AGR	CON	OPN
Random guessing	51.68	50.02	53.06	50.83	51.52
Bert 50	48.05	47.45	56.16	50.75	50.92
Bert 100	52.25	55.26	51.65	54.65	57.36
Bert 200	55.56	**55.56**	53.15	51.95	51.65
Bert 300	57.66	50.15	53.15	50.75	59.76
Bert 400	54.35	54.95	54.05	52.85	59.76
Bert 512	55.56	54.85	**58.26**	**55.86**	**63.57**
Longformer 1024	**57.96**	53.15	51.05	55.56	57.66
Longformer 4096	57.36	48.65	52.55	49.25	54.05

Table 5: Performance comparison of token limit for the language transformer Bert and Longformer on the Essays dataset in the classification setting. The model names stand for which language transformer was employed and the respective token limit. So, for example, Bert 50 would indicate that Bert was the language encoding module and the token limit for Bert was 50. It is important to note that for the Longformer 4096, the model ran only on half-precision since we faced a GPU memory constraint of 16GB. For the same reason, the batch size was reduced from the standard 16 in Bert to 4 for the Longformer-based models. In addition, we accumulated the gradients to combat the highly varying gradients.

Model	E/I	S/N	T/F	J/P
Random guessing	76.96	86.20	54.11	60.41
Mehta et al. variable training	78.30	86.40	74.40	64.40
Ours 4 epochs training Bert 512 tokens	**82.68**	**89.51**	**83.53**	**77.65**

Table 6: Results of the Myers-Briggs type indicators for the PersonalityCafe dataset.

Model	E/I	S/N	T/F	J/P
Random guessing	64.83	74.61	50.38	61.64
Ours				
5 epochs training	**66.14**	**74.80**	**53.54**	**62.20**
Bert 512 tokens				

Table 7: Results of the Myers-Briggs type indicators for our Twitter dataset.

	AVG	EXT	NEU	AGR	CON	OPN
S-BERT seperate	55.84%	61.20%	51.60%	49.00%	55.10%	62.30%
S-BERT united (5)	48.84%	50.36%	48.60%	45.70%	47.34%	52.20%
S-BERT united (10)	50.14%	51.26%	50.30%	48.00%	47.00%	54.14%
S-BERT united (sum)	48.85%	49.95%	49.20%	46.30%	47.40%	51.40%

Table 8: Results of Siamese-based approach on the Essays dataset. (5) - 5 documents for pairing, (10) - 10 documents for pairing

ment do not seem to be noticeable. As discussed in subsection 3.1 the patterns for each personality in the Essays dataset are not distinguishable. This results in a task that is not solvable with current natural language understanding methods.

It would be interesting for future research to explore whether social media activity is a better indicator of personality traits. In particular, what an individual likes and on which posts the user spends the most time. However, one possible difficulty for this line of future research is that current social media companies do not share this data with researchers due to privacy concerns.

References

Michael C Ashton. 1998. Personality and job performance: The importance of narrow traits. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 19(3):289–303.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Deborah A Cobb-Clark and Stefanie Schurer. 2012. The stability of big-five personality traits. *Economics Letters*, 115(1):11–15.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.

Francois Mairesse, Marilyn Walker, Matthias Mehl, and Roger Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res. (JAIR)*, 30:457–500.

Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, E. Cambria, and Sauleh Eetemadi. 2020. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines.

Isabel Briggs Myers. 1962. The myers-briggs type indicator: Manual (1962).

James Pennebaker, Roger Booth, Ryan Boyd, and Martha Francis. 2015. Linguistic inquiry and word count: Liwc2015.

King Pennebaker. 1999. Linguistic styles: language use as an individual difference.

David J Pittenger. 1993. The utility of the myers-briggs type indicator. *Review of educational research*, 63(4):467–488.

Boele Raad. 2000. *The Big Five Personality Factors: The psycholexical approach to personality.*

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Robert P Tett and Dawn D Burnett. 2003. A personality trait-based interactionist model of job performance. *Journal of Applied psychology*, 88(3):500.

Twitter. 2022. Twitter api documentation.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. PRIMER: pyramid-based masked sentence pre-training for multi-document summarization. *CoRR*, abs/2110.08499.

Ninoslav Čerkez, Boris Vrdoljak, and Sandro Skansi. 2021. A method for mbti classification based on impact of class components. *IEEE Access*, PP:1–1.