Exploring the OCEAN Depths: a Look at Personality Trait Classification on Essays

Mateo Hrelja, Dominik Jurinčić, Ivan Linardić

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{mateo.hrelja, dominik.jurincic, ivan.linardic}@fer.hr

Abstract

Capturing personality traits from text is a challenging task wherein current available solutions leave a lot to be desired. In this paper, we explore a variety of approaches in tackling this task. We try various text preprocessing methods in tandem with several different machine learning models of varying complexity. Ultimately, we focus on the interpretability of our results and aim to provide better clarity on the reasoning behind them. We use two different approaches when evaluating interpretability in order to get an even better grasp of this difficult, yet interesting task.

1. Introduction

Human personality has been a fascinating area of research for many years now, as it plays a large part in better understanding human behavior and interactions. The Big Five personality traits (McCrae and Costa, 1985), often referred to as OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) serve as a reliable psychological model for describing the concept of personality for individuals. Over the years, advancements in natural language processing research and methods have made it possible and more reliable to capture human personality traits from text.

In this paper, we dive deeper into understanding how language models infer personality traits from stream-of-consciousness essays. We present several classification methods and models which we used for our results, as well as the effects of various text preprocessing and feature selection methods. As part of our research, we include both simple models which yield highly interpretable results as well as more complex deep models focused on performance.

The main goal of the paper is to delve deeper into the concept of interpretability and better understand certain patterns and insights carried within the writings of individuals, using different methods to interpret the obtained results in tandem with the use of different models.

2. Related Work

Previous work in this field constitutes of the idea that language-based assessments can give valid personality traits measures (Park et al., 2014) and that there exists correlation between language and personality traits which was proven by Schwartz et al. (2013). Also, they have performed experiments which showed that removing stop words harms the performance of their models, which was also useful in our case. Pizzolli and Strapparava (2019) predicted personality traits of characters in stories based on computational linguistics and gave a critic on the Big Five personality traits model we use.

Gjurković et al. (2021) found problems in quality and low number of personality-labeled datasets which induced them to create a new one which included personality and demographics information.

Recent work in deep learning, done on the Big Five personality model using the essays dataset introduced by Pennebaker and King (1999), was done in (Majumder et al., 2017). They used a convolutional neural network (CNN) feature extractor in which sentences were fed to convolution filters to obtain n-gram feature vector. State-of-the-art results on the essays dataset were achieved by Kerz et al. (2022) who took the transformer-based approach. Similar to Kerz et al. (2022), we use the Big Five personality traits and the pre-trained transformer language model BERT (Devlin et al., 2019).

3. Dataset

The essays dataset, introduced by Pennebaker and King (1999), consists of stream-of-consciousness essays written by psychology students. The use of stream-of-consciousness essays for personality trait classification seems appropriate, as the free form nature of the text encourages writers to express their thoughts and use language in a non-restricted way. The dataset contains 2467 essays and each is labeled with the personality scores of the writer for each of the Big Five personality traits. The labels are 'y' and 'n', representing scoring high or low for a given personality trait. The personality scores of students were obtained by filling out the Five Factor Inventory questionnaire (John et al., 1991).

4. Data Preparation and Feature Selection

The first step in the data preparation process was converting the 'y' and 'n' labels into their numerical versions 1 and 0. We then experimented with different text preprocessing techniques and feature extraction methods on the essay texts in order to find the ones which provide the best results. We used spaCy[1] and Natural Language Toolkit (NLTK)[2] to preprocess the raw essay text in several ways. In the end we had 4 versions of the essays: (1) the raw essay text, (2) the

[1] https://spacy.io/
[2] https://www.nltk.org/

Table 1: Words important for scoring high in a personality trait by coefficient magnitude

Trait	Model	Words
Openness	LR	like, cat, ll, too, maybe, re, love, music, of, you
Openness	SVM Linear	crazy, love, like, maybe, you, ll, cat, of, re, music
Agreeableness	LR	least, would, with, right, on, to, really, so, family, have
Agreeableness	SVM Linear	many, worried, would, least, so, right, family, on, with, have
Conscientiousness	LR	today, it, tonight, hope, party, and, the, my, he, to
Conscientiousness	SVM Linear	student, decision, today, couldn, tonight, my, hope, to, he, party
Neuroticism	LR	don, feel, everything, want, scared, money, me, sex, life, stressed
Neuroticism	SVM Linear	worry, this, scared, me, everything, life, boyfriend, sex, money, stressed
Extraversion	LR	is, its, sorority, and, love, boyfriend, fun, all, am, so
Extraversion	SVM Linear	ready, mean, if, love, its, all, am, boyfriend, fun, so

Table 2: Words important for scoring low in a personality trait by coefficient magnitude

Trait	Model	Words
Openness	LR	college, is, to, my, because, school, home, class, have, classes
Openness	SVM Linear	college, because, is, boyfriend, assignment, class, home, tomorrow, game, confuse
Agreeableness	LR	stupid, girlfriend, don, is, damn, read, more, nothing, same, no
Agreeableness	SVM Linear	stupid, girlfriend, is, read, damn, store, same, don, nothing, wont
Conscientiousness	LR	want, hate, don, this, think, re, wake, god, point, chance
Conscientiousness	SVM Linear	want, hate, this, wake, point, re, think, chance, don, music
Neuroticism	LR	its, would, her, many, semester, already, beat, mind, as, texas
Neuroticism	SVM Linear	its, many, her, would, already, pledge, semester, beat, glad, mind
Extraversion	LR	don, there, in, should, want, something, eyes, perhaps, very, mother
Extraversion	SVM Linear	don, should, in, something, there, want, eyes, very, perhaps, real

lowercase and lemmatized essay text, (3) the lowercase essay text in which we removed stopwords and punctuation, (4) the essay text in which we removed punctuation.

4.1. Feature Extraction

We used a modified grid search approach to determine the best feature extraction method for our bag-of-words models. The grid search was performed with the `TfidfVectorizer` and `CountVectorizer` from scikit-learn[3], with different values of the parameters which control the maximum number of features and the n-gram range. All the versions of the vectorizers were used on all 4 versions of the essays. A support vector machine (SVM) model was trained on 80% of the data and tested on the remaining 20% to determine which combination of essay version and text vectorizer was optimal. The vectorizers were fitted only using the training data to avoid data leakage. The best performance was obtained when using the raw essay text with the default `TfidfVectorizer`, which uses unigrams of words and has no limit on the number of features.

The fact that the best performance came from using unprocessed, raw essay text is not surprising. The specific use of punctuation, capitalization and word forms likely carries information about the personality traits of the writer.

We used average word embeddings from essays, using pretrained GloVe[4] embeddings (Pennington et al., 2014), as inputs in our fully-connected models. The inputs to our transformer models were contextualized embeddings from the `BertTokenizer` module from HuggingFace[5]. The contextualized embeddings were made from 200 word subsections of essays, with word overlapping set to 50 words.

5. Models

The given task is framed as a multi-label binary classification problem. As such, we tried several different models befitting of the task, which use the features we had previously extracted, including a random classifier as a baseline model.

Using a variety of models was done in order to tackle the problem from different angles. The simpler models, which use features extracted from the text directly, are more oriented towards the interpretability of the results. The two deep models are meant to provide better results, as well as display the advantage of using word embeddings as opposed to using features from the text - albeit, at the cost of interpretability.

Random Classifier The random classifier is our baseline model for this task. This simple model randomly assigns labels to data without considering any features or

[3]https://scikit-learn.org/stable/

[4]https://nlp.stanford.edu/projects/glove/

[5]https://huggingface.co/

patterns in the text or overall data.

LR The logistic regression (LR) models use the raw essay text converted into unigrams with the `TfidfVectorizer`. The two hyperparameters which were used in the optimization process are the inverse of regularization strength *C*, and the type of regularization which is applied - *L1* (Lasso) or *L2* (Ridge). The hyperparameters were optimized based on the average F1-score they yielded for all personality traits. We used scikit-learn's `LogisticRegression` implementation for our models.

SVM The approach we used for training the SVM models and their hyperparameter optimization is comparable to the LR models. Once again, the raw essay text was converted into unigrams based on the `TfidfVectorizer`, which were used as the models' features. The hyperparameters we optimized were the inverse of regularization strength *C*, and the type of kernel used in the SVM - Linear, Polynomial or Radial Basis Function (RBF). Equivalent to the approach in the LR models, the optimal hyperparameters were chosen based on the average F1-score in regards to all of the personality traits. We used scikit-learn's `SVC` implementation for our models.

FC The fully-connected deep models (FC) consist of (1) the input layer, which uses a 300-dimensional vector as its input features, representing the average value of the word embeddings for a given essay, obtained from GloVe; the output size of the input layer is 500, (2) a rectified linear unit (ReLU) activation function to induce non-linearity, (3) a fully-connected layer with the input size of 500 and output size of 150, (4) a ReLU activation function to induce non-linearity, (5) the output layer: a fully-connected layer with the input size of 150 and output size of 1. The models were trained on each trait separately, using a 60/20/20 train/validation/test split of the data. We used PyTorch[6] for the implementation of our models. The models were trained for 20 epochs using the `Adam`[7] optimizer with the learning rate of $1 \cdot 10^{-3}$.

Transformer Our transformer models are based on the idea presented in Pappagari et al. (2019). The model primarily relies on BERT, but alleviates BERT's weakness of poor performance on longer texts by splitting long input texts into chunks of 200 words with an overlap of 50 words between each. Each of these chunks goes through the BERT model in order to obtain its embedding. We used the aforementioned `BertTokenizer` module from HuggingFace to obtain the contextualized embeddings, as well as the `BertModel` module in tandem with PyTorch for the implementation of these models. The models were fine-tuned for 10 epochs using the `Adam` optimizer with the learning rate of $2 \cdot 10^{-5}$ and without bias correction.

6. Interpretability

The main benefit of using LR and linear kernel SVM models with the bag-of-words representations of text is the interpretability of these models. We performed experiments to see which features the models deemed the most important for classification.

The importance of a feature for a LR or SVM model with a linear kernel can be determined by the magnitude of the feature coefficient (Chang and Lin, 2008). The largest positive values of coefficients correspond to the features which are the most important for scoring high in a certain personality trait, while the largest negative values correspond to the features which are the most important for scoring low in a certain personality trait.

Another way to determine the importance of features is by using Shapley Additive Explanations (SHAP)[8] values, as proposed by Lundberg and Lee (2017). The values represent the impact each feature had on the output of the model for each example.

7. Results

The models used in this paper were tested for interpretability and performance. Interpretability is explained in detail in Section 6.. The performance results of the LR and SVM models were obtained by 5-fold cross-validation, whereas the fully-connected models and the transformer models were tested on 5 separate training and testing runs using different random seeds. The final results were tested for statistical significance using the t-test at the significance level of 5%. All models showed statistically significant improvement over the random baseline.

7.1. Model Performance

The best results were obtained by the transformer models for all personality traits. The LR and SVM models outperformed the fully-connected models on all personality traits except agreeableness, which is perhaps somewhat surprising at first. We conclude that this is the result of average embeddings being a suboptimal feature for a task such as this. The loss of valuable information such as word order and punctuation ultimately led to worse results in the fully-connected models. Exact results can be seen in Table 3.

7.2. Interpretability Using Coefficient Magnitudes

Table 1 shows the 10 unigrams for which the LR and SVM models had the largest coefficient values, while Table 2 shows the 10 unigrams with the smallest coefficient values.

Openness People who are high in openness tend to be more insightful, curious, and have a broad range of interests, while people who score low tend to be more traditional and struggle with abstract thinking. Words like "fun" and "music" were outlined as important in scoring high by both models, while words like "college", "class", "school" and "home", which could be connected to a more scheduled lifestyle without much time for new experiences, were outlined as important in scoring low.

Agreeableness Being high in agreeableness is often characterized by trust, kindness and affection. The inclusion of the word "family" in Table 1 for both models seems to follow this characterization. On the other hand, being low in agreeableness is characterized by competitiveness, manipulation and lack of caring for others. This often

[6] https://pytorch.org/

[7] https://pytorch.org/docs/stable/generated/torch.optim.Adam.html

[8] https://shap.readthedocs.io/en/latest/index.html

manifests by insulting and belittling others, which may be the reason that the words "stupid" and "damn" are considered important for scoring low in both models.

Conscientiousness Scoring high in conscientiousness implies good impulse control, thoughtfulness and organized and goal-oriented decision making. People who score low in conscientiousness are less structured and organized, and often lack impulse control. The words "today" and "tonight", which were deemed important for scoring high in both models, imply structure and orderliness, while words like "want" and "hate", which were deemed important for scoring low in both models, imply low impulse control. Surprisingly, the word "party" appears in Table 1 for both models.

Neuroticism People who score high in neuroticism often experience sadness, moodiness and emotional instability, while people who score low are more stable, emotionally resilient and relaxed. This is very well reflected in the words which were important in scoring high for both models, as most of them are emotionally charged.

Extraversion Being high in extraversion is characterized by sociability, talkativeness, excitability and emotional expressiveness. The words "love", "boyfriend" and "fun", which appear in Table 1 for both of the models, and the word "sorority", which appears only in the LR model, seem to coincide with this. The words related to scoring low in extraversion do not offer such a clear interpretation.

7.3. Interpretability Using SHAP Values

The second part of the interpretation was conducted using beeswarm plots[9] from SHAP values calculated on the test set, which were ordered by several relevance metrics. The SHAP values analysis was performed on the LR models.

Openness We found that high values of the features "home", "semester" and "college" resulted in small SHAP values, while high values of the features "music" and "love" resulted in large SHAP values.

Agreeableness High SHAP values for the agreeableness trait were obtained by high values of the features "family" and "love", while the lowest SHAP values were obtained by high values of feature "mean" and certain profanities. This is to be expected since low agreeableness is often manifested by using more swear words (Ireland and Mehl, 2014).

Conscientiousness High values of pronoun features, such as "he", "she", "her", "we" and "my", obtained very high SHAP values for this trait. This may be explained by the fact that people high in conscientiousness tend to pay attention to detail, so they avoid using unclear words like "someone" to describe people in their stories.

Neuroticism The highest SHAP values were obtained by high values of features "feel" and "stressed". The high value of the feature "blue" also resulted in high SHAP values, which is possibly explained by the use of the phrase "feeling blue" to indicate negative emotion.

Extraversion High values of the features "sleep", "home" and "typing" had the lowest SHAP values, which is expected since they are indicative of a lower desire for socialization. On the other hand, high values of the features "love" and "sorority" had the highest SHAP values.

[9]The mentioned plots, along with other code, can be found here: https://github.com/ilinardic22/TAR-project-2023

Table 3: Performance results for every personality trait expressed in percentages of F1-score. The best results were obtained by the transformer models.

Model	Ext	Agr	Opn	Con	Neu
Random	50.87	50.73	50.94	50.26	50.28
LR	57.27	53.54	61.68	55.88	56.45
SVM	56.95	53.53	62.17	56.28	57.01
FC	53.14	55.48	58.14	55.51	53.60
BERT	**67.21**	**67.31**	**69.16**	**67.18**	**64.36**

8. Conclusion

Personality trait classification based on text has been a challenging, yet vital problem both in the field of natural language processing, as well as modern psychology for quite a long time now. Text is a fickle medium for such a complex task, and the challenge of processing it properly in order to solve such a task has been tackled from various angles and through various means in existing work over the years. We present various methods and their results at attempting to solve this problem, with varying degrees of complexity and success. This paper's main focus, however, is highlighting the importance of interpretability. Being able to understand not only how people think, but how models think and delving deeper into the "why's" and "how's" behind the results was what we deemed to be the most valuable takeaway from this research.

Acknowledgements

We would like to thank prof. dr. sc. Jan Šnajder and mag. ing. Josip Jukić for the advice and help they gave us while writing this paper.

References

Yin-Wen Chang and Chih-Jen Lin. 2008. Feature ranking using linear svm. In Isabelle Guyon, Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander Statnikov, editors, *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, volume 3 of *Proceedings of Machine Learning Research*, pages 53–64, Hong Kong, 03–04 Jun. PMLR.

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

Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. PANDORA talks: Personality and demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online, June. Association for Computational Linguistics.

Molly E. Ireland and Matthias R. Mehl. 2014. 201Natural Language Use as a Marker of Personality. In *The Oxford Handbook of Language and Social Psychology*. Oxford University Press, 09.

O. P. John, E. M. Donahue, and R. L. Kentle. 1991. *The Big Five Inventory—versions 4a and 5*. University of California, Berkeley, Institute of Personality and Social Research.

Elma Kerz, Yu Qiao, Sourabh Zanwar, and Daniel Wiechmann. 2022. Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Robert R. McCrae and Paul T. Costa. 1985. Comparison of epi and psychoticism scales with measures of the five-factor model of personality. *Personality and Individual Differences*, 6(5):587–597.

Raghavendra Pappagari, Piotr Żelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification.

Gregory Park, H. Schwartz, Johannes Eichstaedt, Margaret Kern, Michal Kosinski, David Stillwell, Lyle Ungar, and Martin Seligman. 2014. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108, 11.

J. W. Pennebaker and L. A. King. 1999. Linguistic styles: Language use as an individual differnece. *Journal of Personality and Social Psychology*, 77:1296–1312.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Daniele Pizzolli and Carlo Strapparava. 2019. Personality traits recognition in literary texts. In *Proceedings of the Second Workshop on Storytelling*, pages 107–111, Florence, Italy, August. Association for Computational Linguistics.

HA Schwartz, JC Eichstaedt, ML Kern, L Dziurzynski, SM Ramones, M Agrawal, and et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE 8(9): e73791*, September.