NLP Shared Task

Tell Me Something About Yourself: Employing Machine Learning for Personality Classification

Arash Alborz
arash.alborz@student.uantwerpen.be

Magi Vinov
magi.vinov@student.uantwerpen.be

Iman Van de Velde
iman.vandevelde@student.uantwerpen.be

Victor Van Moer
victor.vanmoer@student.uantwerpen.be

All relevant information can be access through the GitHub and HuggingFace repositories.

1. Introduction

In recent times, human resources management deploys personality evaluation when considering candidates for a job. This is done because online personality tests are considered to be unbiased, efficient and diversity-oriented. One such known example of a personality test is the MBTI, however this is disputed as being unreliable and ungrounded (Stein and Swan 2019). Instead, the big five personality traits are used, which correspond to Openness (to experience), Conscientiousness, Extraversion, Agreeableness, and Neuroticism (as opposed to Mental stability), known as OCEAN (Goldberg 1990). Based on the assumption that mental properties are reflected in language use (Pennebaker and Niederhoffer 2003), each personality trait was labeled as 'low', 'medium' or 'high' - by analyzing a written text, the writer's OCEAN traits can be evaluated, and their personality can be matched to the requirement of a certain position or job. As a way to further the advancements in the field of personality tests, we developed our own OCEAN traits classification model. During the development of our system, we encountered papers describing previous attempts at performing a similar task, such as, (Philip and Devashrayee 2019), (Putra and Setiawan 2022), and (Maharani and Effendy 2022). These papers were taken as inspiration for creating our classification model, and certain aspects of their models were tested out. Specifically, we considered factoring in a count-based dictionary based on Linguistic Inquiry Word Count (Pennebaker and Booth 1999) to our embeddings, in order to improve the results (as used by (Putra and Setiawan 2022)), but eventually a different method was decided upon. This will be further expanded on in our methodology section. To give an extensive report of our model, we will first discuss the dataset used to develop our system, and how we adapted it further for our goal. Next, the model architecture and methodology are discussed. This is followed by our results, and we conclude with the limitations and conclusions of this project.

2. Data

The PANDORA dataset serves as a basis for the train data within this project. This dataset consists of Reddit comments, and each author's corresponding OCEAN traits scores (i.e. Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism). After employing statistical exploratory tools, we concluded that the data set is quite imbalanced. The lengths of comments created by each author varied significantly, ranging from as little as 7 words, to an astonishing 2,844,284 words per author (fig. 1). Additionally, some traits are imbalanced, with one label appearing more frequently than others in particular traits (e.g., 'high' for Openness and Agreeableness or 'low' for Conscientiousness) (see fig.2). This may be related to the nature of these traits.

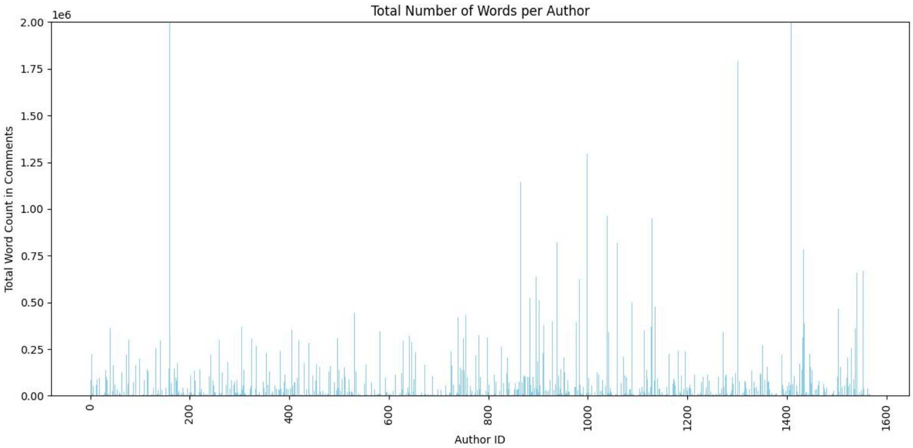

Figure 1
Words per Author in Train Data (limited to 2e6 tokens)

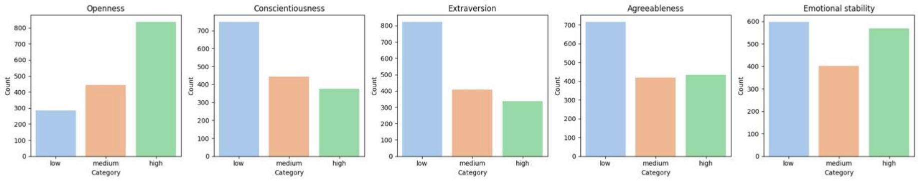

Figure 2
Classes in Train Data

To combat the imbalance, the training data underwent the following changes: (i) The Neuroticism label was changed to Emotional stability (the personality trait used in the validation data) and its scores were inverted (subtracted from 100), based on the fact that Neuroticism is the opposite from Emotional stability (Goldberg 1990). (ii) All numeric scores were converted to categorical classes based on the predetermined corresponding thresholds: scores below 33 were labeled 'low', scores between 33 and 66 were labeled 'medium', and scores above 66 were labeled 'high'. This way, the training data and validation data adhere to the same structure.

The validation data consists of responses to job interview questions that were filled out by our fellow students, together with the corresponding OCEAN traits scores each student acquired. However, due to the limited number of entries and the nature of the

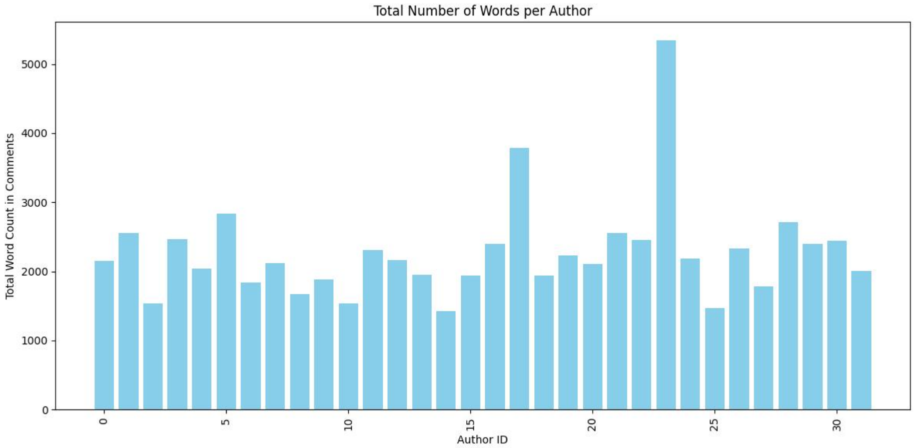

Figure 3
Words per Author in Validation Data

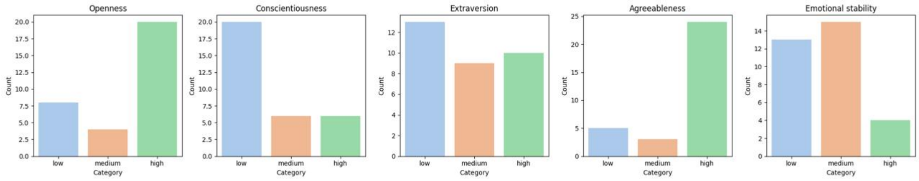

Figure 4
Classes in Validation Data

task, the validation data set is much smaller (fig. 3) and the labels are imbalanced and biased towards certain classes (fig. 4), creating a difficulty in analysing underrepresented classes. Additionally, because the validation data consists of job interview answers, which are usually presented in a more formal tone compared to Reddit comments, our train and validation data had a difference in register. Thus, the training data provided is different in style and disproportionate in length (compare fig.1 and fig. 3) to the validation data, discrepancies that may be the main cause of overfitting in the model.

To resolve the mismatch in genre and register, synthetic data was added to the train data. This synthetic data was generated based on the validation data, using the GPT model. This was done in two steps. (i) First, GPT4 was used to create synthetic answers based on the question and answer pairs of the 3 job interview questions, which are the following:

1. Please describe a situation where you were presented with a problem outside of your comfort zone and where you were able to come up with a creative solution.
2. Tell us about a time when you have failed or made a mistake. What happened? What did you learn from this experience?
3. Describe a situation in which you got a group of people to work together as a team. Did you encounter any issues? What was the end result?

Then, a few-shot prompting was employed, where two examples of the answers from the validation data were added to the prompt, as well as adjusted 'presence penalty' and 'frequency penalty' (0.6 and 0.4, respectively) to avoid generating similar outputs. (ii) Secondly, GPTo3 was used to generate the corresponding 5 OCEAN traits labels to the synthetic text. However, the generated scores were similar and consisted mostly of labels 'high' and 'medium'. This may be explained by the fact that the people answering job interviews try to answer as positively as possible, resulting in predominantly open, conscientious, extraverted, agreeable, and emotionally stable answers. It was decided to include these positive-oriented answers in our data, seeing that the majority consists of PANDORA train data, which will provide a quantity of 'low' labels for our model to be trained on. Finally, the PANDORA and generated text and labels were merged to be used as the final 'train data'.

3. Methodology

3.1 Text embeddings

The first step of developing our model was to embed the text fragments. For the initial try, we used tf-idf (as encouraged by (Philip and Devashrayee 2019)) to embed the comments from the PANDORA data set. However, this method failed to capture the importance of context (Pennebaker and Niederhoffer 2003) and semantic relations of words and the associated personality they denote. Moreover, the varying lengths of the train and validation sets proved that tf-idf frequency-based embeddings came short when compared to contextual embedding. Thus, we used the model 'distilbert/distilbert-base-cased-distilled-squad'. This BERT model processes up to 512 tokens at a time, so longer text was divided into multiple chunks with the purpose of including all Reddit comments. During the embedding step, two possibilities were explored for combatting the difference in register: mean pooling embeddings and the addition of LIWC features to the embeddings. Unlike [CLS]-based sentence embeddings, mean-pooled embeddings may be less sensitive to differences in register or style, due to the averaging of all embedding vectors (Doganca-Cetin 2023). However, the development of the model was eventually done with [CLS] token embeddings, seeing that synthetic data was introduced to our train data as a method to include both registers (i.e. formal and informal). After only using the text embeddings and encountering disappointing results, LIWC features were brought to our attention - as linguistic features that provide psychological information based on stylistic and topical choices (Putra and Setiawan 2022). Therefore, LIWC features were integrated into our embeddings, which resulted in an increase in performance. The combination of LIWC features and BERT-based [CLS] embeddings were used as the final data set the model was trained on.

3.2 Model training

During the development of the personality classification model, many different Machine Learning methods were tested out. The results for Naive Bayes, Random Forest, Support Vector Machines, and an ensemble of these machine learning methods (inspired by, among many, (Philip and Devashrayee 2019), (Putra and Setiawan 2022), and (Maharani and Effendy 2022)), as well as the performance of neural networks are reported below (Table 1). In hopes to improve accuracy and prediction scores, we thought of fine-tuning an already existing model. The 'BERTForSequenceClassification' model was chosen for that. This model embedded the text fragments, and fine-tuned them based on the results of the classification (per trait), but results were disappointing. This circled us back to using traditional Machine Learning approaches, namely - 5 different models, one for each trait, based on the method that provided the best results per trait. Looking at the average F1 and Accuracy scores of all tested methods, Random Forest performed the best. As the rest of the approaches led to unsatisfactory results, all personality traits were predicted using Random Forest. Since the difference in size of both data sets was substantial - Grid Search did not yield significant contributions. We opted to experiment with the parameters manually, changing each trait's 'numbers of estimators' and 'max depth' and fixing the parameters that give the best results. The final model and its evaluation are specified in the Results section (see Table 2).

Model	**Accuracy**	**F1 score**	**Weighted average**
Random Forest	0.43	0.33	0.37
Ensemble (RF+GB+MLP+SVM)	0.39	0.22	0.29
Ensemble (RF+GB+MLP)	0.35	0.20	0.26
Gradient Boosting	0.34	0.23	0.29
MLP Classifier	0.33	0.20	0.25
Logistic Regression	0.32	0.28	0.29
SVM	0.29	0.25	0.27

Table 1
Methods Employed per Personality Trait

4. Results

To adhere to an acceptable scientific and fair research framework, we chose a model that is not biased toward the majority class (which would result in a high accuracy score), but instead one that can effectively predict all classes. This decision was especially important given the imbalanced nature of our dataset, where relying solely on accuracy would obscure poor performance on underrepresented classes. We prioritized metrics that better reflect a model's performance across all categories, such as accuracy and macro-averaged F1 score. This approach ensures that minority classes are not neglected and that the model's predictions are more equitable and reliable. The evaluation scores, which reflect this balanced perspective, are shown in Table 2.

Evaluation Metric	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Emotional Stability**
Accuracy	0.62	0.62	0.47	0.38	0.53
F1-score	0.47	0.48	0.44	0.36	0.47

Table 2
Final Results per Personality Trait

These results highlight the model's ability to generalize across all classes, rather than overfitting to dominant patterns in the data (e.g. simply predicting the majority class all the time), supporting the validity of our chosen approach.

5. Conclusions and limitations

This project explored the application of machine learning for personality classification using the Big Five (OCEAN) model, combining Reddit-based data, synthetic interview-style data, and formal validation entries. While we aimed to build a model that prioritizes fairness and avoids bias toward majority classes, we encountered notable challenges - most prominently, the genre and size mismatch between our training and validation sets. To mitigate this and bridge the stylistic gap, we generated synthetic data using GPT. Although this method improved performance on validation data, it introduced limitations that affect the ability to generalize and reproducibility. The synthetic data likely introduced a third, distinct register (GPT-generated language), which risks the model overfitting to artificial stylistic cues rather than genuinely learning personality traits (Alexander and Oswald 2020). Additionally, GPT-generated labels differ from those derived via self-assessment questionnaires, as used in both the PANDORA dataset and our validation data, neglecting important individual insights such as self-perception (Ding and Bing. 2023). The reliance on synthetic data also complicates reproducibility, even with controlled parameters such as 'temperature' and 'random seed'. Despite these challenges, our combination of BERT-based embeddings, LIWC features (Pennebaker and Booth 1999), and Random Forest classifiers yielded promising results without relying solely on high accuracy scores that often reflect bias toward dominant classes. Future research would benefit from access to larger, context-consistent datasets with verified labels - ideally annotated by the text authors themselves - even if that data is harder to collect and subject to personal biases. Using synthetic data is acceptable in this academic setting, but alternative methods may be preferred in professional or business environments where interpretability and accountability are paramount.

References

Alexander, Evan Mulfinger, Leo and Frederick L. Oswald. 2020. Using big data and machine learning in personality measurement: Opportunities and challenges. *European Journal of Personality*, 34(5):632–648.

Ding, Chengwei Qin Linlin Liu Yew Ken Chia Boyang Li Shafiq Joty, Bosheng and Lidong Bing. 2023. Is gpt-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics 1*.

Doganca-Cetin, Serenay. 2023. *Exploring the effectiveness of BERT using different pooling strategies on SVM for news classification*. Ph.D. thesis, department of cognitive science artificial intelligence, Tilburg university, Tilburg.

Goldberg, Lewis R. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Maharani, Warih and Veronikha Effendy. 2022. Big five personality prediction based in indonesian tweets using machine learning methods. *International Journal of Electrical and Computer Engineering (IJECE)*, 12(2):1973–1981.

Pennebaker, Martha Francis, James W. and Roger Booth. 1999. *Linguistic inquiry and word count (LIWC)*. Erlbaum Publishers, Mahwah, NJ.

Pennebaker, Matthias R. Mehl, James W. and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54:547–577.

Philip, Dhvani Shah Shashank Nayak Saumik Patel, Joel and Yagnesh Devashrayee. 2019. Machine learning for personality analysis based on big five model. *Advances in Intelligent Systems and Computing*, 839:345–355.

Putra, Rahadian Perwita and Erwin Budi Setiawan. 2022. Roberta as semantic approach for big five personality prediction using artificial neural network on twitter. In *International Conference on Advanced Creative Networks and Intelligent Systems (ICACNIS)*, pages 1–6.

Stein, Randy and Alexander B. Swan. 2019. Evaluating the validity of myers-briggs type indicator theory: A teaching tool and window into intuitive psychology. *Social and Personality Psychology Compass*, 13(2).