Language Models Can Be Used to Approximate the Perceived Personality of Famous People

Stanford CS224N Custom Project

Xubo Cao
Graduate School of Business
Stanford University
xcao@stanford.edu

Abstract

The perceived personality of famous people (e.g., artists, politicians, business leaders) have drawn great interest of social science researcher. However, obtaining this type of personality data can be costly and time consuming. I hypothesize that the perceived personality of famous people are well encoded in the language that describes them, and therefore language models pre-trained on large corpus likely contain information about people's perception about certain famous individuals. In this project, I aim to 1) explore whether embedding features obtained from pre-trained language models can be used to predict perceived personality of famous people 2) compare the performance of models based on different language models (e.g., Word2Vec, RoBERTa, GPT-3). I found that embedding features obtained from these pre-trained language models predict perceived personality of famous people with moderate to high accuracy, and GPT-3 considerably outperformed the other two models in this task.

1 Key Information to include

- Mentor: Ethan Chi
- External Collaborators (if you have any): Michal Kosinski (My PhD advisor. He provided feedback on this research project outside the class, although he did not participate in this class project.
- Sharing project:

2 Introduction

Research over the past few decades has shown that the myriad of potential personality and value constructs can be reliably captured by five essential personality traits, the so-called Big Five or the Five-Factor Model (FFM) [1]. The five underlying dimensions include (a) Extraversion, (b) Agreeableness, (c) Conscientiousness, (d) Neuroticism, and (e) Openness to Experience. These personality traits can be used to efficiently describe individuals' behavioral patterns and reliably predict a wide range of individual and social outcomes [2]. Therefore, the perceived personality of famous people (e.g., artists, politicians, business leaders) have drawn great interest of social science researcher in various disciplines. For example, organizational researchers have found that the personalities of top executives influence organizational culture and organizational effectiveness (e.g., [3]); political scientists have found that the perceived personality of politicians influences voters' preferences, legislator approvals, and congressional behaviors (e.g., [4]); marketing researchers have found that the perceived personalities of the celebrity endorsers of a brand can influence brand recognition, brand attitudes, and purchase intention (e.g., [5]); even the public personality of artists has been found to influence the preferences of music listeners (e.g., [6]). Celebrity personality data are often collected by surveying the general public or a certain group of qualified informants. The

traditional survey methods, however, can be costly and time-consuming. In this paper, I propose an alternative, complementary approach to assessing celebrity personality. Specifically, I test whether semantic embeddings obtained from pre-trained language models can be used to approximate the perceived personality of celebrities.

I hypothesize that the perceived personality of famous people are well encoded in the language that describes them. Therefore, language models pre-trained on large corpus likely contain information about people's perception about certain famous individuals. In this project, I aim to 1) explore whether embedding features obtained from pre-trained language models can be used to predict perceived personality of famous people 2) compare the performance of models based on different language models (e.g., Word2Vec, RoBERTa, GPT-3).

3 Related Work

Although NLP techniques, especially word embeddings, have been increasingly used in social science research, most of the studies focus on using cosine similarity between group concepts (e.g., gender words, occupation words) to reveal macro-level patterns of human judgment and perception. For example, Garg and colleagues have used the cosine similarity between gender words and occupation words to demonstrate the historical change of stereotypes [7]. Similarly, Lewis and Lupyan showed that female words were closer to family-related words while male words were closer to career-related words in the semantic space, reflecting people's implicit gender associations.

However, there has been limited studies on using language models on a micro-level, like understanding people's perception about a certain individual. Most relevant to the my project is Bhatia and colleagues' line of research. In their 2019 work, Richie, Bhatia, and colleagues showed that word embeddings can be used to predict people's perception of famous figures' warmth and competence, which are two universal dimensions of the perception of human beings [8]. Importantly, they found that a regression-based approach is more effective than cosine similarity to predict people's ratings of a certain target [9]. The authors fitted regularized regression models with word vectors of human names (e.g., 'Bill Clinton') as features and the trait of interest (e.g., "warmth") as the criterion variable. The model was able to learn the weights that map word vectors onto the trait and make highly accurate predictions. Extending this line of work, Bhatia and colleagues utilized the same method to predict leadership perception with fairly high accuracy, such that the Pearson correlation between predicted ratings and human-ratings reached .78 [10].

Bhatia and colleagues' work indicates that word embeddings contain rich information of human perception. The approach can be potentially extended to measure other psychological traits. In my project, I test whether this approach can be used to predict the personality perception.

One important limitation of Bhatia and colleagues' work is that it is solely based on Google's Word2vec model, because it is one of the few, if not the only pre-trained static embedding models that include word vectors of human names. Thus, the authors' approach is constrained by the vocabulary of Word2Vec and cannot be updated. In addition, Word2vec is not the best-performing word embedding models, plus that static embeddings models are not the state-of-art in NLP. It is possible that the authors' method can be improved by using more updated models. This limitation motivates my attempt to use RoBERTa and GPT-3 to improve their approach.

4 Approach

My task is to create a model that, given a famous person's name, can return the predicted personality ratings of this individual. I fitted regression models with embedding features as input. The approach has been tested in previous research [10] and is also one of the recommended way to use semantic embeddings according to GPT-3's user guide. Specifically, the approach involves two steps: (1) obtaining semantic embeddings that represent the target from a language model; (2) feed embeddings into a regularized regression model to learn the weights that map embedding features to personality traits.

4.1 Obtaining Embeddings

I used three pre-trained language models to generate semantic embeddings: Word2Vec, RoBERTa, and GPT3. The performance of the three different models will be compared. One important assumption is that the embedding of a certain person's name include rich information about people's perception of the person. Therefore, it is possible to use a person's "name vector" to predict their perceived personality. However, because the three language models have different structure or API, different approaches are need to obtaining the embedding that represents a certain person. I briefly describe the methods in the following section:

First, I utilized the pre-trained Word2Vec model to generate word embeddings. Because Word2Vec is a static word embedding model that has a key-value structure, one can directly retrieve a vector that represents an individual using the person's name, which takes the form of multi-word phrases separated by an underscore (e.g., "Barack_Obama"). Importantly, only a limited group of people are represented as vectors in Word2Vec, who are likely the most famous people on the world. Bhatia and colleagues searched the Word2Vec vocabulary and identified 6627 famous people who were present in the embedding vocabulary[1]. I referred to this dataset and used the Python module "gensim" to obtain word embeddings from Word2Vec.

Next, I utilize RoBERTa to generate semantic embeddings. RoBERTa takes free texts as input and turn them into numerical values. Therefore, the RoBERTa-based approach is not limited by a specific vocabulary (although a person's name may not be recognized and represented by an "unknown token", this does not seem to be a problem in my experiments). Two decisions are to be made when using RoBERTa to generate embeddings: (1) the prompt to use and (2) how to aggregate the information in the hidden layers. In my initial experiments, I simply use the target's name as the input, and experiment with two different ways to aggregate the output. (1) the CLS approach, which simply extracts the embedding of the [CLS] token in the final hidden layer and uses it as the embedding features for regressions; (2) the mean approach, which takes the average of all embeddings in the last hidden layer and uses it as the embedding features for regressions. I found that neither of these methods yielded satisfying results (see results section for more details). I then experimented with different prompts. Considering the large number of possible combinations of prompts and aggregating methods, I only present the most intuitive and successful attempt here, the so-called "personality embedding" approach[2]. Because RoBERTa is a contextual embedding model instead of a static embedding model like Word2Vec, its embedding representation is sensitive to the context of the input sequence. In order to "inform" the model of its task, I included the word "personality" following the target's name (e.g., "Barack Obama's personality"), and extract the embedding of the word "personality" and used it as the input. This "personality embedding" approach performed significantly better than

Finally, I utilized GPT-3 to generate semantic embeddings. Although GPT-3 is not public accessible currently, its developer, OpenAI, provides an API that researchers and developers can apply to use. The API includes a Python module "openai" can be used to easily turn a sequence into a semantic embedding using GPT-3's models, like one can do with the 'transformer' module. In the case of GPT-3, however, no much experimentation is possible because the module does not return all the hidden layers like the "transformer" module does. Therefore, I simply use the target's name as the input to obtain the embedding that represents the target.

4.2 Weights Learning

After obtaining vectors that represent the targets, I trained Ridge regression models to learn the weights that map high-dimensional embedding features onto personality traits. Ridge regression models are similar to linear regression models (or a linear layer in neural networs) except that an L2 penalty is added to the coefficients. The loss function of a Ridge regression model is as follows:

$$L_{ridge}(\hat{\beta}) = ||Y - \hat{\beta} * X||^2 + \lambda||\hat{\beta}||^2 \quad (1)$$

Here, the first absolute value term is the loss function of the traditional OLS regressions, the sum of squared error. The second term is an L2 penalty term that constraints the magnitude of coefficients.

[1] https://osf.io/52w7r/

[2] Similarly, I also experimented with different layers of RoBERTa output, but it seems that the embeddings in the last layer always outperform other layers. Therefore, I do not report the relevant results in this paper

The penalty helps address the multicollinearity between predictors and reduce overfitting by rewarding a more parsimonious model. λ is a parameter that regularizes the magnitude of the penalty, which is commonly referred to α in Statistics and Psychometrics. While it is possible to experiment with different value of α to optimize the model's performance or to use cross-validation to find an optimal alpha for the training set, it is not the focus of this paper (Similarly, one can experiment with L1 penalty instead of L2 penalty). Here, for the sake of simplicity and comparability between studies, I use the alpha value of 1, which was suggested by [10] and proved to work well in his leadership perception study.

5 Experiments

5.1 Data

I have collected a sample of perceived personality of from Prolific. 600 participants were recruited to rate 10 famous individuals on their Big-Five personality traits on a 7-point Likert scale using the Ten-Item-Personality Inventory (TIPI) [11]. The 10 targets presented to each participant were randomly selected from a pool of 300 famous individuals, whose Wikipedia pages received the most views [12]. To guarantee the validity of the ratings, participants were allowed to skip rating a person they were not familiar with them. As a result, each target received different number of ratings from participants. To guarantee the reliability of the ratings, I only used the 218 targets that received ratings from more than 10 different participants as my analytical sample. The means of these ratings were calculated and used as the criterion variables. All of the 218 individuals can be found in the dataset published by [10] and are therefore present in the Word2Vec vocabulary. This is not a coincidence because we generate our list of targets using a data source as [10], the Pantheon dataset [12]. Also, these 218 individuals received most ratings because they are most famous, which is also the reason why were included in the Word2Vec vocabulary. Thanks to this, all of our models are compared using the same sample.

5.2 Evaluation method

To recap, my analytical sample includes a list of 218 famous individuals. The names of these people are the input in my task. I collected people's ratings on these targets' Big-Five personality, which constitute the labels in my training and testing data. Because of the limited sample size and the exploratory nature of this research, I did not explicitly divide the data into training, validation, and test sets. Instead, I applied leave-one-out cross validation to generate out-of-sample predictions. In other words, after I obtained 218 vectors that represent the 218 targets, I use 217 of them to train a regression model and used the model to predict the personality traits of the remaining person. The process was repeated 218 times, resulting in out-of-sample predictions for the whole analytical sample. I then calculated the Pearson correlation between the predicted personality score and the average human ratings as the metric for evaluation. Correlation coefficients are commonly used as the metric for regression tasks, especially in psychometrics.

5.3 Experimental details

In terms of model settings. For Word2vec, I used the 300-dimension version pretrained model. For RoBERTa, I used the "roberta-base" provided by Huggingface. For GPT-3, I use the "davinci-similarity" engine to generate embeddings. No additional training is performed for any of the model. In terms of Ridge regression, I used an alpha equal to 1, which is recommended in [10].

5.4 Results

Table 1 shows the performance of the different models. Although there is no clear baseline for our analysis, I expected that (1) the results with Word2Vec model would be comparable to the 0.78 in Bhatia and colleagues' study on leadership perception and (2) the results with RoBERTa and GPT-3 would be better than the results with Word2Vec. Not all of the expectations were met.

First, none of the predictive correlations in my results exceed the 0.78, and many of them are much lower than the threshold. The difference may simply reflect the difference between personality perception and leadership perception. Leadership perception may be more predictable because it is a

Model	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
Word2Vec	0.63	0.45	0.58	0.55	0.67
RoBERTa (CLS)	0.42	0.22	0.46	0.21	0.60
RoBERTa (Mean)	0.49	0.39	0.35	0.37	0.61
RoBERTa (Personality Embedding)	0.63	0.48	0.49	0.51	**0.74**
GPT-3	**0.72**	**0.62**	**0.62**	**0.68**	0.72

Table 1: Summary of model performance on with leave-one-out-cross-validation.

one-dimension trait and is largely influenced by the targets' occupation. It is easy to predict that an US president is generally perceived as more leader-like than a pop singer. In contrast, personality is multidimensional and more subtle. In fact, even within the Big-Five personality, some traits are more predictable than others. Regardless of the model used, agreeableness and openness were considerably more predictable than conscientiousness, extraversion, and neuroticism. In addition to the different between traits of interest, difference in data quality may also contribute to the worse results in my study.

Second, compared to the baseline Word2vec model, using GPT-3 embeddings yielded significant better results. However, RoBERTa performed worse than even Word2Vec when the target's name was used as the prompt. The personality embedding approach largely fixed problem, and significantly improved the performance of RoBERTa-based models. This finding demonstrates the importance of prompt design when using RoBERTa embeddings in trait predictions. As has been discussed earlier, because RoBERTa is a context-sensitive model, including the word "personality" helps direct the model's attention. Interestingly, it is the embedding of the word "personality" that can be used to predict personality traits best instead of the embedding of the target's name like in Word2Vec and GPT-3. However, even the personality embedding approach of RoBERTa was only able to perform comparably but not much better than Word2Vec. It is possible that there are better prompt design and aggregation method that can further improve RoBERTa's performance in this personality prediction task, but my data indicates that RoBERTa may not contain much richer information about individual personality than Word2Vec. On the contrary, GPT-3 considerably outperformed both models in this task. This may be attributed to the higher dimension of GPT-3 embeddings (12240 compared to 768 in RoBERTa and 300 in Word2Vec). The difference may also be explained by the more complex structure and larger training sample of GPT-3.

6 Analysis

In order to further investigate the weights learnt by the models and examine whether they actually captures personality and are plagued by any stereotypes, I run the so-called "pseudo-rating" analysis developed in [10]. Instead of feeding word vectors of human names into the regression model to generate leadership ratings, one can also feed the model with word vectors of human traits (e.g., "innovative") to generate pseudo-ratings of these traits. These pseudo ratings represent the importance of the corresponding trait in leadership perception.

Here, I collected 525 person-descriptive adjectives that have been used in personality studies [13] and performed a pseudo rating analyses on these adjectives with the best performing GPT-3-based model. To summarize the results, Table 2 shows the top 5 rating and the bottom 5 rating adjectives for each of the five personality traits.

It can be observed that the ranking of the adjectives are highly intuitive. For example, the word "graceful" received highest rating on the dimension of agreeableness and the word "abusive" received the lowest rating. This confirms that the regression models were indeed able to capture people's perception of an individual on their different traits. However, a closer inspection also shows some biases or heuristics. For example, the word "democratic" was ranked second lowest in openness to experience. Admittedly, political ideology is a covariate of individual personality, the association seems to be excessively amplified by the model (or the human raters). Also, for extraversion, one of the top rating words is "obnoxious" and one of the bottom rating words is intellectual, reflecting a false association between extraverted individuals with obnoxiousness and low intelligence.

Personality Trait	Top5	Bottom5
Agreeableness	Graceful, Gracious, Warm-hearted, Kind-hearted, Lovely	Evil, Corrupt, Hostile, Dishonest, Abusive
Concientiousness	Athletic, Admirable, Gracious, Inspirational, Intelligent	Irresponsible, Disgusting, Incompetent, Disorganized, Awful
Extraversion	Entertaining, Hilarious, Cocky, Obnoxious, Glamorous	Depressed, Withdrawn, Quiet, Intellectual, Thoughtful
Neuroticism	Corrupt, Terrible, Evil, Awful, Disgusting	Gracious, Admirable, Appreciative, Inspirational, Athletic
Openness to experience	Artistic, Creative, Romantic, Imaginative, Talented	Conservative, Democratic, Narrow-minded, Corrupt, Close-minded

Table 2: Pseudo ratings of 525 person-descriptive adjectives

Because our model uses human ratings as the training data, which are inherently highly subjective and biased, it is hard to argue that the model is "making an mistake". Instead, it more likely reflects people's lay definition of personality traits. However, when researchers attempt use human ratings or the approximate ratings generated with language models, it is important to recognize these inherent biases. From this persepctive, the analysis of these pseudo ratings provide a window for researchers to understand these stereotypes in a more nuanced way.

7 Conclusion

Summarize the main findings of your project, and what you have learnt. Highlight your achievements, and note the primary limitations of your work. If you like, you can describe avenues for future work.

Overall, my results indicate that it is possible to extract information about individual personality from language models. This can serve as a tool for social science researchers to collect some convenient personality perception data. The pseudo rating analysis can also provide insights into lay definition of personality and potentially used to generate a more nuanced "stereotype profile" for a social group (e.g., how are males and females generally perceived on these five dimensions.)

In terms of model comparison, I found that high dimensional embedding features like GPT-3 embeddings likely contain richer information about human personality. In contrast, RoBERTa embeddings and Word2Vec embeddings performed almost equally well. However, a critical lesson is that the performance of a RoBERTa-based model is largely dependent on prompt design and output aggregation method. Here, I found that the "personality embedding" approach significantly outperformed other RoBERTa approaches in this trait prediction task. Future study should continue to explore whether there are better way to utilize RoBERTa in this type of tasks.

References

[1] Robert R. McCrae and Paul T. Costa. Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1):81–90, 1987.

[2] Daniel J. Ozer and Verónica Benet-Martínez. Personality and the Prediction of Consequential Outcomes. *Annual Review of Psychology*, 57(1):401–421, 2006.

[3] Charles A. O'Reilly, David F. Caldwell, Jennifer A. Chatman, and Bernadette Doerr. The promise and problems of organizational culture: CEO personality, culture, and firm performance. *Group and Organization Management*, 39(6):595–625, 2014.

[4] Jonathan D Klingler, Gary E Hollibaugh, and Adam J Ramey. What I Like About You : Legislator Personality and Legislator Approval. *Political Behavior*, 41(2):499–525, 2019.

[5] Debasis Pradhan, Israel Duraipandian, and Dhruv Sethi. Celebrity endorsement: How celebrity–brand–user personality congruence affects brand attitude and purchase intention. *Journal of Marketing Communications*, 22(5):456–473, 2016.

[6] David M. Greenberg, Sandra C. Matz, H. Andrew Schwartz, and Kai R. Fricke. The self-congruity effect of music. *Journal of Personality and Social Psychology*, 121(1):137–150, 2021.

[7] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 115(16):E3635–E3644, 2018.

[8] Amy J.C. Cuddy, Susan T. Fiske, and Peter Glick. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map, 2008.

[9] Russell Richie, Wanling Zou, Sudeep Bhatia, Simine Vazire, and Simine Vazire. Predicting High-Level Human Judgment Across Diverse Behavioral Domains. *Collabra: Psychology*, 5(1):1–12, 2019.

[10] Sudeep Bhatia, Christopher Y. Olivola, Nazlı Bhatia, and Amnah Ameen. Predicting leadership perception with large-scale natural language data. *Leadership Quarterly*, (November 2019), 2021.

[11] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003.

[12] Amy Zhao Yu, Shahar Ronen, Kevin Hu, Tiffany Lu, and César A. Hidalgo. Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific Data*, 2016.

[13] Gerard Saucier. Effects of variable selection on the factor structure of person descriptors. *Journal of Personality and Social Psychology*, 73(6):1296–1312, 1997.