UNIVERSITY OF MASSACHUSETTS AMHERST
COMPUTER SCIENCE

December 15, 2023

CS 485 Final Project: Using Convolutional Neural Networks to Predict Personality

by

Ryan Zaid Nicholas Santos

Abstract:

The main motivation of this paper is to measure the efficacy of using various convolutional neural networks in predicting the Big 5 personality traits. To achieve this, an annotated data set of essays displaying the Big 5 personality traits were used to create feature vectors. There is 4 main models that were tested, with them being a bag of words model, a model with GloVe pre-embedded word vectors, a model with LIWC features, and a model with both GloVe and LIWC vectors. The bag of words model performed the worst, predicting a positive class almost unilaterally. The LIWC model performed slightly better, maintaining an average accuracy of .52. Our GloVe model performed the best, reaching .64 accuracy in predicting openness, and maintained above a .5 accuracy for every personality trait, something not done in any other model tested. Our attempt at combining the LIWC and GloVe model did not results in improvements to the GloVe model, though improvement was seen in comparison to the LIWC model.

1 Introduction

Personality is a very strange area of research within the psychological field. Many researchers have wildly differing definitions of what it is, what it entails, and how to measure it. Some choose to study personality through a biological lens, studying the neural substructures behind what makes us who we are. However, the bulk of current psychological research takes a primarily statistical, linguistic approach to the study of personality. When it comes down to it, personality is just a linguistic category, a set of associations about what we as a people think a person consists of. So, to determine what a person consists of, we can just administer millions of questions about the self to millions of different people. The resulting factor analysis of these questions is known as The Big 5 personality measure, a cross cultural standard of metrics that define the bulk of who you are. While other measures exist, this is by far the most statistically valid of them.

The Big 5 defines personality across 5 unique dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Openness is ones ability to experience new things, and is also highly correlated with intellect, or ones tendency to think abstractly. Conscientiousness is how orderly and hard working someone is. Extraversion is how outgoing a person is, and is highly correlated with feelings of positive emotion. Agreeableness is how polite and compassionate a person is. And finally, neuroticism is one's susceptibility to negative emotion.

By training a convolutional neural network to predict these five personality traits, the goal is to be able to quickly gauge someone's personality using a source text that they create. This means understanding how something as specific and nuanced as personality can be reduced in order to be easily identified by a model. Therefore, the creation of our feature vectors are incredibly important, as they need to encapsulate more than just the words themselves. Hence, a secondary goal is to understand which method of creating feature vectors encodes the most relevant information. These goals are much more complicated than a simple classification problem due to the five dimensions of output allowed by the Big 5 personalities.

In relation to the goal of quickly assessing someone's personality, one object of experimenting with these models is to beat the performance of rudimentary classification methods with the use of more complicated ones.

2 Related Work

Over the last 20 years, there has been a plethora of research pertaining to training models to classify personality. One of the most recent papers come from Edward P. Tighe titled " Personality Trait Classification of Essays with the Application of Feature Reduction".[3] Similar to our methods, Tighe used LIWC for feature creation. However, Tighe lacked the use of a Convolutional Neural Network as one of the algorithm the paper tested. Tighe saw mediocre results in the metrics of the models. However, much was learned about how feature reduction techniques can be used to increase classification performance.

There has also been other classification attempts with other personality metrics than the Big 5.[1] However, it was done using other techniques than the ones being used in our experiments.

3 Data

The data set that is being used to train the convolutional neural networks is a set of 2400 *stream of consciousness* essay passages written by 34 students over 4 years. These essays were then annotated to flag a binary value if a certain one of the Big 5 personalities were exhibited in the given text. This removed the need to do human annotation. This data is being provided in a CSV format.

In addition to the annotations done on the original data set, in order to train the LIWC convolutional neural network, additional annotations were added. These included the 69 different features that are identified by the LIWC model. This was done to hopefully capture more psychological meaning behind the texts than just the binary annotations first given.

To test for over fitting to the specific essays themselves, we created 2 of our own data sets using chatGPT. These data sets consist of 25 sentences each, along with personality indicators and LIWC scores that match with their respective sentences.

For one of the data sets, 5 sentences were generated for each personality trait designed to be emblematic of that personality trait. Personality scores for that personality trait would be marked 1, and all other traits would be marked 0 for

ease of analysis.

For the other, 5 sentences were generated for each personality trait designed to specifically not show that personality trait. Personality scores for that personality trait would be marked 0, and all other traits would be marked 1 for ease of analysis. Using these data sets we will be able to test our best model's ability to predict personality positively and negatively on external data.

3.0.1 Data Set Analysis

Various methods of analysis were done on the data set in order to learn more insights into motivations behind the data set and what to expect from the output of our models.

First, a heat map (Figure 1) was created to to show correlation between a positive value in one of the Big 5 personalities and the other four personalities. An interesting takeaway, and validation point for the annotations, is that there is a negative correlation between Neuroticism and the other 4 personalities. This is typically seen in relation to extraversion and agreeableness, as they are somewhat opposing personality traits.

We also performed an analysis of word counts across essays marked with each personality trait. Typically, people high in openness tend to use larger words than others. We did not see this in our data set, as all distributions were roughly equal.

These two explorations into our data set are suggestive of fairly poor validity. Features usually seen in the general population are not seen here. This is likely due to the binary classification of personality present in this data set. Personality is typically defined across a 0-100 scale, not a 0 or a 1, leaving nuance out entirely. Unfortunately no other data set could be obtained, as more psychological research needs to be conducted for properly valid data. (Figure 2)

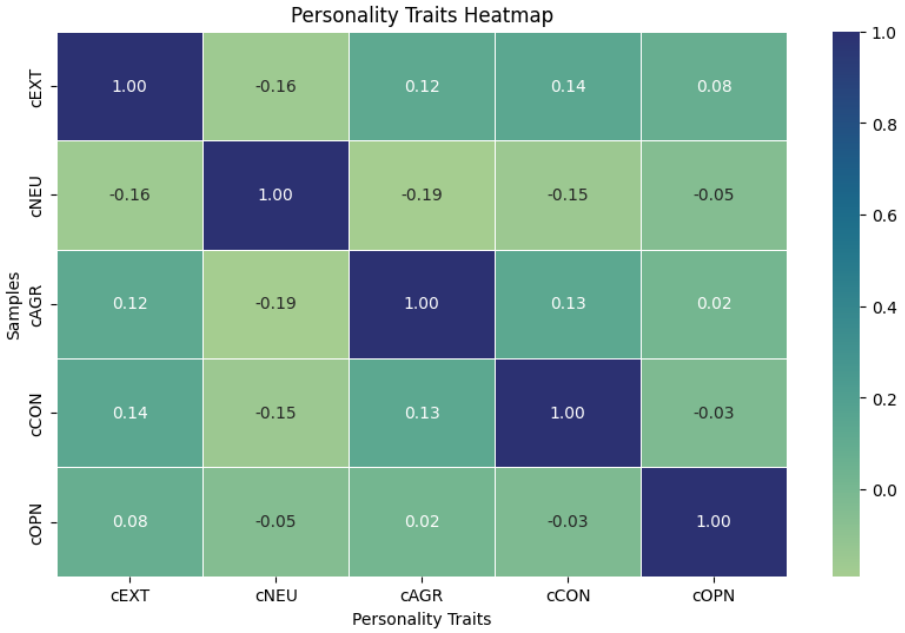

Figure 1: A Heat Map of the Big 5 Personalities

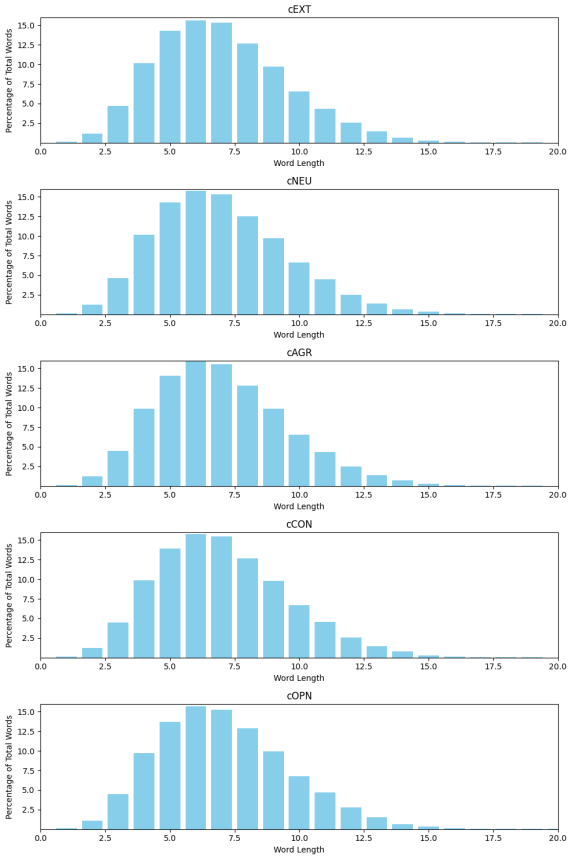

Figure 2: Percentage of Word Lengths respective of their flagged personality

4 Method

4.1 Convolutional Neural Network

We will be using a Convolutional Neural Network as our predictive model. This is because Convolutional Neural Networks excel at natural language processing sentiment classification tasks. For example, Microsoft is currently using convoloutional neural networks along with sentiment analysis for information retrieval. [2]. Therefore, using a convolutional neural network will allow for easy inclusion of semantic meaning, as this is a common practice in the use of convolutional neural networks. [4]

4.2 Training

With the previously mentioned data set, the model and analysis used comprised of a 75/25 training and test split. Therefore, 75% will be used for training each model and 25% will be used to test the models.

4.3 Feature Vectors

To learn more about how well convolutional neural networks will preform at classifying personality, four differently trained models will be examined for their capabilities. The differences in these models will come from how we create the feature vectors that will be used to train the convolutional neural network.

4.3.1 Bag of Words

The first, and most simple, method of creating features is to use a Bag of Words method. This is simply a count of the words that appear in the essays. The count is over the total number of unique words in the whole corpus. Along with the annotations that are already provided in the data set, it correlates what words are used and whether the annotator has flagged that text as potraying a certian personality feature.

4.3.2 GloVe

GloVe "is an unsupervised learning algorithm for obtaining vector representations for words"[1]. Using this GloVe model, there is pre-trained word vectors that are available for use. These are massive vectors that encapsulate semantic meaning for each word. Many computations can be done with these word vectors and they have unique properties such as allowing to mathematically calculate how semantically close two words are. Using these vectors, we create an embedding matrix of each word in the corpus. Then, those matrices are used as the input to the convolutional neural network. Once again, the annotations provided for the data set allow us to correlate the binary value of each personality to the embedding matrices.

4.3.3 LIWC

LIWC is a text analyzation tool that over 100 dictionaries to create features that capture specific emotional and psychological meaning. Specifically, it iterates over the words in the corpus and see if they correspond with any of the words inside the dictionaries. Then, using these frequencies, creates around 70 variables corresponding to various emotional and psychological meanings. These frequencies are what comprises our feature vectors. The usage of LIWC is an attempt to encode more emotional information into the input of our convolutional neural networks.[2]

4.3.4 GloVe and LIWC

These feature vectors constitute both the word embeddings from GloVe and the variables that are output by LIWC. Doing this is an attempt to encode the maximum amount of information into the input vectors. It is done with the hope of combining the semantic analysis given by GloVe and the emotional information provided by LIWC.

[1] https://nlp.stanford.edu/projects/glove/
[2] https://www.liwc.app/

4.4 Hyper Parameter Tuning

After all of the feature vectors are done being created, we train each model with various hyper-parameters to ensure that the model is working to it's full efficacy. To do this, we randomize a value for the *Random State* parameter and fix the *Epoch* parameters in intervals of five. Then, we test the accuracy of each model with these hyper-parameters. When the values of each hyper=paremeter are found that maximize the accuracy, we then use it in the final result metrics.

5 Results

Our baseline method was the bag of words model. One of the main objectives was to outpreform our baseline method.

5.1 Metrics

To calculate the results, various metrics were obtained from each model. From each model, values for accuracy, prevision, recall, and F1 scores were calculated for each of the personalities.

——	Accuracy	Precision	Recall	F1
cEXT	0.5405	0.5405	1	0.7017
cNEU	0.4737	0.4737	1	0.6429
cAGR	0.5547	0.5547	1	0.7135
cCON	0.5304	0.5430	0.8277	0.6558
cOPN	0.5081	0.5146	0.9572	0.6694
Averages	0.52148	0.5253	0.95698	0.67666

Table 1: Bag of Words Metrics

——	Accuracy	Precision	Recall	F1
cEXT	0.5202	0.5284	0.6914	0.5990
cNEU	0.5223	0.5182	0.5772	0.5462
cAGR	0.5668	0.5796	0.7228	0.6433
cCON	0.5223	0.5308	0.4498	0.4870
cOPN	0.6478	0.6681	0.6040	0.6345
Averages	0.55588	0.56502	0.61904	0.582

Table 2: GloVe Metrics

——	Accuracy	Precision	Recall	F1
cEXT	0.5243	0.5213	0.6411	0.5750
cNEU	0.5000	0.4909	0.6722	0.5674
cAGR	0.5526	0.5931	0.6255	0.6088
cCON	0.5324	0.5381	0.4309	0.4781
cOPN	0.5223	0.4896	0.5108	0.5000
Averages	0.52652	0.5248	0.5761	0.54586

Table 3: LIWC Metrics

——	Accuracy	Precision	Recall	F1
cEXT	0.4939	0.5097	0.6172	0.5583
cNEU	0.5385	0.5536	0.3780	0.4493
cAGR	0.5202	0.5688	0.4644	0.5113
cCON	0.5162	0.5191	0.5462	0.5323
cOPN	0.6093	0.6863	0.4200	0.5211
Averages	0.53562	0.5675	0.48516	0.51446

Table 4: GloVe and LIWC Metrics

5.2 Generalization Testing

Using the essays that were generated using ChatGPT, metrics were generated for the GloVe and LIWC models to test how well the models could be generalized. These essays were generated with the specific instructions to show or omit one of

the personality traits clearly. Accuracy was therefore only assessed on predicting the personality trait shown or not shown in the specific essay. Only GloVe and LIWC were tested due to the poor performance of other models.

——	True Positives	False Positives	Accuracy
cEXT	0	5	0.0
cNEU	1	4	0.2
cAGR	3	2	0.6
cCON	4	1	0.8
cOPN	2	3	0.4
Average	2	3	0.4

Table 5: LIWC classification results on positive ChatGPT Essays

——	True Negatives	False Negatives	Accuracy
cEXT	5	0	1.0
cNEU	4	1	0.8
cAGR	1	4	0.2
cCON	3	2	0.6
cOPN	2	3	0.4
Average	3	2	0.6

Table 6: LIWC classification results on negative ChatGPT Essays

——	True Positives	False Positives	Accuracy
cEXT	2	3	0.4
cNEU	1	4	0.2
cAGR	3	2	0.6
cCON	3	2	0.6
cOPN	0	5	0.0
Average	1.8	3.2	.36

Table 7: GloVe classification results on positive ChatGPT Essays

——	True Positives	False Positives	Accuracy
cEXT	1	4	0.2
cNEU	2	3	0.4
cAGR	3	2	0.6
cCON	3	2	0.6
cOPN	4	1	0.8
Average	2.6	2.4	0.52

Table 8: GloVe classification results on negative ChatGPT Essays

5.3 Analyzing Metrics

5.3.1 testing on original data set

All models tested performed somewhat poorly. The Bag of Words model had a recall value of 1 for 3 of the personality traits, suggesting a single class was predicted for every test essay. The GloVe model performed the best, with balanced accuracy, precision, recall, and F1 scores. These scores were the highest (or of almost equal value) of the 3 non-BOW models across every one of these metrics

5.3.2 testing on outside data

To analyze the performance of our outside models, we will look at the accuracy of each model in correctly identifying both when an essay is emblematic of a personality trait and when an essay is not emblematic of a personality trait. If a model has particularly high performance in both, it is able to distinguish between having that trait and not having that trait well.

Our LIWC model had high accuracies in predicting both presence and absence of conscientiousness. It predicted presence for extraversion and neuroticism much more than absence.

Our GloVe model had identical accuracies in predicting presence and absence of agreeableness and conscientiousness, implying some degree of differentiating ability for these traits. These accuracies are only .6 however, meaning they are not very statistically significant.

5.3.3 overall performance

Overall, our more complex models performed better than our BOW model, reaching a higher accuracy while actually distinguishing between positive and negative instances of personality traits. LIWC's ability to predict conscientiousness was interestingly high, both in generalization and for the original data set. Combining models did not prove to be an effective technique for increasing performance, and just seemed to perform between their respective models. The best accuracy of any trait of any model was performed by GloVe on openness, with a .65 accuracy and .67 precision, far better than any other model for any other trait. This is likely due to the nature of trait Openness, a trait that is highly suggestive of using complex words that would likely be weighted highly within the GloVe dictionary. This is quite an interesting find and one that can definitely be studied in further work.

6 Further Work

It is clear from the research done in this paper that more work has to be done in order to create more accurate models. The metrics that were found using our models is on par with other metrics seen in other contemporary papers. [3] However, this cannot be the ceiling of performance that can be achieved. Our team has two points for future works to improve on. Just as highlighted in Tighe's paper, a more extensive data set with ranges for the personality traits instead of binary values is needed.[3] This will be incredibly difficult, as annotating such a complicated topic requires plenty of time and experience. Second, it is possible that using a BERT model will increase accuracy and performance in classification.

References

[1] AMIRHOSSEINI, M. H., AND KAZEMIAN, H. Machine learning approach to personality type prediction based on the myers–briggs type indicator®. *Multimodal Technologies and Interaction 4*, 1 (2020), 9.

[2] BRITZ, D. Understanding convolutional neural networks for nlp. *URL: http://www. wildml. com/2015/11/understanding-convolutional-neuralnetworks-fornlp/(visited on 11/07/2015)* (2015).

[3] TIGHE, E. P., URETA, J. C., POLLO, B. A. L., CHENG, C. K., AND DE DIOS BULOS, R. Personality trait classification of essays with the application of feature reduction., 2016.

[4] WANG, W., AND GANG, J. Application of convolutional neural network in natural language processing. In *2018 International Conference on Information Systems and Computer Aided Education (ICISCAE)* (2018), pp. 64–70.