Title: A Language Model of the Myers-Briggs Personality Type Index
Authors: Ian Knight, Ian Naccarella, Chase Navellier

Background

The Myers-Briggs test is personality classification system which indexes a person into one of two different classifications within four orthogonal dimensions. These personality dimensions are be delineated as follows:

Personality Dimension[1]	Option 1	Option 2
"Favorite World"	Extraversion (E)	Introversion (I)
"Information"	Sensing (S)	Intuition (N)
"Decisions"	Thinking (T)	Feeling (F)
"Structure"	Judging (J)	Perceiving (P)

These dimensions tend to be signified by the letters I vs. E, N vs. S, F vs. T, and P vs. J, which is a standard we will adhere to throughout this paper. The four binary choices lead to a total of 16 (2^4) personality types, referred to by juxtaposition of the four classifications (e.g. INTJ). As we will describe in the next section, our project will explore various modeling and learning methodologies for these personality types within the framework of natural language.

Task Definition

Our focus for this project was building a Myers-Briggs personality classifier capable of sorting people into their Myers-Briggs Type Index (MBTI) personality type based on text samples from their social media posts. The motivations for building such a classifier are twofold. First, the pervasiveness of social media means that such a classifier would have ample data on which to run personality assessments, allowing more people to gain access to their MBTI personality type, and perhaps far more reliably and more quickly. There is significant interest in this area within the academic realm of psychology as well as the private sector. For example, many employers wish to know more about the personality of potential hires, so as to better manage the culture of their firm.[2] Our second motivation is centers on the potential for our classifier to be more accurate than currently available tests as evinced by the fact that retest error rates for personality tests

[1] More information on this topic can be found at the MBTI website (http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/home.htm?bhcp=1)

[2] In fact we were approached by Kevin Lee of Pear Ventures about this exact application of our program

administered by trained psychologists currently hover around 50%. That is, the there is a probability of about half that taking the test twice in two different contexts will yield different classifications. Thus, our classifier could serve as a verification system for these initial tests as a means of allowing people to have more confidence in their results. Indeed, a text-based classifier would be able to operate on a far larger amount of data than that given in a single personality test. Furthermore, the ability to produce an accurate text-based classifier has significant potential implications for the field of psychology itself, since the connection between natural language and personality type is non-trivial. This is as seen in multiple studies, such as those by Scherer et al.[3]

We will dive into the details of our approach in the next section, but our general aim was to feed in a series of social media posts and have our classifier correctly identify the MBTI personality type of the user who wrote said posts. As we will be detailed in later discussion, we had two main measures of success against which to judge our results, namely the retest proficiency of trained psychologists (~50%) and the current classification accuracies found in the literature. Given the machine learning nature of our problem we emphasized classification accuracy above time and memory constraints since we did not run into significant issues regarding those. It should also be noted that as a paradigm for solving this problem, we consistently broke it down into four sub-problems by classifying each dimension of personality separately. This was done to leverage the orthogonal structure of the personality index, as well as to provide granularity into testing accuracy between dimensions.

Infrastructure

Our training/development and the majority of our test sets came from the enclosed Kaggle data set which contains a mapping of 8600 individuals' Myers-Briggs personality types to 50 of their social media posts on the website personalitycafe.com, which focuses on discussion of personality psychology. For example, this particular individual is classified as an INFP and the following are some of their posts (which would be used as input):

- "Basically, my main questions are: What do you believe in? What are the values that matter the most to you? (Sorry if I tend to be confusing)"
- "One of those days when you feel like the world is your oyster :)"
- "Hey, The last movie I saw was The Neon Demon. Quite creepy if you ask me... Not a movie I would watch twice. But anyway, to each his own."

Because this website has a sample bias of the population's personality types (i.e. there are more introverts than extroverts in the world),[1] we cleaned the dataset to equally represent the frequency of each personality dimension. Therefore, the baseline accuracy for any of the four dimensions is random chance (50%). As we noted earlier, because retest accuracy among professionals for one

[3] Scherer, K. R. (1979). Personality markers in speech. In Scherer, K. R., & Giles, H. (Eds.), Social markers in speech, pp. 147–209. Cambridge University Press.

of the 16 categories is ~50%, our oracle accuracy within one dimension is ~85% ($0.85^4 = 0.5$). As we will see, it was actually possible for our model to beat this oracle, which lends itself to interesting questions that we address at the end of this report. Concretely, the changes to the infrastructure include cleaning personality data points to represent each of the 16 groups evenly as well as cleaning the 4 indices to represent that specific dimension evenly. We will also address how the particular cleaning methodology affected our results.

Approach

Overall, we explored two different models for learning: a linear classifier and a neural network. For all of these models, particularly the linear model, developing a feature extractor to capture the psychologically relevant nuances of natural language in a high-dimensional space was a key focus. For each feature vector, we normalized by the number of words so that the absolute magnitude doesn't vary. We used our results from the linear model to consider the effectiveness of each following feature(s): bag of words, part of speech frequency (using nltk's pos_tag), sentiment analysis (using nltk's textblob), magnitude of sentiment, word length, frequency, relative amount of punctuation usage, number of quantitative words, number of derogatory words, number of misspelled words, and character n-grams.[4] Following this, we will discuss the modeling and training of a neural network that uses both the previously discussed features as well as nltk's word2vec feature which uses both the preceding and following words to map a word to a high dimensional space.

Testing Paradigm

For testing, we further separated a given test set into "batches" of a certain variable size (e.g. 100). Specifically, this mean grouping posts all of the same MBTI type, classifying each post, and classifying the whole batch as one of the binary classifications depending on which classification was more common within the batch testing. For example, when testing 10 posts in a batch, say 7 are classified with +1 and 3 with -1. This would lead us to classify the whole batch (i.e. the user who wrote the posts in said batch) as +1, because +1 is more common than -1. Following this paradigm, we test each user and present the accuracy from these final batch classifications.

Error Analysis

Our chosen algorithm was stochastic gradient descent due to the relatively large size of the data set we were training on. We began by optimizing the hyperparameters eta and number of iterations with a validation set equal to 15% of our data for each new type of feature vector. We proceeded to divide the remaining data into test and training datasets containing 85% and 15% of the data respectively and sampled at random. For our final results, we performed a randomized cross-validation by averaging our results from ten different runs.

---

[4] Poria, Soujanya, et al. "Common sense knowledge based personality recognition from text." *Mexican International Conference on Artificial Intelligence*. Springer, Berlin, Heidelberg, 2013.

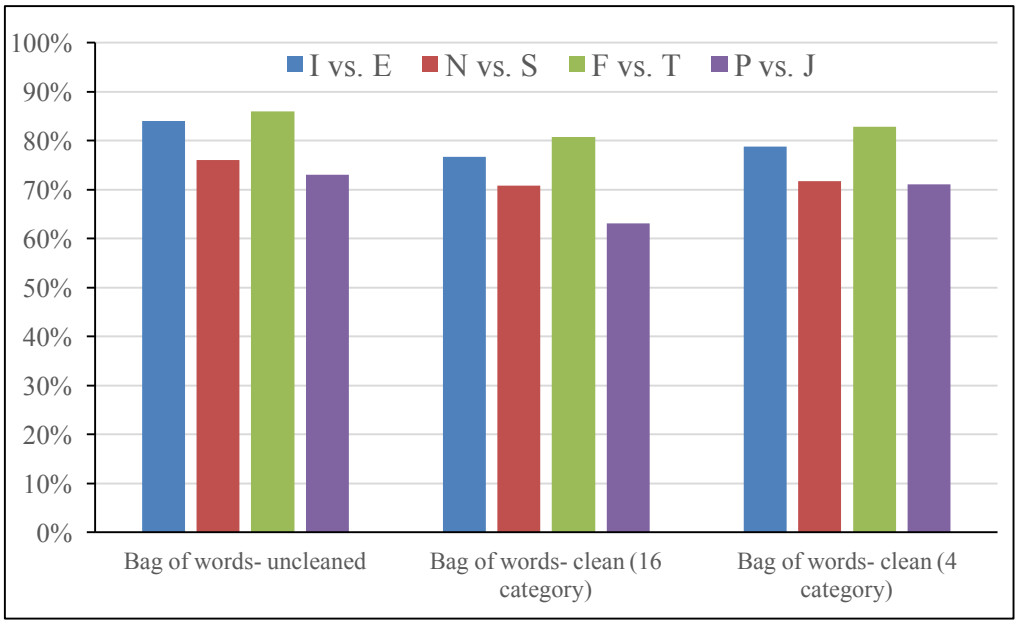

Figure 1 | Graph comparing classification accuracies on different datasets differentiated by uniformity of personality classification.

Due to the distribution bias in our original data set, using uncleaned data leads to classification accuracies of 84%, 76%, 86%, and 73% for I vs. E, N vs. S, F vs. T, and P vs. J respectively. While using the bag of words feature to compare versions of the data cleaned in different manners, we considered two different approaches: elimination of data to make each of the 16 categories (i.e. INTJ, ENTP, ...) equal in size and elimination of data to make each of the four orthogonal indices (i.e. I vs. E, ...) equal in size. Overall, we can see that the cleaning only within one index at a time preserves significant accuracy, largely due to the ability to retain significantly extra data. We see from the uncleaned data that by leaving a bias we get much higher test accuracy since the baseline is already well over 60% in all four indices.

With regard to our features and overall learning optimization, we considered ways to avoid over-fitting while rationally extracting the most useful predictors for the model. In order to test the utility of various features, we implemented them each separately (regardless of overall vector size) in order to gauge how significantly about the model would be able to train on that specific feature. We also tested various combinations of the features to create a top performing feature extractor.

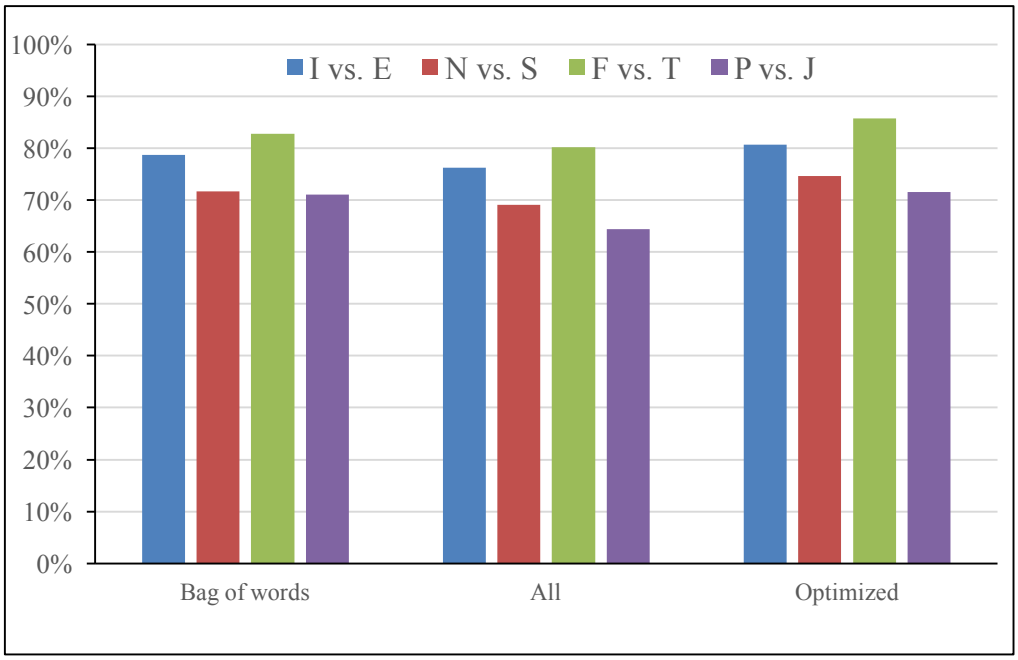

Figure 2 | Graph comparing classification accuracies of different feature combinations

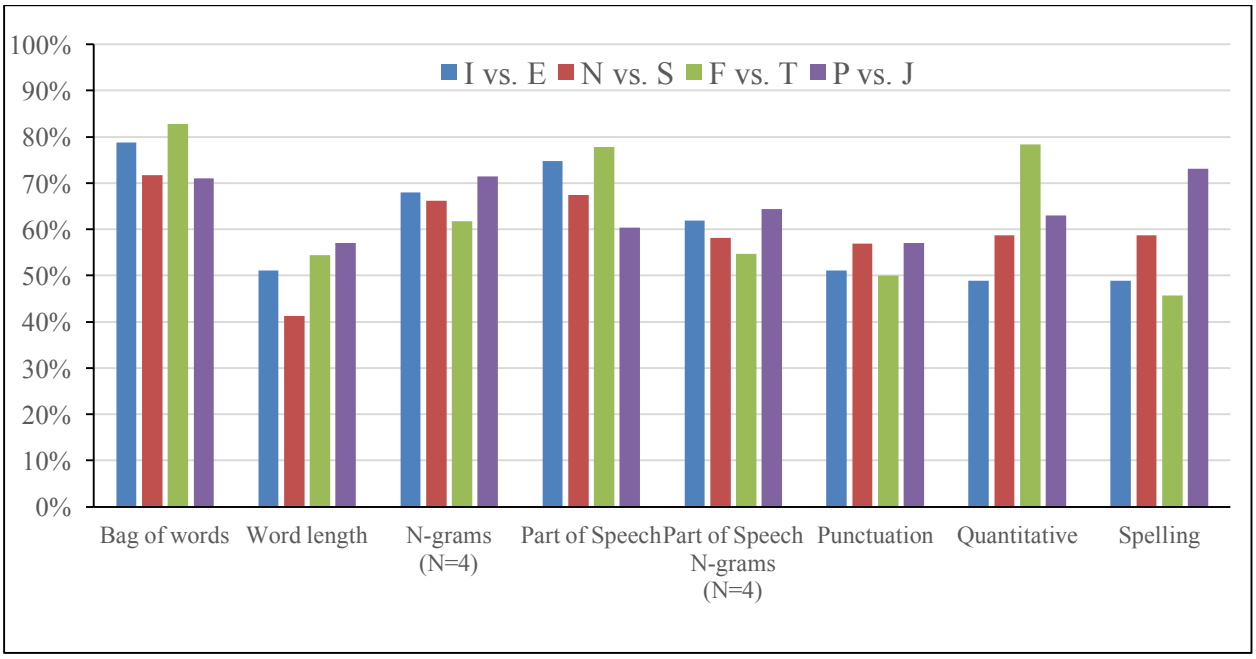

Figure 3 | Graph comparing the classification accuracy of each of the features tested

Interestingly, we see that bag of words performs the best as a single feature and better than using all features together. This indicates some level of over-fitting when all possible features are used. We also note that for T vs. F, relative number of quantitative words had significant accuracy for a single-dimensional feature vector. For P vs. J, we note that relative number of misspellings also had significant accuracy. Part of speech was useful across all four indices, but particularly for T vs. F. N-grams was also successful across all four indices, and we ultimately chose n=4 due to an

investigation of both the literature as well as seeing similarly optimal values between n=3 to n=6 in our empirical findings. Word length and punctuation usage were not significantly above 50% for most of the indices. Our highest yielding feature vector ultimately incorporated bag of words, 4-grams, part of speech, quantitative fraction of words, and misspelled fraction of words.

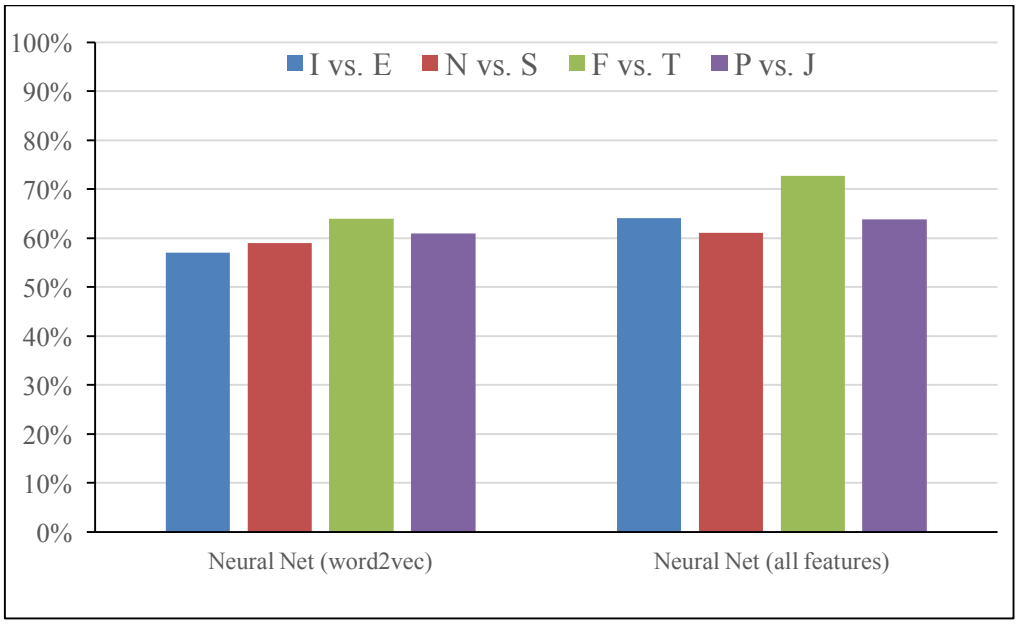

Figure 4 | Graph comparing classification accuracies using a non-recurrent neural network with word2vec average across the post and manually extracted features

For the neural network model, we used a vanilla neural net with a single hidden layer using the scipy toolkit. We trained it using a relatively simple input and as well as with all of the features considered in our linear model. The simple input was modeled as word2vec averaged across all the words in a post. This of course is a poor way to represent the meaning of natural language in a high dimensional space, but nonetheless provided an initial starting point. We then trained the neural network with an extensive feature vector which ultimately had surprisingly promising results. We also used our optimized linear classifier vector as another data point, but it did not have significantly different results from using all of the features at once. The optimized linear classifier therefore still had superior accuracy.

Literature Review

There is a long history of researchers attempting to classify people based on things they have written. Prior work from multiple different sources indicate that it is possible to determine

the identity, gender, and even education level of an author from textual analysis.[5,6,7] Building off of this, in recent years several researchers have attempted to go even further and classify people into their personality type based on similar analyses. There exists ample evidence that how a person writes can provide insight into their mental state, as shown by Rude et al. when they discovered that people who are depressed tend to use more first-person singular pronouns.[8] Thus far, researchers have, for the most part, remained within two major systems of personality classification, namely MBTI and "Big Five," which correlate highly with each other. We chose to use the MBTI system as mentioned previously because of its greater visibility in the community which made it easier to find data pertaining to it and also allowed us to compare our results more easily to those found in the literature.

The vast majority of the literature we found regarding personality classification utilized linear classifiers with a variety of features tuned to optimize classification accuracy. Interestingly, work by Juola and Ryan seems to indicate that N-grams are amongst the best performing of the tested feature vectors.[9] In our own analysis, we utilized a variety of the linear predictors most often used in the literature, including N-grams, as well as some of our own invention. There are, however, notable differences between our work and the literature. First, we were attempting to classify people into their personality type based on a series of short social media posts whereas most papers we found utilized paragraphs of text.[6] We would argue that our extension to this classification problem is more useful given the explosion of social media data over the past few years. Second, we also utilized a neural network to help with our classification which appears to be unique amongst people attempting to solve the specific problem of personality classification. Unfortunately we did not get exceptional results using the neural net but we believe with proper modification, it could be a useful tool for personality classification.

Conclusion

In this section, we will compare our results to what we believe are the highest classification accuracies currently attained in the literature.

Category	Our Results - Linear Classifier (%)	Our Results - Neural Network (%)	Noecker, Ryan, and Juola (2013) (%)[10]	Luyckx and Daelemans

[5] Binongo, J. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. Chance, 16(2): 9–17.

[6] Juola, P. and Baayen, R. H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. Literary and Linguist Computing, 20(Suppl): 59–67.

[7] Koppel, M., Argamon, S., and Shimoni, A. (2002). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4): 401–12.

[8] Rude, S., Gortner, E., and Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. Cognition and Emotion, 18: 1121–33.

[9] Juola, P. and Ryan, M. (2008). Authorship Attribution, Similarity, and Noncommutative Divergence Measures. Selected Papers from the Chicago DHCS Colloquium. Chicago, IL: Chicago Colloquium on Digital Humanities and Computer Science.

[10] Noecker Jr, John, Michael Ryan, and Patrick Juola. "Psychological profiling through textual analysis." *Literary and Linguistic Computing* 28.3 (2013): 382-387.

				(2008) (%)[11]
Attitudes (E versus I)	80.32	64.12	76.35	65.52
Perceiving (N versus S)	74.01	61.08	76.54	62.07
Judging (T versus F)	85.65	72.64	80.47	73.79
Lifestyle (J versus P)	65.37	63.82	84.45	82.07
Average	76.34	65.42	79.45	70.86

As you can see, our linear classifier model's accuracy fell right around the accuracies of the current records for this type of classification. Some of this is attributable to our dataset which most likely lends itself to personality assessments, but at the same time we can note that we came very near one of our evaluation metrics.

One thing that it is interesting to note is that that n-grams (where n=4) was the most effective feature for Juola and Ryan (2008), whereas bag of words was more effective for us. This is potentially a function of the source of our data versus theirs. They used arbitrary text from speeches and books whereas our data was pulled from a website directed towards psychology. It is therefore conceivable that the subject matter of the text from our data set was more relevant to the task at hand, therefore leading to bag of word's relative success.

After completing our model, we decided it would be interesting to consider how our model performed while classifying text from different data sources. As a fun activity, we scraped 30,000 of Donald Trump's tweets and classified his personality type as ESTJ, which is generally the most popular assumption of his personality type per several sources.[12] While this is in no way definitive proof of the accuracy of our model, it is nice to see that it has potential real-world applications.

Thus, we can see that we were able to achieve high accuracies relative to those achieved in the literature as well as when compared to our baseline and oracle. Furthermore, we were able to learn more about which features are most useful for personality classification. Lastly, we implemented a real-world application of our classifier and saw that it holds up when utilized on a different dataset. Overall, while there are definitely improvements that need to be made, we

[11] Luyckx, K. and Daelemans, W. (2008a). Personae: A Corpus for Author and Personality Prediction from Text. Proceedings of the 6th Language Resources and Evaluation Conference. Marrakech, Morocco: International Conference on Language Resources and Evaluation.
[12] http://www.personalitypathways.com/personality-type/donald-trump-estj-personality/

believe that this represents a potential opportunity for validating professional personality assessments and could find applications in the near future.