Parker Williamson

5/9/2018

Springboard Data Science Career Track

Capstone Project 2 Final Report – Text Mash

**Problem**

Improve interactions between people and improve user experience. Matching people and finding fun activities are very labor intensive manual process and very important for people’s life experiences. Understand who people are and what they are likely to want is one of the most important things to identify to really make a difference in most people’s lives.

**Client**

Any apps would be able to use the personality traits to improve the experience. Personality traits can be used to pair like-minded people for dating, assistance or finding what activities all people in a group would be most likely to enjoy based on all of their personalities. It could also be used to present information in the way each type of person finds most appealing. It could even be used to understand how an email would make you seem (optimistic, extroverted, ext). If the solution is strong enough it could be useful enough to start a business around.

**Dataset**

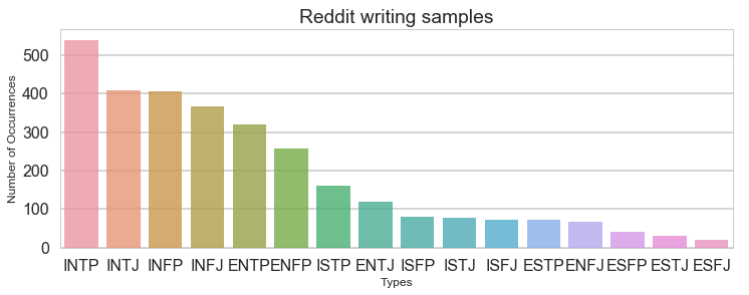
I focused on accurately predicting MBTI personality types. The MBTI raw data I will get from Kaggle’s Myers Briggs dataset (<https://www.kaggle.com/datasnaek/mbti-type>). Some personality types didn’t post as much so I will try to find more data to make a large enough sample, so that can have more examples of the types less vocal on forums. That extra data I will collected from Reddit’s MBTI subbreddits which have posters self labelled type and separate subreddits for different types. Text data was normalized using (<http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/>).

**Findings**

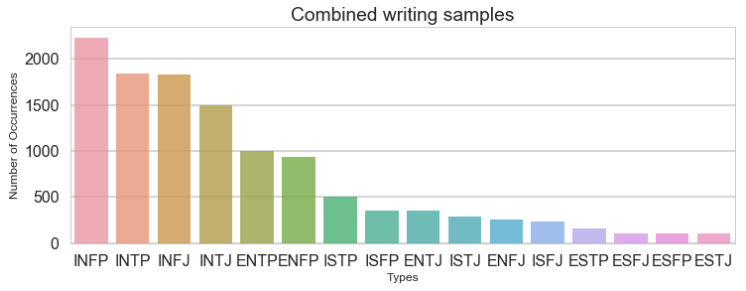
The initial Kaggle dataset is biased towards introversion, and so are the subreddits which means some personality types are more likely to post on forums. Although that means that the personality traits are more likely to influence what people enjoy, that makes it harder to collect an unbiased dataset. To try and achieve an unbiased dataset I combined the Kaggle dataset with the data I collected from reddit.

1750 
1500 
1000 
E 750 
E 500 
INFP 
INF] 
INTP 
ENTP 
ENFP 
ISFP ENTJ 
Types 
IST] 
ISFJ 
ESTP 
ESFP 
ESFJ 
EST] 

The Kaggle dataset’s bias is show above with the clear variation between the introverted and extroverted types. The reddit database’s distribution is shown below, it is biased as well but also helps to raise the number of examples of the uncommonly posting types.

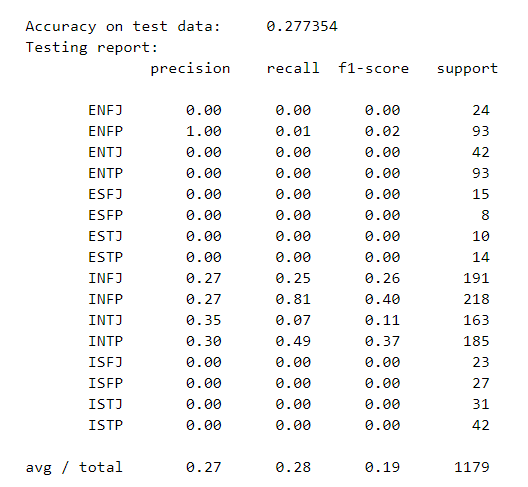


The combined samples with some extra Reddit are shown below.

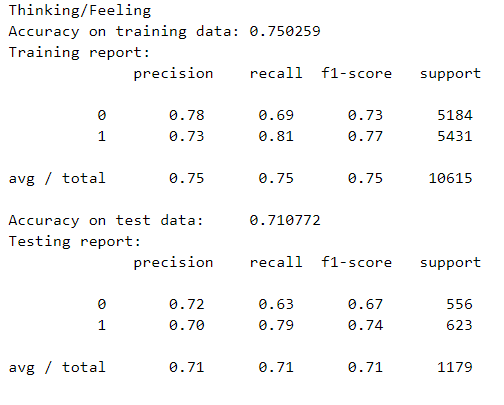


To clean the text data I removed English stop words, replaced URLs with <URL> and replaced all the 4 letter personality types with <MBTI>. For the existing Kaggle dataset sample the <MBTI> replacement factored in at about 20% of some of the models, so eliminating that means there may need more data or better initial text analysis. The accuracy general accuracy of predicting the test type from the original data without the Myers Brigg class in the text is shown below.

I tried stemming the word (excluding suffixes), but it made two of the types go up and two go down by nearly the same amount. Since stemming is slow I then excluded it in future models.

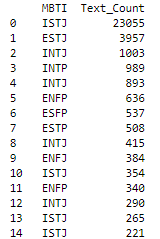


Based on the image above is clear that the model is not good at predicting all the personality traits at once. The predictions for each individual traits are 60-75% accurate depending on the training test split.

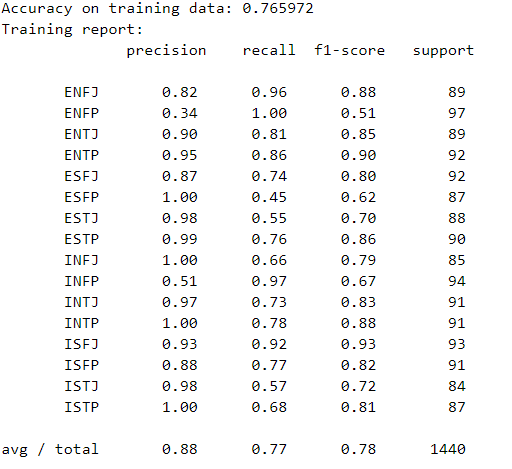


I tested a number of different models; the main difference between the different models was how much they predicted the less common class. Both XGboost and linear SVC models suffered because they predicted the more common class too much even with an inverse weighting. The other three models, MultinomialNB, Logistic Regression and SGD classifier were all fairly similar, but Logistic Regression had the highest F1 score of .71. That shows that Logistic Regression is the optimal model to use since CNN and RNN over fit as well.

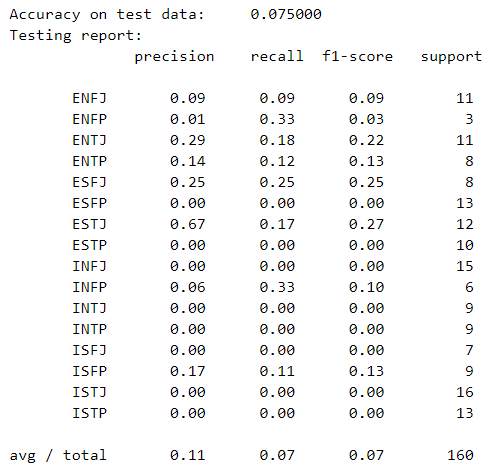
I also tested the accuracy of predicting my and a couple close friend’s personality types from my texting history and compared them to their stated MBTI. The difference between texting and forum posts made the predictions all the same until I normalized the results to a dataset of texting I found online.



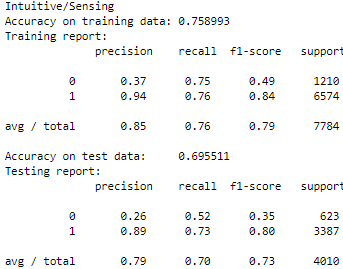
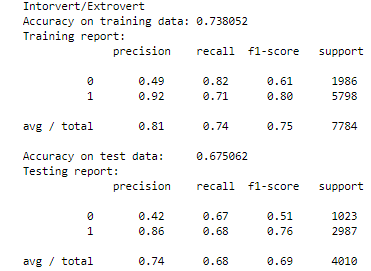
When I only use 100 different users from each type there was a lot more variance and although the training accuracy was greater the testing accuracy was very unhelpful. It had been estimating the most likely and there was not enough new info to accurately estimate using.

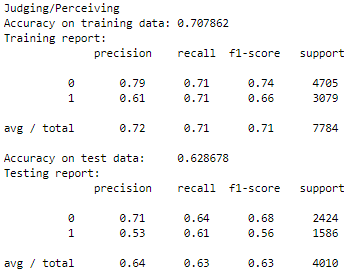
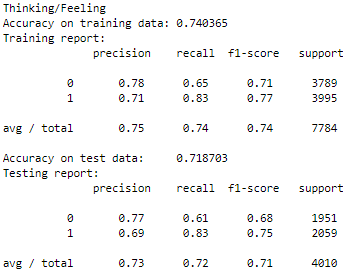


The accuracy on the test data to each personality type is low for the smaller sample size of 100 users per type.

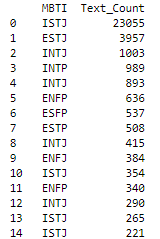


Each estimation individually is around 65% on its own, just all 4 are hard to get right at the same time.



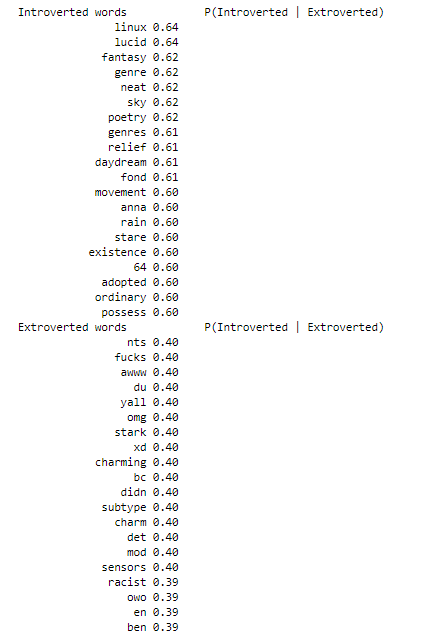


The types assigned based on texting history vary a lot but none of the 4 people I know have an accurate assignment.



Including the reddit data decreased the test accuracy by about 2%, but the Kaggle MBTI data was not as applicable to average texting, because it is based on text where people are talking about their personalities. Adding the Reddit data helps prevent overfitting, for example, one of the most strongly weighted words (towards extroversion) is socionics. Socionics is not a word used in many texting conversations, but it was used fairly commonly in the MBTI forum drawn from for the Kaggle dataset.

The introverted words do tend to make sense as more intellectual or emotional.



Overall it is clear that the MLP classifier is an okay, but not accurate enough model for personality prediction based on text. It is around 65-70% accurate per Meyers Briggs personality trait. To be in line with the normal method of testing personality it would need to be around 85% accurate. This does show how Reddit can be used to expand the dataset some, and how different types of posts tend to indicate different personality types.