Personality Prediction

Harsh Sharma (B20CS017)

Abstract

The focus of this project is to use machine learning to create a classifier that can classify people into personality types in the MyersBriggs Type Index (MBTI) based on text samples from social media posts. There are two motivations for creating such a classifier. First, the proliferation of social media has sufficient data for such classifiers to carry out personality assessments, allowing more people to access MBTI personality types, potentially much more reliably and quickly. It means to make it accessible. There is considerable interest in this area, both in the academic and private sectors of psychology.For example, many employers want to know more about the personality of potential employees in order to better manage their company's culture. Our second motivation is that the classifier is more accurate than the tests currently available, as evidenced by the retest error rate of personality tests performed by trained psychologists, which is currently around 0.5. Focusing on the possibilities. In other words, running the test twice in two different contexts is about half as likely to result in different classifications. Therefore, our classifier can act as a validation system for these first tests and increase confidence in the results. In fact, text-based classifiers can process much larger amounts of data than a single personality test.

1 Introduction

In the scientific field of psychology, the concept of personality is powerful, but is considered an inaccurately defined construct. Therefore, psychologists will greatly benefit from the development of more specific and empirical measurements of existing models of personality. Our project aims to better understand one such model, the Myers Briggs Type Indicator (MBTI). We plan to use machine learning to create a classifier that takes text (such as a social media post) as input and produces a prediction of the MBTI personality type of the creator of that text as output. A successful implementation of such a classifier demonstrates MBTI's strong linguistic foundation and, in some cases, its general character. In addition, the relationship between natural language and personality types is important, so the ability to create accurate text-based classifiers has significant potential impact on the psychology field itself.

2 Background/Related Work

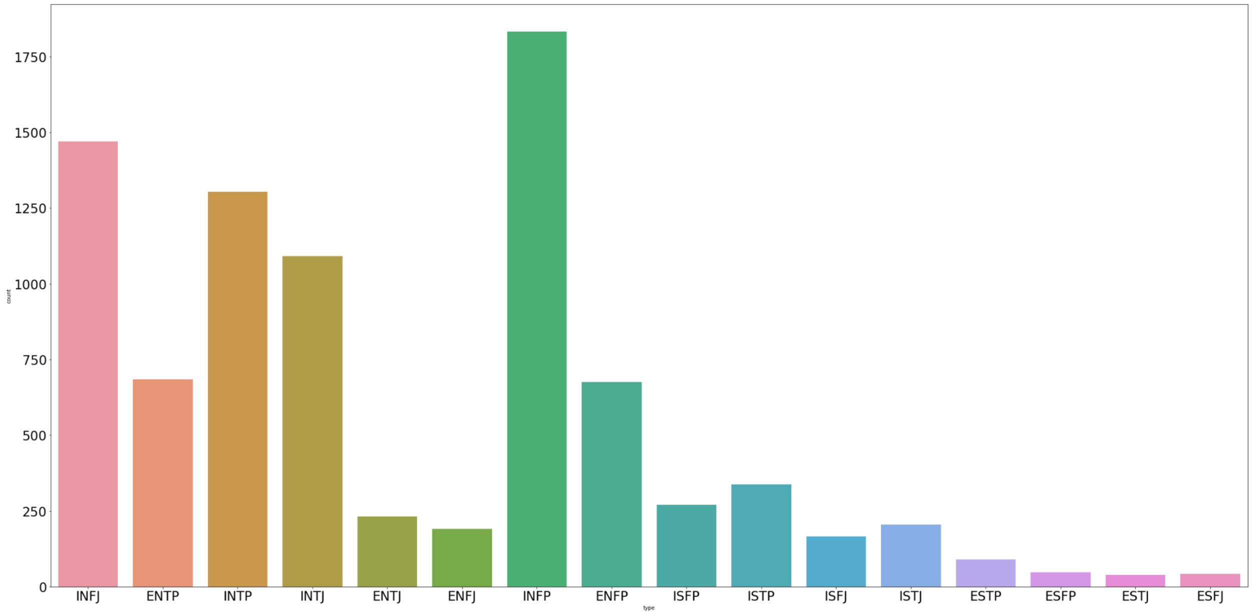
The Myers Briggs Type Indicator (or MBTI for short) is a personality type system that divides everyone into 16 distinct personality types across 4 axis:

- Introversion (I) – Extroversion (E) - a measure of how much an individual prefers their outer or inner world.
- Intuition (N) – Sensing (S) - a measure of how much an individual processes information through the five senses versus impressions through patterns.
- Thinking (T) – Feeling (F) - a measure of preference for objective principles and facts versus weighing the emotional perspectives of others.
- Judging (J) – Perceiving (P) - a measure of how much an individual prefers a planned and ordered life versus a flexible and spontaneous life.

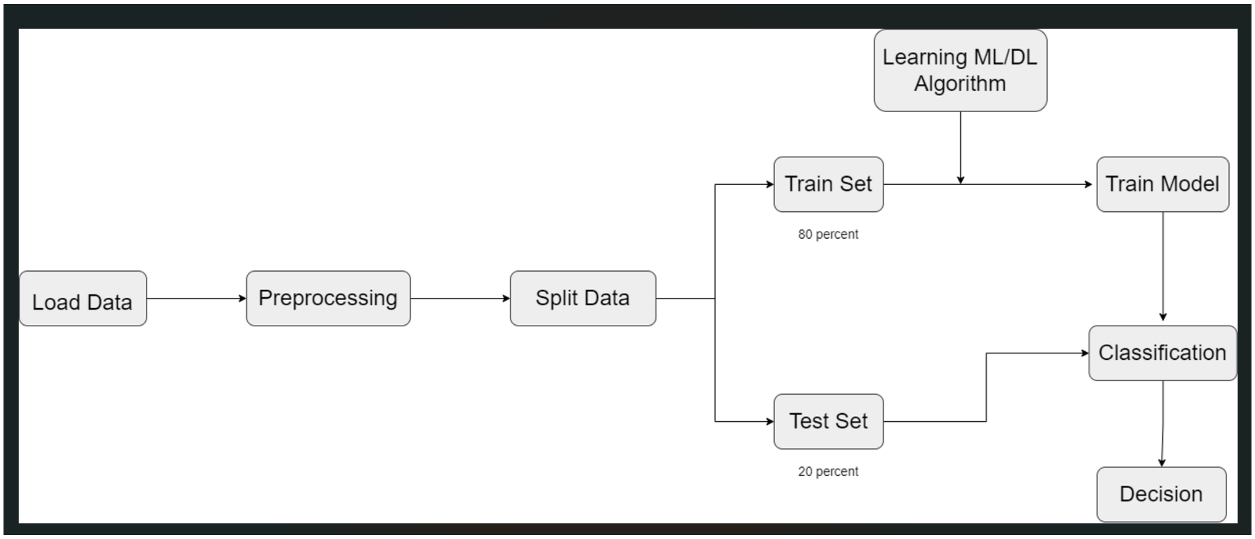
Dataset

This dataset contains over 8600 rows of data, on each row is a person's:

- Type (This persons 4 letter MBTI code/type)
- A section of each of the last 50 things they have posted (Each entry separated by '|||' (3 pipe characters))



3 Approach



3.1 Preprocessing

3.1.1 Selective Word Removal

Since the dataset is from an internet forum where individuals communicate only through written text, it was clearly necessary to remove the word. For example, there were multiple instances of a data point that contained a link to a website. To generalize the model to English, we have removed all data points, including links to websites. Then I removed the so-called "stopwords" from the text to make each word in the data as clear as possible (for example, very common filler words such as "a", "the", "or").).). NLTK for Python. Finally, the specific dataset we are dealing with is from a personality model, especially from a site for explicit discussion of MBTI, so types (eg "INTI", "INFP", etc.). Has been deleted. That is to prevent the model. From "Cheating" by learning to recognize MBTI references by name.

3.1.2 Lemmatization

I used nltk.stem.WordNetLemmatizer to transliterate the text. That is, inflections of the same stem have been converted to dictionary format (for example, "walking", "walking", and "walking" are all "walking"). This allows us to take advantage of the fact that inflections of the same word still have a common meaning.

3.1.3 Tokenization

Tokenized the most frequent words in the lexicalized text. That is, the most common word is 1, the second most common word is 2, and so on. The text at this point is in the form of a list of integers, as all other words in the lexicalized text have been removed.

3.2 Model

3.2.1 Multinomial Naive Bayes

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. The accuracy obtained is 20.8 .

3.2.2 Random Forest Classifier

The Random Forest classifier uses the boosting ensemble method to train on various decision trees and generate aggregated results. This is one of the most commonly used machine learning algorithms. The Random Forest classifier was used after preprocessing of the classification using the default parameters and achieved an accuracy of 45.07 (approximate).

3.2.3 LGBM Classifier

LightGBM is a decision tree-based gradient boosting framework for increasing model efficiency and reducing memory usage. The Lightgbm model was used for training and achieved an accuracy of 66.68.

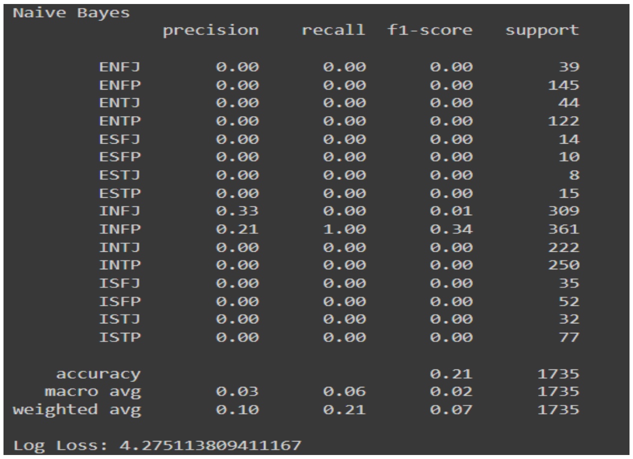
3.2.4 Logistic Regression

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. Logistic regression model was used for training and an accuracy of 58.44 without tuning was obtained and 67.55 with tuned parameters.

4 Evaluation

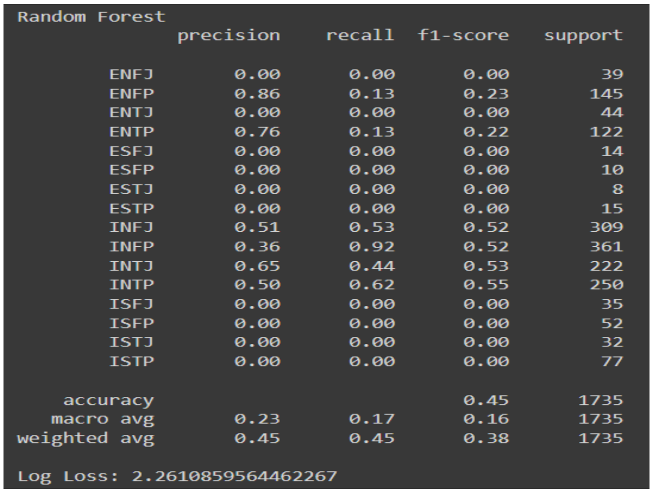
4.1 Class Report

a) Multinomial Naive Bayes



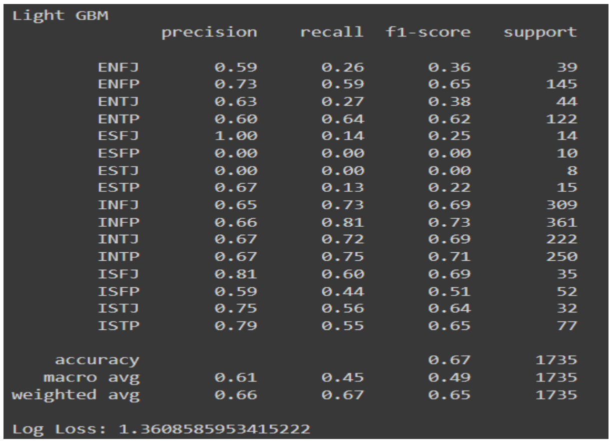
```
Naive Bayes
              precision    recall  f1-score   support

        ENFJ       0.00      0.00      0.00        39
        ENFP       0.00      0.00      0.00       145
        ENTJ       0.00      0.00      0.00        44
        ENTP       0.00      0.00      0.00       122
        ESFJ       0.00      0.00      0.00        14
        ESFP       0.00      0.00      0.00        10
        ESTJ       0.00      0.00      0.00         8
        ESTP       0.00      0.00      0.00        15
        INFJ       0.33      0.00      0.01       309
        INFP       0.21      1.00      0.34       361
        INTJ       0.00      0.00      0.00       222
        INTP       0.00      0.00      0.00       250
        ISFJ       0.00      0.00      0.00        35
        ISFP       0.00      0.00      0.00        52
        ISTJ       0.00      0.00      0.00        32
        ISTP       0.00      0.00      0.00        77

    accuracy                           0.21      1735
   macro avg       0.03      0.06      0.02      1735
weighted avg       0.10      0.21      0.07      1735

Log Loss: 4.275113809411167
```

b) Random Forest



```
Random Forest
              precision    recall  f1-score   support

        ENFJ       0.00      0.00      0.00        39
        ENFP       0.86      0.13      0.23       145
        ENTJ       0.00      0.00      0.00        44
        ENTP       0.76      0.13      0.22       122
        ESFJ       0.00      0.00      0.00        14
        ESFP       0.00      0.00      0.00        10
        ESTJ       0.00      0.00      0.00         8
        ESTP       0.00      0.00      0.00        15
        INFJ       0.51      0.53      0.52       309
        INFP       0.36      0.92      0.52       361
        INTJ       0.65      0.44      0.53       222
        INTP       0.50      0.62      0.55       250
        ISFJ       0.00      0.00      0.00        35
        ISFP       0.00      0.00      0.00        52
        ISTJ       0.00      0.00      0.00        32
        ISTP       0.00      0.00      0.00        77

    accuracy                           0.45      1735
   macro avg       0.23      0.17      0.16      1735
weighted avg       0.45      0.45      0.38      1735

Log Loss: 2.2610859564462267
```

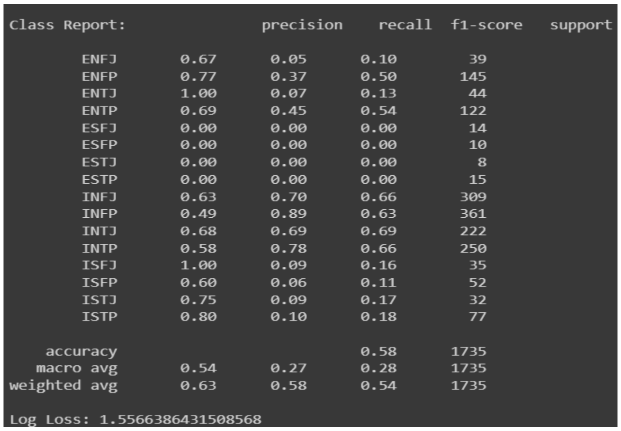
c) Light GBM



```
Light GBM
              precision    recall  f1-score   support

        ENFJ       0.59      0.26      0.36        39
        ENFP       0.73      0.59      0.65       145
        ENTJ       0.63      0.27      0.38        44
        ENTP       0.60      0.64      0.62       122
        ESFJ       1.00      0.14      0.25        14
        ESFP       0.00      0.00      0.00        10
        ESTJ       0.00      0.00      0.00         8
        ESTP       0.67      0.13      0.22        15
        INFJ       0.65      0.73      0.69       309
        INFP       0.66      0.81      0.73       361
        INTJ       0.67      0.72      0.69       222
        INTP       0.67      0.75      0.71       250
        ISFJ       0.81      0.60      0.69        35
        ISFP       0.59      0.44      0.51        52
        ISTJ       0.75      0.56      0.64        32
        ISTP       0.79      0.55      0.65        77

    accuracy                           0.67      1735
   macro avg       0.61      0.45      0.49      1735
weighted avg       0.66      0.67      0.65      1735

Log Loss: 1.3608585953415222
```

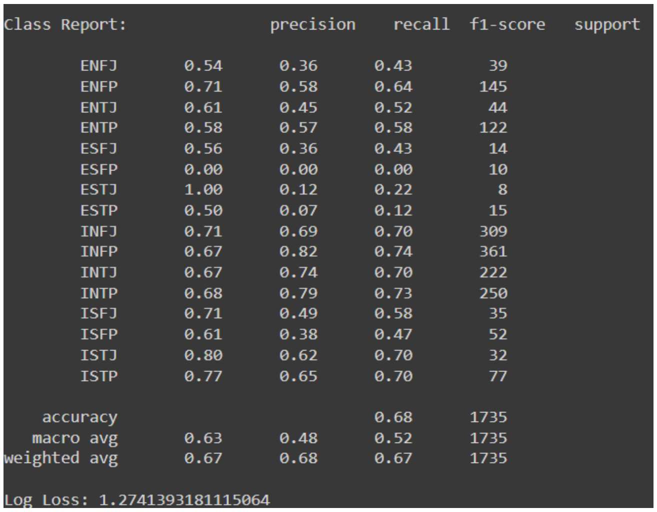
d) Logistic Regression

With default parameter



```
Class Report:              precision    recall  f1-score   support

        ENFJ       0.67      0.05      0.10        39
        ENFP       0.77      0.37      0.50       145
        ENTJ       1.00      0.07      0.13        44
        ENTP       0.69      0.45      0.54       122
        ESFJ       0.00      0.00      0.00        14
        ESFP       0.00      0.00      0.00        10
        ESTJ       0.00      0.00      0.00         8
        ESTP       0.00      0.00      0.00        15
        INFJ       0.63      0.70      0.66       309
        INFP       0.49      0.89      0.63       361
        INTJ       0.68      0.69      0.69       222
        INTP       0.58      0.78      0.66       250
        ISFJ       1.00      0.09      0.16        35
        ISFP       0.60      0.06      0.11        52
        ISTJ       0.75      0.09      0.17        32
        ISTP       0.80      0.10      0.18        77

    accuracy                           0.58      1735
   macro avg       0.54      0.27      0.28      1735
weighted avg       0.63      0.58      0.54      1735

Log Loss: 1.5566386431508568
```

With tuned parameter



```
Class Report:              precision    recall  f1-score   support

        ENFJ       0.54      0.36      0.43        39
        ENFP       0.71      0.58      0.64       145
        ENTJ       0.61      0.45      0.52        44
        ENTP       0.58      0.57      0.58       122
        ESFJ       0.56      0.36      0.43        14
        ESFP       0.00      0.00      0.00        10
        ESTJ       1.00      0.12      0.22         8
        ESTP       0.50      0.07      0.12        15
        INFJ       0.71      0.69      0.70       309
        INFP       0.67      0.82      0.74       361
        INTJ       0.67      0.74      0.70       222
        INTP       0.68      0.79      0.73       250
        ISFJ       0.71      0.49      0.58        35
        ISFP       0.61      0.38      0.47        52
        ISTJ       0.80      0.62      0.70        32
        ISTP       0.77      0.65      0.70        77

    accuracy                           0.68      1735
   macro avg       0.63      0.48      0.52      1735
weighted avg       0.67      0.68      0.67      1735

Log Loss: 1.2741393181115064
```

5
Results and analysis

Among all the models implemented logistic regression and light gbm performs nearly the same with highest accuracy that is less loss . Also we can say that the tree based approach outperforms/outwits other classifying algorithms .