Pattern Recognition and Machine Learning Bonus Project Report
Personality Prediction of person with social media posts

Aditi Tiwari B20EE005 (tiwari.15@iitj.ac.in)

May 1, 2022

Abstract

The Myers Briggs Type Indicator is a personality type system that divides a person into 16 distinct personalities based on introversion, intuition, thinking and perceiving capabilities. You need to identify the personality of a person from the type of posts they put on social media. In this project We particularly focused on feature engineering techniques for text data and provide an in-depth look at the logic, concepts, and properties of the Multilayer Perceptron (MLP) model, an ancestor and the origin of deep neural networks (DNNs) today. I also provide an introduction to a few basic machine learning models as Random Forest, Logistic Regression and K Neighbor.

Index Terms

text, nltk, tokenizer,KNeighborsClassifier, RandomForestClassifier,Logistic Regression, Sequential(CNN) Personality prediction

1 Introduction

In this project we are going to predict the personality of a person from the type of posts they put on social media as (Extraversion (E) vs. Introversion (I), Sensing (S) vs. Intuition (N), Thinking (T) vs. Feeling (F), udging (J) vs. Perceiving (P)). For the project we had used dataset provided in link Link for dataset.

2 Table of Contents

Here I gave the overview of content done in project:

Id	Feature	Parts of feature
1	Feature extraction	Text Cleaning, Visulization
2	Classical ML Model	LogRegression, RandomForests, KNeighbor
3	WordCloud	Wordcloud Plot
4	DeepNeuralNetwork (DNN)	Convolution NeuralNetwork

Important Libraries to work with text:

- nltk: Natural Language Processing with Python provides a practical introduction to programming for language processing.
- Tensorflow: TensorFlow provides a collection of workflows to develop and train models using Python or JavaScript, and to easily deploy in the cloud, on-prem, in the browser, or on-device no matter what language you use.
- Sklearn: Scikit-learn is an open source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection, model evaluation, and many other utilities.

3 Data Description, Visualization and Preprocessing

3.1 Data Description

We're going to use the mbti data for personality prediction(16 different Personality) of the MBTI dataset. Data has 16 differnt classes from (Extraversion (E) vs. Introversion (I), Sensing (S) vs. Intuition (N), Thinking (T) vs. Feeling (F), udging (J) vs. Perceiving (P)) repeating two classes with a variety of emotions We get 1088 rows (2 columns * (type, posts)).

3.2 Data Preprocessing

The first thing we will do is to Clean the text of posts with nltk library. This way we can remove question mark, music, https link, images, split, exclamation .

We are cleaning the posts of data with the english.stopwords . Stop words are common words like 'the', 'and', 'I', etc. that are very frequent in text, and so don't convey insights into the specific topic of a document. We can remove these stop words from the text in a given corpus to clean up the data, and identify words that are more rare and potentially more relevant.

Secondly we converted the spectrograms to csv data file by extracting the features(matrix representation) and further separated data in to training and testing set for better visualization of accuracy.

3.3 Data Visualization

Here is visualization for dataset, the classes appears to be unbalanced. INFP class appers more number of times and class ESTG appers very less number of times .

for our Dataset we had considered [map1 = "I": 0, "E": 1 , map2 = "N": 0, "S": 1, map3 = "T": 0, "F": 1, map4 = "J": 0, "P": 1]. **Plot to show class distribution -**

We made different bar plot to show class distribution.

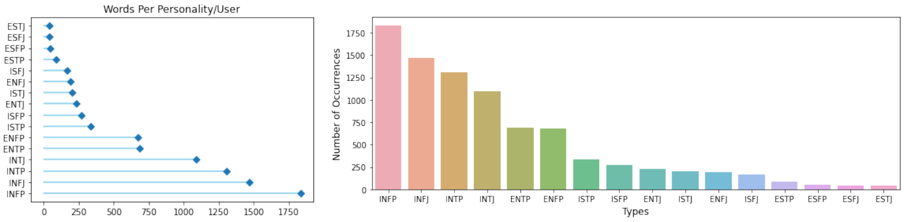


Figure 1: Dataset distribution for different classes available

- **N**umber of posts for each personality.
- (http per comment, music per comment, question per comment, img per comment, excl per comment, ellipsis per comment)

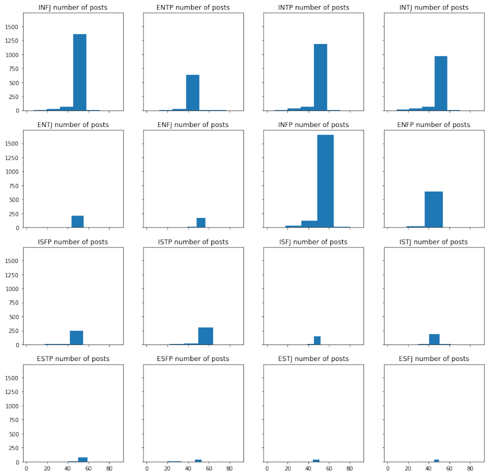


Figure 2: Number of posts for each personality

- **A**verage Parts of Speech Used by Each Personlity.
- We can see from plot that class personality INFP is posting highest posts while class personality ESFJ is posting very less posts.

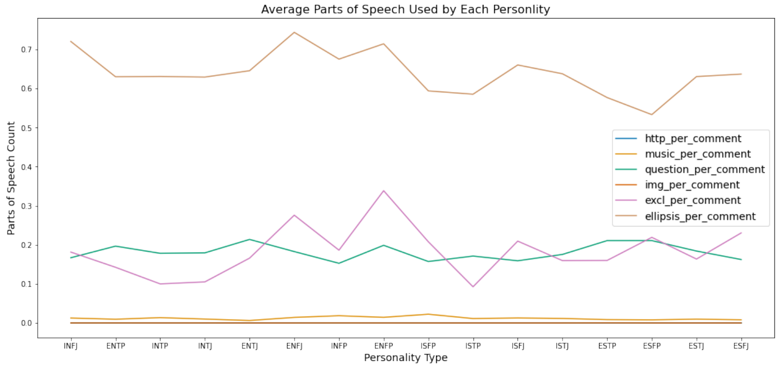


Figure 3: Parts of Speech Used by Each Personlity

4 Word Cloud and most commond words used -

wordCloud :- A word cloud (also called tag cloud or weighted list) is a visual representation of text data. Words are usually single words, and the importance of each is shown with font size or color. We had finded 10 most common word used in posts and make barplot from them. We had also finded most common word used foe every personality person's posts and also make barplot from them. Similarly we did for all classes.

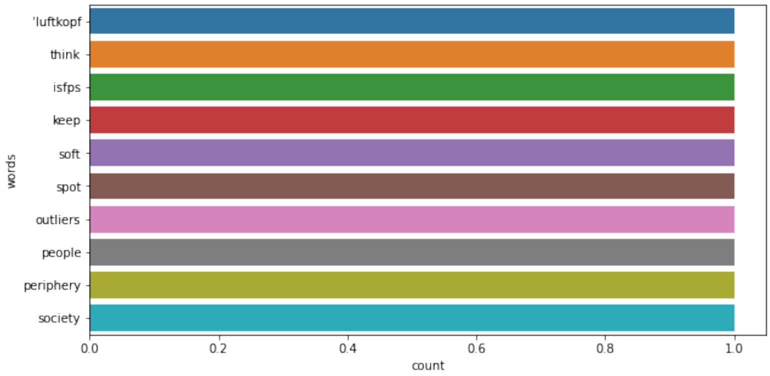


Figure 4: Most common word used in posts

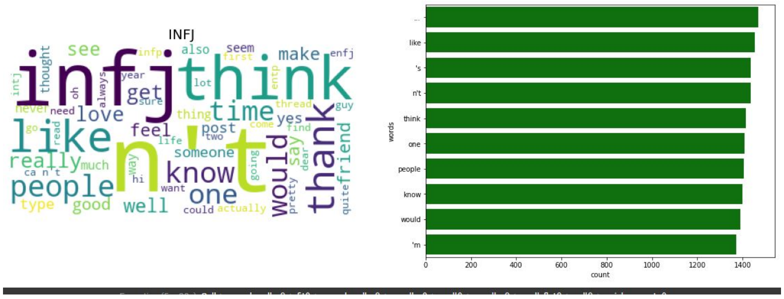


Figure 5: Most common word used in INFJ personality posts

5 Training The Model

5.1 Machine Learning Models

Random Forest Classifier
Random forests algorithm is general technique of bootstrap aggregating to tree learners. Given a training set X with responses Y, bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples. We find the best accuracy of 16.9% for our model.

Logistic Classifier
Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.We find the best accuracy of 24.16% for our model.

K Nearest Neighbor
KNN is one of the simplest forms of machine learning algorithms mostly used for classification. KNN classifies the new data points based on the similarity measure of the earlier stored data points. We performed hyper-parameter tuning using GridSearchCv and have taken best parameter, leaf_size=1, p=1 and n_neighbors=1. Achieved accuracy of 14.02% over testing.

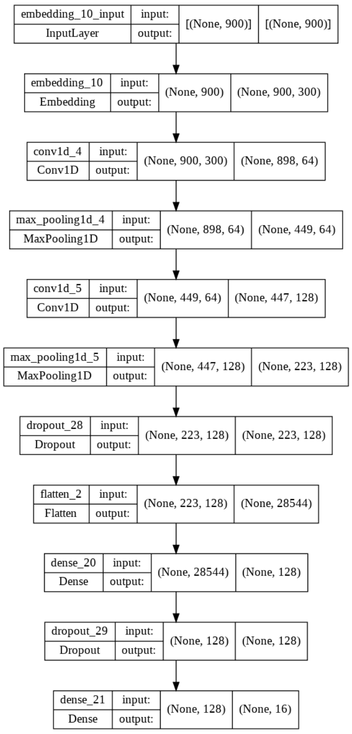
Accuracy for different Simple models like KNN, RFC, XGB

Classifier	KNeighborsClassifier	RandomForestClassifier	XGBoost Classifier
train Accuracy	91.57%	23.63%	47.53%
test Accuracy	16.9%	24.16%	14.02%

5.2 Deep Learning Models

Convolutional Neural Network
A ten hidden layers Neural Network was trained. The activation function for the first layer was Relu (Rectified Linear Unit), for second it was also relu and the last hidden layer utilized Softmax activation function. A Dropout layer with dropout rate of 0.02 was added. The Dropout layer randomly sets input units to 0 at each step during training time, which helps prevent overfitting.The model is then compiled with loss='sparse$_{c}ategorical_{c}rossentropy', metrics =' accuracy'andoptimizer =' adam'andrunnedfor100epochs. Itachievedatrainingaccuracyof97.12\%andtrainingaccuracyof31.11 whichindicatesthatthemodelwasnotabletogeneralisewell.$

CNN Architecture


embedding_10_input | InputLayer | input: | output: | [(None, 900)] | [(None, 900)]
embedding_10 | Embedding | input: | output: | (None, 900) | (None, 900, 300)
conv1d_4 | Conv1D | input: | output: | (None, 900, 300) | (None, 898, 64)
max_pooling1d_4 | MaxPooling1D | input: | output: | (None, 898, 64) | (None, 449, 64)
conv1d_5 | Conv1D | input: | output: | (None, 449, 64) | (None, 447, 128)
max_pooling1d_5 | MaxPooling1D | input: | output: | (None, 447, 128) | (None, 223, 128)
dropout_28 | Dropout | input: | output: | (None, 223, 128) | (None, 223, 128)
flatten_2 | Flatten | input: | output: | (None, 223, 128) | (None, 28544)
dense_20 | Dense | input: | output: | (None, 28544) | (None, 128)
dropout_29 | Dropout | input: | output: | (None, 128) | (None, 128)
dense_21 | Dense | input: | output: | (None, 128) | (None, 16)

Loss and accuracy plot -

In this section
1.we had plotted loss plot vs val loss plot for training and testing model.(loss for training sample is below than testing sample)
2.we had plotted accuracy plot vs val accuracy plot for training and testing model.(accuracy for training sample is higher than testing sample)

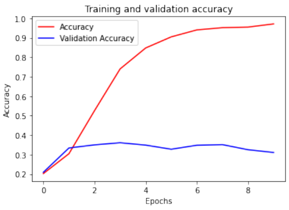


Figure 6: Loss plot

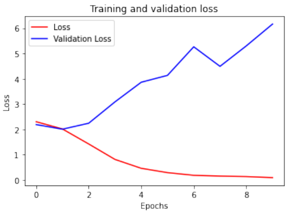
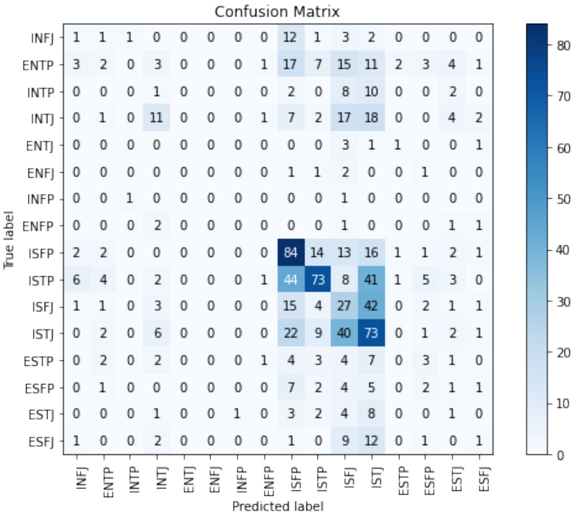


Figure 7: Accuracy plot

Confusion matrix plot -

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.
We can see from the confusion matrix plot highest number of true prediction are in classes (ISFP, ISTP, ISTG).



	INFJ	ENTP	INTP	INTJ	ENTJ	ENFJ	INFP	ENFP	ISFP	ISTP	ISFJ	ISTJ	ESTP	ESFP	ESTJ	ESFJ
INFJ	1	1	1	0	0	0	0	0	12	1	3	2	0	0	0	0
ENTP	3	2	0	3	0	0	0	1	17	7	15	11	2	3	4	1
INTP	0	0	0	1	0	0	0	0	2	0	8	10	0	0	2	0
INTJ	0	1	0	11	0	0	0	1	7	2	17	18	0	0	4	2
ENTJ	0	0	0	0	0	0	0	0	0	0	3	1	1	0	0	1
ENFJ	0	0	0	0	0	0	0	0	1	1	2	0	0	1	0	0
INFP	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
ENFP	0	0	0	2	0	0	0	0	0	0	1	0	0	0	1	1
ISFP	2	2	0	0	0	0	0	0	84	14	13	16	1	1	2	1
ISTP	6	4	0	2	0	0	0	1	44	73	8	41	1	5	3	0
ISFJ	1	1	0	3	0	0	0	0	15	4	27	42	0	2	1	1
ISTJ	0	2	0	6	0	0	0	0	22	9	40	73	0	1	2	1
ESTP	0	2	0	2	0	0	0	1	4	3	4	7	0	3	1	0
ESFP	0	1	0	0	0	0	0	0	7	2	4	5	0	2	1	1
ESTJ	0	0	0	1	0	0	1	0	3	2	4	8	0	0	1	0
ESFJ	1	0	0	2	0	0	0	0	1	0	9	12	0	1	0	1

Figure 8: confusion matrix

6 Conclusion

Classical machine learning models such as KNeighborsClassrifier (KNN), Logistic Regression (LR), and Random Forests have distinct advantages to deep neural networks in many tasks but do not match the performance of even the simplest deep neural network in the task of personality classification.
We're going to have to explore more complicated deep learning methods to get real performance on this dataset. **Convolutional Neural Networks (CNNs) is a DNN candidates for personality classification: CNN is giving more accuracy thann any simple Machine Learning model.It is giving accuray for train as 97% and on test data as 31%.**

type	Classifier	Acc(%)
ML	RandomForest LogisticRegressor KNeighbours	16.87 24.16 14.02
DL	Convolution NeuralNetwork	31.11

References

[1] https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/
https://www.16personalities.com/personality-types
https://www.psychologyjunkie.com/2018/06/19/how-to-spot-each-myers-briggs-personality-type-in-conversation/
https://medium.com/@makingbusinessmatter/the-ultimate-guide-to-myers-briggs-29253737a966
https://www.greeneresources.com/blog/culture/personality-tests-in-hiring/