**MBTI Personality prediction using BERT**

1. **Introduction:**

MBTI is a personality Myers Briggs Type Indicator is a personality type system that divides personalities into 16 distinct types across 4 axes:

Where each person will have a personality abbreviated from the combination of all 4 axis i.e. someone who is introverted, relies more on intuitions, feeling, and perceive rather than judge will be be labelled as an INFP. The combinations of all these axes creates a more complex traits, strengths, and weaknesses.



It is one of, if not the, the most popular personality test in the world. It is used in businesses, online, for fun, for research and lots more. A simple google search reveals all of the different ways the test has been used over time. It’s safe to say that this test is still very relevant in the world in terms of its use.

1. **Objectives:**

The objective of our project is:

* To predict personality of a person based on the given post.

1. **Methodology:**

Data collection

Data pre-processing

Model building

Testing and Prediction

1. **Work progress / plan & Implementation**

The project can be divided into Six phases

1. Phase-1: Data collection and loading the data

* At this Phase we collected dataset from [Kaggle](https://www.kaggle.com/datasets/datasnaek/mbti-type/download?datasetVersionNumber=1) ,and loaded into our notebook using pandas.

1. Phase-2: Preprocessing data and importing necessary libraries

* At this Phase we preprocess the data which contains URL’s, Multiple spaces, Non words, Multiple full stops, Stop words etc.
* We will import necessary libraries such as TensorFlow, NumPy, Transformers etc.
* Here we will preprocess Target feature also, i.e. we will assign o and 1 to each class i.e. [I=0, E=1], [N=0, S=1], [F=0, T=1], [J=0, P=1] based on the combination of these values, we will derive Target feature for example, if it has [“INFJ”] then its value will be [0 0 0 0].

1. Phase-3: Splitting data into Train and Test

* At this phase we will split our data into Train and Test using Sklearn , our training size will be 80 % and test size will be 20 %.

1. Phase-4: Passing preprocessed data to BERT Tokenizer

* At this phase we will pass the X\_train and X\_test to BERT Tokenizer to preprocess the text (internally text is converted into numbers).

1. Phase-5: Building and Compilation of Model

* At this phase, we will make use of pretrained BERT model along with transformers both helps in increasing accuracy.

1. Phase-6: Training and Predicting

* At this phase we will train the developed model based on number of EPOCH’s (here I used 15 in order to avoid overfitting) and predict the result.
* The predicted result will be in terms of Scores ranges from 0 to 1, we will choose personality which has score greater than 0.5, and do mapping according to that.

1. **Libraries Used**

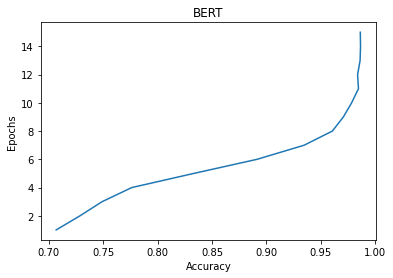
* Pandas - for loading dataset
* NLTK - for text preprocessing
* NumPy - for mathematical operations
* Re - for text filtering
* Tensor flow - for model building
* Scikit Learn - for splitting dataset, required metrics
* Transformers - to load previously trained model
* TensorFlow addons
* Matplotlib - for plotting Graphs and charts

1. **Approach**

* At the beginning, I’ll load my dataset into data frame and then I’ll binarize (One Hot Encoding) to Target variable since it’s a multiclass classification. According to this structure we will map our target feature [I=0, E=1], [N=0, S=1], [F=0, T=1], [J=0, P=1], for example if our target feature is [INFJ] its binarize/One hot Encoded value will be

[0 0 0 0].

* By Exploratory Data Analysis (EDA), we got to know that, our data (Posts) contains text along with URL’s, non-words, multiple spaces/full stops/repeating words, stop words, along with target class names in feature, so we need to remove that.
* At this time preprocess work is done, next we will Split our data into train and test based on split ratio generally in machine learning , we will take 70:30 i.e. we will use 70% of our data to train our model and 30% to test, based on the scenario this will change , if our model is underfitting then we will use 80:20, at worst case we will use 90:10 or if our model is overfitting then we will decrease our training data i.e. 60:40 at worst case will remove some instance/records.
* Here we had used 80:20 as our split ratio, out of 8675 samples, we had use 6940 for training and 1735 for testing.
* From Transforms module we will import pretrained BERT model and Bert Tokenizer which helps in Converting preprocessed text into array of numerical data.
* Then we will initialize our BERT model, and design our model at the output layer we used “Sigmoid” activation function, since it’s a Binary Classification at this point of time. Then we will take Summary of our designed model by “model.summary()”.
* At next step we will compile and fit the model, and save the Weights/Model.
* We will load the Weights file and try to predict it with real word data. By all these steps
* we got an Training accuracy around 98% and Test accuracy around 70% for 15 Epochs.



* Now we will try to predict by passing a real-world data, the below tweet is tweeted by sir “Anand Mahindra”: "A fascinating list that shows how entrepreneurs are flexible & ‘pivot’ when opportunity arises. But it provides a great lesson even for our personal lives: Don’t be afraid of change. You don’t have to stay wedded to what you originally started out to do. Evolution is life".
* Our model Predicted that it’s a [‘E’,‘N’,‘F’,‘J’], internally it works like if we provide data, it will return Confidence/Scores ranges between 0 – 1,if the score is greater than 0.5 ,then it will be marked as 1 else 0. With respect to the given data my model gave result as mentioned below:

-- comment: A fascinating list that shows how entrepreneurs are flexible & ‘pivot’ when opportunity arises. But it provides a great lesson even for our personal lives: Don’t be afraid of change. You don’t have to stay wedded to what you originally started out to do. Evolution is life.

-- personality: ['E', 'N', 'F', 'J']

-- scores: [9.7446334e-01, 3.4945118e-05, 9.7951674e-01, 4.6276513e-01]

Personality is: Skilled Communicator, Caring and Idealistic.

* Based on my prediction we can conclude that, the persons personality is: “IDEALISTIC and SKILLED COMMUNICATOR”.