**HDSC ’22 Capstone Project Presentation: Personality Prediction**

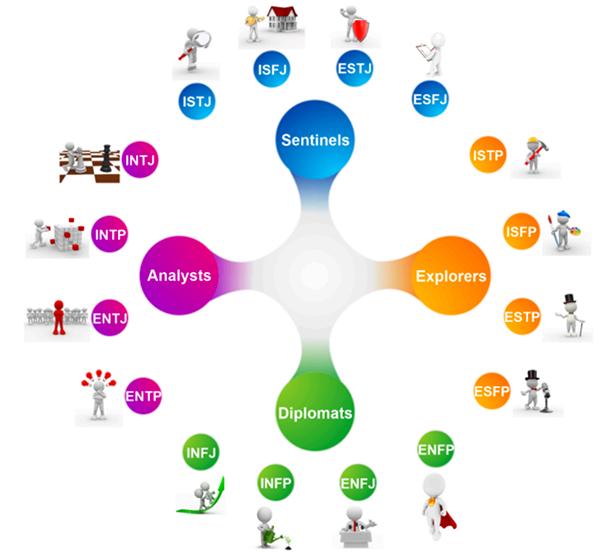
***A Project by Team StandardScaler***

****

**Introduction**

One quick google search about MBTI (Myers Briggs Type Indicator) personality type redirects you to the website where after a set of behavioural questions you are assigned one of the 16 personalities. Some companies have also started using this test to shortlist applicants based on their personality thus aiding them in strategic hiring and getting the right person and thereby increasing productivity in the workplace.

A division of several classes of people into 16 distinct personality types across 4 axis based on the Myers Briggs Type Indicator is illustrated below:



* Introversion (I) – Extroversion (E)
* Intuition (N) – Sensing (S)
* Thinking (T) – Feeling (F)
* Judging (J) – Perceiving (P)

**Dataset Description**

The dataset used for this project was collected through the [Personality Cafe forum](https://www.personalitycafe.com/login/), which includes a lot of people’s MBTI type and also contains content written by them.

The dataset has over 8600 rows of data and on each row we have a person’s type which is his/her 4 letter MBTI code type and also a section of each of the last 50 things posted by them.

**Objective**

The objective of this project is to build a machine learning model that predicts individual personality based on the given dataset.

**Methodology**

**1. Data Cleaning**

It is necessary to clean the text data to make it more understandable and to gain more insights from it. The posts column in the dataset which contains text was cleaned before using the data to create an NLP (Natural Language Processing) model. The text in the data was converted to lowercase characters and links, punctuation marks, other symbols including digits while excluding characters in the text, and unnecessary data was removed from the text using regular expression. Also, emojis and emoticons were replaced with their meanings while stop words which have no significant contribution during prediction of untrained data were removed.

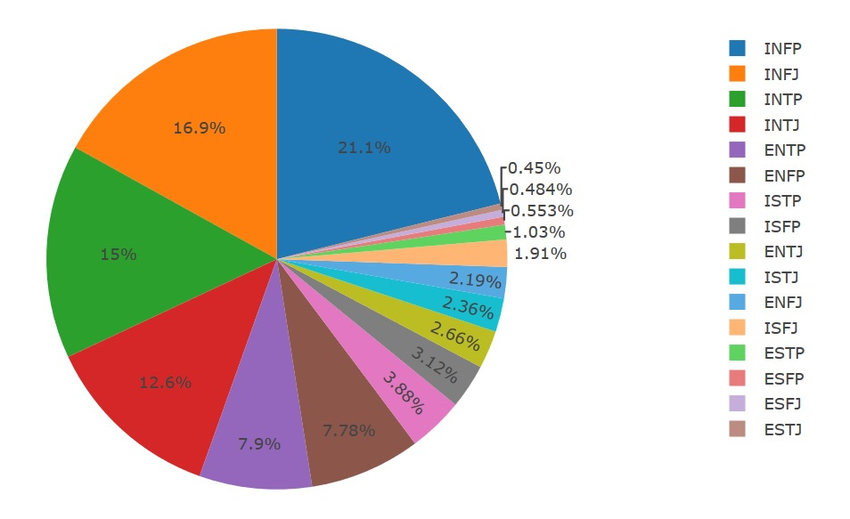
Although, most of the common words in the data were stop words which had a frequency of above 100,000. The posts majorly contain stop words like I, I'm, me, if, so, etc., which is safe to form an assumption that they do not provide much useful information when training our ML model. Tokenization, Stemming, Lemmatization and Vectorization were used to convert these posts into understandable formats.

.

**2. Exploratory Data Analysis**

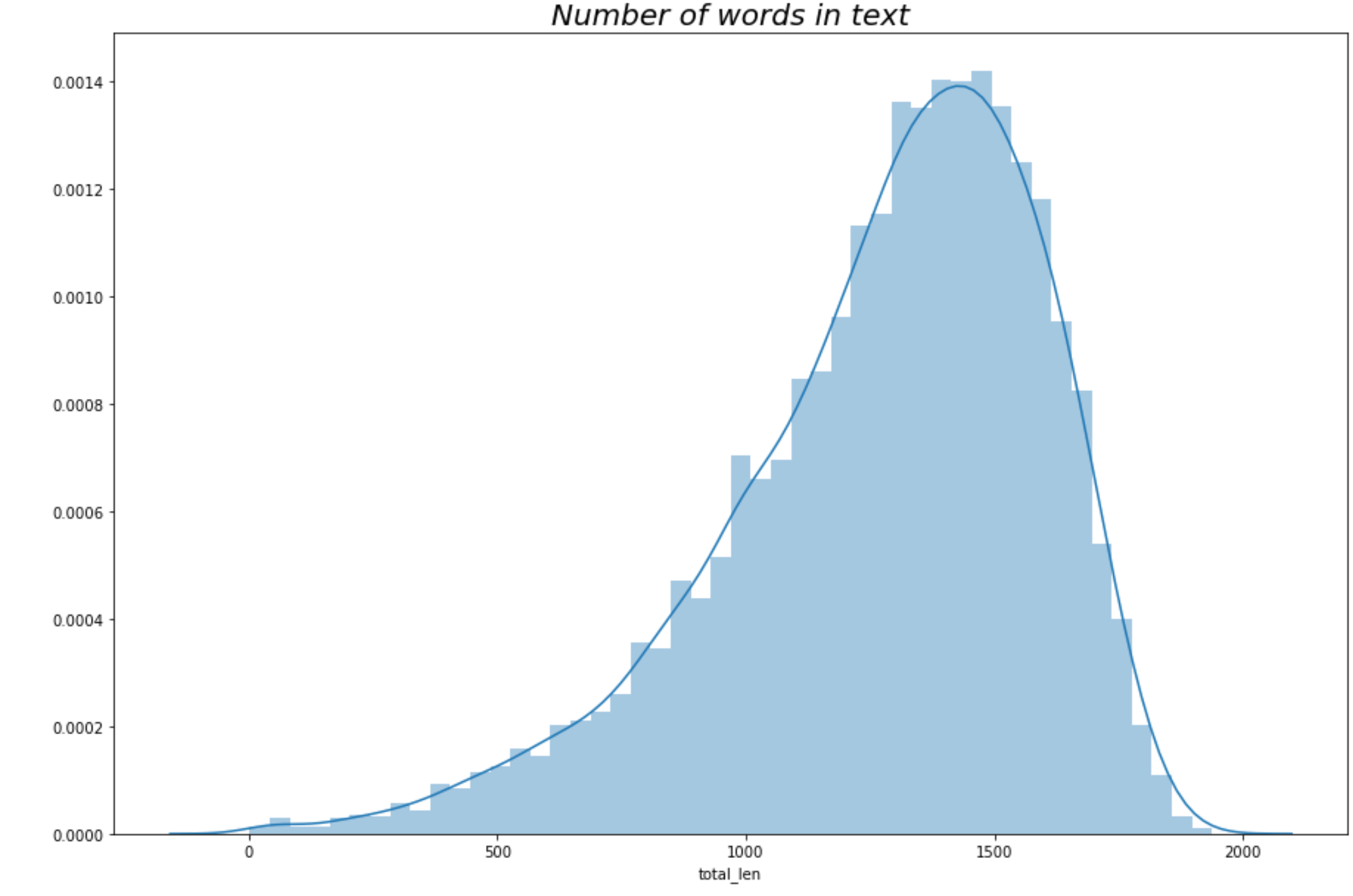
* ***Pie Chart***

A bar plot was used to visualise the dataset, which clearly showed that the dataset is clearly unbalanced throughout the different classes. It was observed that some of the personality types have a lot more data than others, the most common personality is INFP (Introverted Intuition Feeling Perceiving). The use of a pie chart revealed similar results as the box plot with the same personality having a huge chunk in the pie.



* ***Distribution plot***

The distribution plot showed that the most lengthy posts have between 7000 to almost 9000 words. The plot is thus skewed to the left where it has a longer tail.



**3. Modelling**

After preprocessing and feature extraction, the resulting dataset is split into a training and test datasets and XGBoost, Logistic Regression, ADABoost, Random Forest, Bagging, and Gradient Boost were used for the classification of the personality types. Both Logistic Regression and XGBoost performed best on the test data each with an accuracy of 64.58% and 64.20%.

Hyperparameter tuning was carried out on both of these models and the accuracy of Logistic Regression rose to 67.15% with an F1\_macro score of 57.50% while that of XGBoost 66.31% and an F1\_macro: 51.48%

**4. Conclusion**

From the vectorization, words were encoded into numbers and this created a lot of features that needed to be used during the training of the models that were used. This required a lot of computing power and only a few of the hyperparameters were tuned.

From the results, Logistic Regression performed best on the model and had the best F1 score.

**5. Future Goals**

* More Hyperparameters can be tuned to increase the accuracy of the models.
* Oversampling can be carried out since the classes are imbalanced (This will require more computing power).