INTRODUCTION

This project titled '**MBTI Test Data Analysis**' is based on data analysis performed on the very famous Myers-Briggs Type Indicator, an introspective questionnaire that indicates various psychological perspectives in which people view the world and interact with it.

This test includes casting forth situational questions to people who respond to them in degrees of agreement, disagreement, or neutrality in their case. The indicator then compiles the results and generates a four-lettered word, a personality type.

The MBTI test works on eight measuring attributes:

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

The combination of these alphabets generates 16 unique personality types, each associated with character traits that can define an individual's personality.

The possible sixteen personality types are:

ISTJ	ISTP	ISFJ	ISFP
INTJ	INTP	INFJ	INFP
ESTJ	ESTP	ESFJ	ESFP
ENTJ	ENTP	ENFJ	ENFP

In this project, a dataset has been extracted from Kaggle. The dataset consists of data extracted from Twitter. It consists of 8765 user entries with their last fifty tweets from their forum 'PersonalityCafe' and the personality type they think they have.

To generate the results put forth in the conclusion segment of this report, techniques such as Natural Language Processing, Opinion Mining etc. have been used alongside my existing knowledge of Python libraries such as NumPy, Matplotlib, Pandas and Seaborn.

The project has a segment of Aspect-Based Sentiment Analysis within it. That has been performed using 'SentimentIntensityAnalyzer' from NLTK (Natural Language ToolKit).

The operations performed on the data can be categorised into broad categories:

1. Data Cleaning and Organization
2. Tokenizing the Data
3. Sentiment Analysis
4. Generation of Hypothesis

SCOPE OF THE STUDY

The four dichotomies i.e., extraversion v/s introversion, sensing v/s intuition, thinking v/s feeling and judging v/s perceiving define an individual's personality.

They can be explained as follows:

1. **Extraversion v/s Introversion**: the way in which people draw their energy.
   Extroverts are social while introverts tend to be quieter.

2. **Sensing v/s Intuition**: the way in which people gather information.
   Sensors gather information from their environment using their 5 senses while intuitivists look for the wider scope.

3. **Thinking v/s Feeling**: the way in which people make decisions.
   Thinkers tend to make decisions based on logic while feelers tend to make decisions based on emotions and values.

4. **Judging v/s Perceiving**: the way in which people organize.
   Judgers prefer details while perceivers are open and tend to be flexible.

Problem Statement:

We seek to analyse data about the personality types and draw insightful conclusions about them based off the textual information of the last fifty posts that the people have shared on a social media platform.

PROJECT WALK-THROUGH

1. **Data Cleaning and Organization:**

We go through the data and take note of recurring data, missing/null (NaN) values, entries containing links etc. and filter them out from the useful data.

This step includes importing the initial functionalities required for the set-up for the project and some libraries or modules required for the future purposes.

The shape and structure of the data-frame is analysed. The '|||' separated values are listed into a textual format. Posts are segregated according to their personality types to create a 'Bag of Words' model.

The clean and organized data is thus, ready to be operated upon.

1. **Tokenization:**

This segment includes the usage of Natural Language ToolKit (NLTK) that contains tools for working with human language data. It is widely used for tasks such as:

Tokenization: breaking text into tokens (words / phrases / symbols / other meaningful elements).

Stemming: reducing a word into its base or root form.

Tagging: assigning a part of speech to each word in a sentence.

Parsing: analysing the grammatical structure of a sentence.

Semantic reasoning: process of understanding the meaning of text based on its context.

In this project, we create a function to tokenise the words.

The generated words, called tokens are then lemmatised using NLTK's lemmatizer. Out of the several widely used lemmatization algorithms, we use the 'WordNetLemmatizer'.

Lemmatization: a process of grouping together the different inflected forms of a word so that they can be used as a single item.

2. **Sentiment Analysis:**

The sentiment analysis has been done using NLTK's 'SentimentIntensityAnalyzer'.

It analysed token text data into four categories: compound, negative, positive, and neutral.

There are many ways to extract the sentiment scores from the text data:

(i) Wordcount method
(ii) Usage of ML models such as 'NaiveBayesClassifier'
(iii) Usage of Lexicon-based models such as VADER (Valence Aware Dictionary and Sentiment Reasoner)

Here, VADER has been used. The ‘polarity_scores’ method is used as a metric for the sentiment scores. The values of sentiment scores lie between -1 and 1.

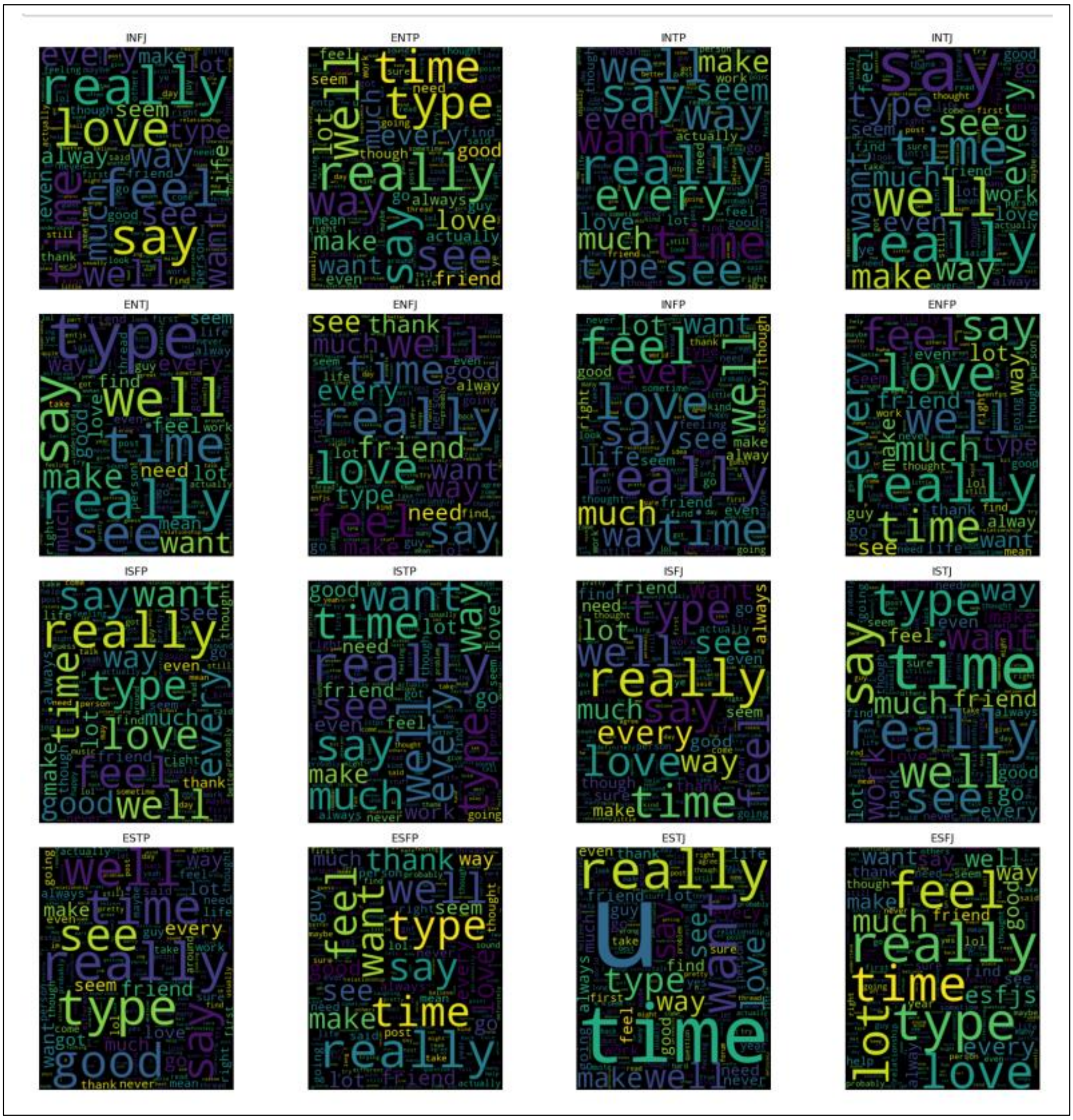
As it was found that there was a class imbalance, the data was further subjugated into eight balanced classes: ‘extrovert’, ‘introvert’, ‘sensing’, ‘intuitive’, ‘thinking’, ‘feeling’, ‘judging’, and ‘perceiving’.

3. Visualisations:

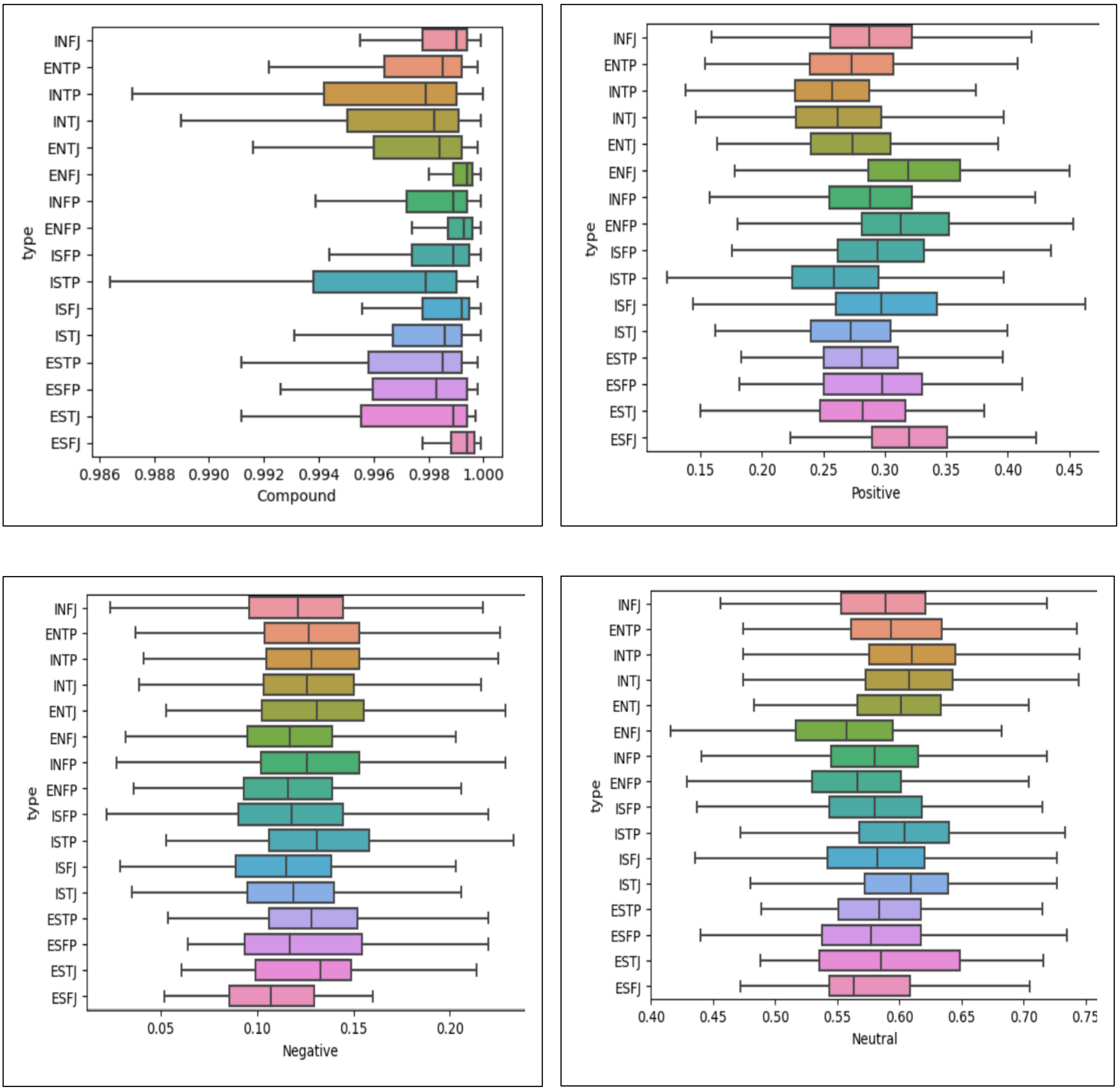
(i) Making Word Clouds:

To delve deeper into the analysis, a word cloud has been prepared. The word cloud contains the frequently used words by each unique MBTI type in their posts.

The words initially displayed in large fonts were: ‘one’, ‘thing’, ‘think’, ‘know’. Since, they were not descriptive of the qualities and were recurrent, they were removed.



‘Compound’, ‘Positive’, ‘Negative’ and ‘Neutral’ for each of the 16 personality types.



POSTMORTEM

1. There is a high imbalance in the provided data with the greatest number of people identifying with ‘INFJ’ and ‘INFP’ personality types.

2. The above occurrence was the reason for some approximate classification for the ‘Sensitive’ and ‘Intuitive’ types.

3. The personality type ‘INFP’ has the greatest number of whilst ‘ESTJ’ has the least number of posts.

4. Information such as recurring words, images, links, abbreviations, stop words etc. was removed during the data clean-up.

5. The highest number of posts had an average word count of about 27.5 words.

6. The variance of ‘Compound’ sentiment is highest for ISTP (0.993-0.999) and INTP (0.994-0.999). It is lowest for ESFJ (0.9991-0.9998).

7. The variance of ‘Positive’, ‘Negative’ and ‘Neutral’ sentiments is almost similar for all the personality types respectively.

8. Top words for the personality types can be found in the word cloud:
   e.g.: INFJ: ‘time’, ‘feel’, ‘love’, ‘say’.
   Which match the general depictive traits of INFJ people. They are thinkers, intuitive, highly inquisitive, feelers, speak selectively and have high standards in relationships. Similar conclusions can be drawn for the other personality types.

Remarks:

Since the data is of ‘original’ nature i.e., extracted from a forum existing in reality, it has certain shortcomings. It doesn’t follow the global data trends, is not uniform and restrictive in terms of representation of different socio-cultural backgrounds.

Using the data on a much larger scale such as that of Facebook and Instagram would lead to generation of much more interesting trends and if trained using ML models, could do better predictive analysis with an application in making personality indicators for text-based data.

As an individual progresses in life, the evolution of his/her personality types would be interesting to analyse in the future.

BIBLIOGRAPHY

Dataset taken from Kaggle:

https://www.kaggle.com/datasets/datasnaek/mbti-type

The MBTI type can be checked from this website:

www.16personalities.com