Personality Type Prediction Using Myers-Briggs Type Indicator from Social Media Posts

Megico Mejora Maria Nayagam
200910992
Mr. Bilal Hassan
MSc Big Data Science
Queen Mary University of London

Abstract—Understanding and predicting personality type can help to understand the preferences of people and how or why they might be different. Understanding the preferences increases flexibility and adaptability that needed for healthy development depending on the individuals personal and professional life. Personality type prediction is vastly useful for recognizing how we lead, influence, communicate, collaborate, negotiate business and manage stress. There is significant interest in personality type prediction in the field of psychology and in corporate sector. This study aims to build a classifier that can predict the personality of the author by analysing the texts extracted from his/her social media posts. The model used for personality analysis is Myers-Briggs Type Indicator (MBTI) method. MBTI model is the most popular and is widely used owing to the fact that hundreds of studies over the past 40 years have proven this instrument to be both valid and reliable. Furthermore, this paper proposes a deep learning architecture combined with Bidirectional Encoder Representations from Transformers (BERT), a pre-trained language representation model used to extract contextualized word embeddings from textual data for author's personality detection. This project uses multiple social media data sources namely PersonalityCafe and Reddit and produces a predictive model for each trait using Natural Language Processing (NLP) and bidirectional context feature methods. The experiment is related to research with four binary classifiers that classifies MBTI type classification.

***Keywords*—MBTI, personality prediction, social media posts, binary classification, Deep learning, Natural Language Processing, BERT.**

I. INTRODUCTION

There have been many attempts in the field of psychology to define a reliable model for personality yet there is none to precisely define personality. Hence psychologist will benefit from a more reliable model. Some of the most popular models are Myers-Briggs Type Indicator (MBTI), Big Five and DISC. In this project I intend to build four binary classifiers using deep learning-based approach combined with BERT language model by taking texts from social media posts as inputs and give a prediction of MBTI personality type of the author of the text as output. A successful implementation of such a classifier would demonstrate a strong linguistic basis for MBTI. The datasets I propose to use in this study are the MBTI dataset collected from PersonalityCafe forum available in Kaggle (PersonalityCafe platform) and MBTI9K- a subset of Corpus of the Reddit comments and posts labeled with MBTI personality types (Matej Gjurković et al).

The Myers-Briggs Type Indicator (MBTI) is a widely used personality classification tool that attempts to identify behavioural and cognitive patterns in people. The theory that people's seemingly unpredictable variation in behaviour and thought patterns can actually be interpreted as consistent and orderly became increasingly popular in the early and mid-1900s. This idea was pioneered by Carl Jung, a renowned psychologist and student of Sigmund Freud. In his book, Psychological Types (Carl Jung), Jung designed a psychological typology that he believed could reduce people's personalities into a set of classifications, which he believed was key to understanding other people and their motivations. Using this as the backbone of their research, Isabel Briggs Myers and Katharine Briggs iterated on Jung's original theory, formulating a set of hypotheses, which were then used to create the four dichotomies that are currently viewed as the main categories for the test. This allowed for a new wave of personality research since it became easier to study personality in a quantitative manner. The MBTI system is divided along four binary orthogonal personality dimensions, comprising a total of 16 distinct personality types. The dimensions are as follows: Extraversion (E) vs Introversion (I): a measure of how much an individual prefers their outer and inner world. Sensing (S) vs Intuition (N): a measure of how much an individual processes information through five senses versus impressions through patterns. Thinking (T) vs Feeling (F): a measure of preference for objective principles and facts versus weighing the emotional perspectives of others. Lastly, Judging (J) vs Perceiving (P): a

measure of how much an individual prefers a planned and ordered life versus a flexible and spontaneous life (I. B. Myers).

II. PROBLEM DEFINITION

The objective of this work is to analyze the social media posts of the author and create a personality profile by scoring the person's personality using the MBTI four scales. This personality profile helps to raise awareness for the individuals to own their preference and recognize where the other styles could add value and how to achieve this.

III. RELATED WORKS

Current research on predicting MBTI personality type from textual data is sparse. However, a few works have been done in this area in the recent past. Zeeshan Mushtaq et al in their work propose a way to analyse the user's data posted on social media, combined two existing machine learning algorithms, such as K-Means Clustering and Gradient Boosting, in order to predict user personality type. Moreover, this research helps to analyse the empirical relation between the user's data posted on social media and the user's personality. This study shows an average accuracy of 86.3%. Muhammad Nurfauzi Sahono et al in their work classified human personalities based on the MBTI method that focused on Extrovert and Introvert class, seen from their tweets. Through this work they put forth that such a classifier would enable humans to better understand and improve themselves by recognizing their weaknesses and strengths. The study used SVM classifier and achieved an accuracy of 84.07%.

Further, the behaviour differences from one person to other was analysed using Extreme Gradient Boosting by Mohammad Hossein Amirhosseini and Hassan Kazemian. The study has compared the performance of Extreme Gradient Boosting with the recurrent neural networks such as RNN, GRU, LSTM, and Bidirectional LSTM by using the same dataset and same pre-processing techniques. The obtained accuracy was higher than the recurrent neural networks analysed from the previous research. The MBTI is a famous personality revealing way designed to identify a person's personality type, strength, and preferences (Z. Mushtaq et al). Bharadwaj et.al are of the view that user's data such as essays, posts, statuses, and blogs have been analyzed to find out the user's personality by using different machine learning algorithms such as Naive Bayes, SVM and Neural networks with an accuracy of 88%.

Recent study by Ninoslav Cerkez et al, compared LSTM model with CNN model. Both the models were trained with CM and CECI. CNN model outperformed (67%) the LSTM model for multiclass classification (23%).

All these researches have used only one dataset that is available from the Kaggle platform collected from PersonalityCafe Forum. Also, mostly only machine learning approaches were used to predict author's personality and only a few studies have proposed neural networks using MBTI multiclass classification.

Compared to all the above-mentioned approaches, this paper uses data from multiple data source namely PersonalityCafe and Reddit social media platforms. Also, this study aims to enhance the research scenario by using the state-of-the-art BERT language model combined with NLP features and Deep Learning architecture. Four binary BERT classifier models were built that classifies text as either Introvert/Extrovert traits, Sensing/INtuition, Feeling/Thinking and Judging/Perceiving.

IV. METHODOLOGY

A. The implementation of this research project is carried out in five main steps:

1. Data collection.
2. Converting the dataset into four binary datasets.
3. Text mining to transform the unstructured input text into a structured data by parsing and cleaning the data by removing unimportant elements.
4. Building the model.
5. Evaluating the model.

B. Data

This project uses two datasets:

1. The first dataset used is publicly available MBTI dataset from Kaggle. This data has been collected from the PersonalityCafe forum. This dataset contains 8675 rows of data. In this dataset each row is a person's MBTI personality type, a combination of four labels and a section of fifty posts obtained from the individual's social media. Each post has been separated by three pipe characters.

2. The second dataset used is MBTI9K – Corpus of the Reddit comments and posts labeled with MBTI personality types. It has 9149 rows and 7 columns from which only around 5000 rows and 2 columns namely 'comments' and 'type' were used for this research. This dataset was acquired on request from Matej Gjurković and Jan Šnajder. MBTI9K dataset is a subset of dataset containing the comments of authors who contributed with more than 1000 words and comments of each user annotated with the MBTI type of the author from the Reddit social media platform.

C. *Proportionality of the Dataset*

Distribution of the MBTI personality types for the two datasets are shown in “Fig 1, 2”.

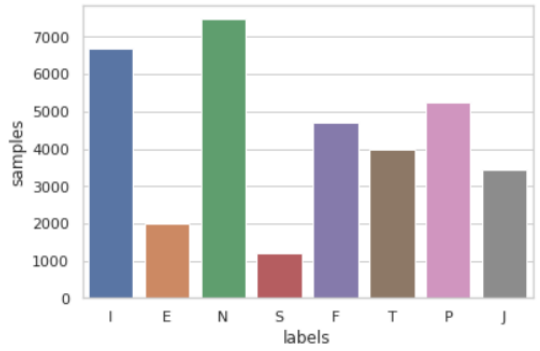

Fig.1. Distribution of classes in Personality Café MBTI dataset

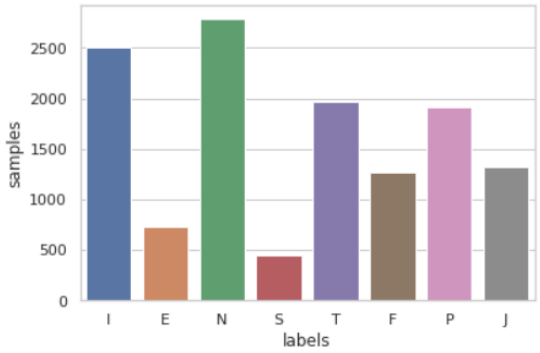

Fig.2. Distribution of classes in MBTI9K Reddit MBTI dataset

Figure 1 and 2 show a non-uniform representation of MBTI personality types in the datasets which is severely imbalanced compared with the actual proportions of MBTI types in the general population.

D. *Splitting dataset into four binary datasets*

This project’s research is predominantly focused to build four binary classifiers that classifies each of the individual’s post as Introvert(I)/Extrovert(E), Sensing(S)/Intuition(N), Thinking(T)/Feeling(F) and Judging(J)/Perceiving(P). From Fig.3 Sample dataset, the column ‘type’ is the combination of MBTI traits labelled for each individual’s posts. In order to obtain four binary classifiers, it is imperative to split each dataset into four binary datasets.

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...

Fig.3. Sample dataset

Using Python string split() method, column ‘type’ is split into four columns. Then created a first separate binary dataset dichotomy of Introvert(I) Vs Extrovert(E) by concatenating first column of ‘type’ with the column ‘posts’. The second binary dataset dichotomy of Sensing(S) Vs Intuition(I) was created by concatenating second column of ‘type’ with the column ‘posts’. The third binary dataset dichotomy of Thinking(T) Vs Feeling(F) was created by concatenating third column of ‘type’ with the column ‘posts’. The final binary dataset dichotomy of Judging(J) Vs Perceiving(P) was created by concatenating the fourth column of ‘type’ with the posts. The sample of partitioned binary datasets of (I/E), (S/N), (T/F) and (J/P) are shown in Fig.4, 5, 6 and 7.

```
The shape of binary dataset (I/E) :  (8675, 2)
```

	types	posts
0	I	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	E	'I'm finding the lack of me in these posts ver...
2	I	'Good one _____ https://www.youtube.com/wat...
3	I	'Dear INTP, I enjoyed our conversation the o...
4	E	'You're fired.|||That's another silly misconce...

Fig. 4 Sample binary dataset of (I/E) MBTI traits

```
The shape of binary dataset (S/N) :  (8675, 2)
```

	types	posts
0	N	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	N	'I'm finding the lack of me in these posts ver...
2	N	'Good one _____ https://www.youtube.com/wat...
3	N	'Dear INTP, I enjoyed our conversation the o...
4	N	'You're fired.|||That's another silly misconce...
...	...	...
8670	S	'https://www.youtube.com/watch?v=t8edHB_h908||...

Fig.5 Sample binary dataset of (S/N) MBTI traits

The shape of binary dataset (T/F) : (8675, 2)

	types	posts
0	F	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	T	'I'm finding the lack of me in these posts ver...
2	T	'Good one _____ https://www.youtube.com/wat...
3	T	'Dear INTP, I enjoyed our conversation the o...
4	T	'You're fired.|||That's another silly misconce...

Fig. 6 Sample binary dataset of (T/F) MBTI traits

The shape of binary dataset (J/P) : (8675, 2)

	types	posts
0	J	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	P	'I'm finding the lack of me in these posts ver...
2	P	'Good one _____ https://www.youtube.com/wat...
3	J	'Dear INTP, I enjoyed our conversation the o...
4	J	'You're fired.|||That's another silly misconce...

Fig.7 Sample binary dataset of (J/P) MBTI traits

E. Text Mining using Natural Language Processing Techniques

Text mining uses Natural Language Processing to extract valuable insights from the unstructured input data and transform it into structured data that machines can understand.

a) *Preprocessing:*

As the dataset was collected from social media, it contains lot of insignificant words which may not be a meaningful feature to identify a person's personality and hence needs to be cleaned. The flow of the preprocess for the datasets are illustrated in Fig. 8.

- Removing URL's: The comments contain number of websites. Since the model is to be generalized to English language, the links to the websites are removed.
- Removing punctuation and special characters: Using Regular expression, punctuation and special characters are removed which appear rarely in the sentences to improve the performance of the model.
- Removing numbers: Numbers are removed as it doesn't provide any useful information for the text.
- Expanding a contraction in the sentence such as the use of 'you're' to make it 'you are'.

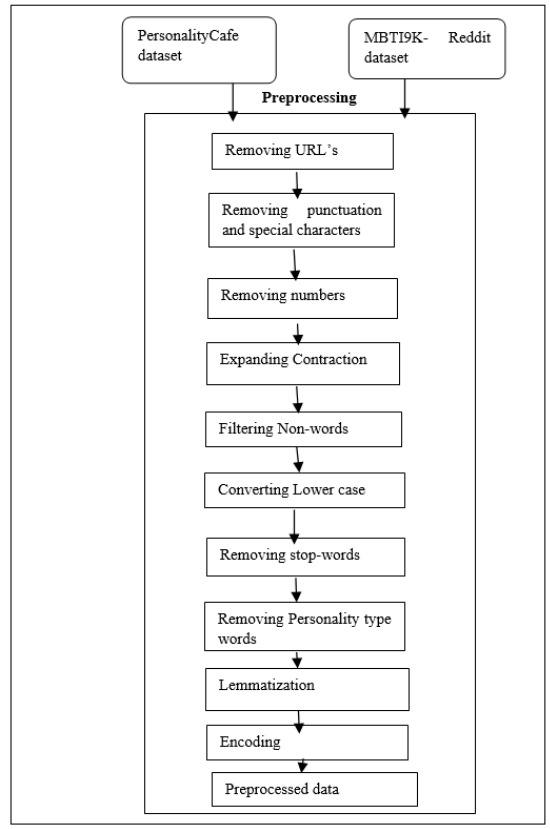

Fig.8. Preprocessing stage

- Filtering non-words: Words without meaning or no occurrence in any text corpus or dictionary are removed using NLTK's word punkt tokenizer.
- Converting each sentence into lower case.
- Removing stop-words: removing the stop-words (e.g., commonly used filter words like 'a', 'the', 'is', etc.,) from the text using Python's NLTK text processing library and giving more focus to the important words.
- Removing Personality type words: Since the datasets used in this project come from the websites intended for explicit discussion of MBTI type personality, it is important to remove the personality types from the posts to get valid model accuracy estimation for unseen data.
- Removing words with length less than 3.
- Lemmatisation: Lemmatisation is done by using WordNetLemmatizer that reduces words to its root word. Lemmatisation is done because when reduced to root word it produces a dictionary meaning word (e.g., words 'played', 'plays', 'playing' all become 'play'). This inflected forms, can be analyzed as a same word carrying one shared meaning and so as to provide more accuracy.

- Encoding: Encoding converts the binary class labels associated with each data into 0/1.

The two datasets were preprocessed using the above explained techniques.

F. Handling Imbalanced dataset

Fig 1 and 2, show the disproportionality present in both the datasets. When classes are imbalanced, classifiers predict the majority classes and the minority classes may be failed to predict. To avoid this, data augmentation is done for the training data using 'nlpaug' library for the MBTI Reddit dataset. The 'nlpaug' library provides Contextual Word Embeddings for BERT language model. It has two actions namely Insertion and Substitution. Insertion is predicted by BERT language model rather than picking one word randomly and Substitution use surrounding words as a feature to predict the target word. For this study, the Substitution action was chosen and generated four augmented binary datasets. Data augmentation doubled the minority class present in the training data.

V. CLASSIFICATION METHODOLOGY

A. Model Architecture

Fig.9. describes the architecture of the BERT model used in this study. BERT is a pre-trained deep bidirectional language representation from unlabelled text by combining the context of each token in sentences from left to right and from right to left on all layers. In this work BERT BASE model is chosen. BERT BASE model has 12 encoder layer, 768 hidden units and 8 attention heads.

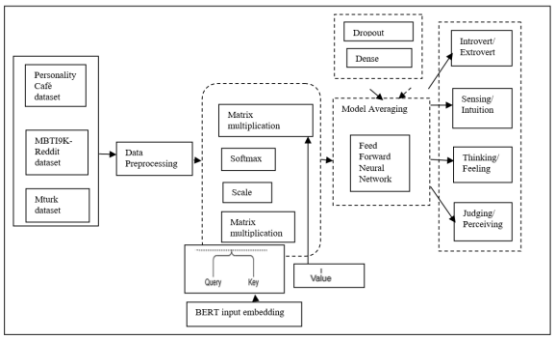

Fig. 9 Model Architecture

B. Model Inputs

BERT has preprocessing and encoding model. Preprocessing model tokenizes each sentence before sending it to the encoder component. The first input token is provided with the special token [CLS] (stands for classification) and a special token is provided at the end of the sentence [SEP] (stands for separator). The purpose of these special tokens is for classification tasks and for separating each input sentence respectively. The tokenization is done by the WordPiece tokenization. BERT tokenizer converts the input words into WordPieces. Each word that is tokenized will be mapped to the WordPiece vocabulary. All the sentences will be encoded by prepending [CLS] token and appending [SEP] token and that forms a single 'input_word_ids'. The 'input_word_ids' has unique id for each input word mapped from the vocabulary. These unique ids are the word embeddings. The other inputs are 'input_mask' and 'input_type'. The 'input_mask' allows the model to differentiate between the content and the padding. The mask has the same shape as the input_word_ids, and contains a 1 anywhere the input_word_ids is not padding. The 'input_type' also has the same shape, but inside the non-padded region, contains a 0 or a 1 indicating which sentence the token is a part of. These three are the inputs to the BERT model that keep flowing up the encoding stack. The BERT encoder model has two sublayers namely Self-Attention and Feed-Forward Neural Network. The encoder receives a list of vectors as input and it processes these inputs by passing it to the self-attention layer and then into the feed-forward neural network. The output from this encoder layer is sent upward to the next encoder layer.

Each of the word embeddings extracted from the pre-trained model will be first passed to the self-attention layer. This layer helps the encoder to look into the other words in the sequence while encoding a specific word. To calculate self-attention for each word, Query vector(q), Key vector(k) and Value vector(v) are created. Next is to score each word of the input sequence against the specific word. The score decides which word should be given more focus in the input sequence for the certain word. The score for each word is calculated by doing dot product of query vector with key vector. Then the scores are scaled down by taking the square root of the dimension of key vector for having stable gradients and the results are passed to the Softmax function. The softmax normalizes the score between 0 and 1. Next step is to multiply each value vector by the softmax score. This step is done to eliminate the irrelevant words which has less score. The final step is to sum up the weighted value vectors. This will be the output for the self-attention layer for the first word. Formula for calculating the self-attention is,

$$Z = \text{softmax}\left[\frac{Q \times K^T}{\sqrt{d_k}}\right] \quad (1)$$

After each sublayer there comes Residual connection and layer normalization. The residual connection is obtained by adding the initial vector to the result with each sublayer feed forward network. Each position outputs a vector of size 768 hidden-size. That vector will be used as the input to the classifier. Two layer neural networks are used as the classifier. The first layer is a Dropout layer, it randomly drops the neurons at each training step. It's added to reduce the overfitting and generalization error of the model. Next, Dense layer with sigmoid activation function is added so that the sigmoid function is applied to the input and the outputs in the interval of (0, 1). The formula to calculate the sigmoid function is,

$$\sigma(x) = 1/(1+\exp(-x)) \quad (2)$$

Finally, the output from the feed forward network will be included in the averaging model function. For loss function, Binary-Cross Entropy is used. The formula for Binary Cross entropy is,

$$\text{Binary cross entropy} = \frac{1}{N}\sum_{i=1}^{N} -(y_i * \log(p_i) + (1-y_i) * \log(p_i)) \quad (3)$$

For fine-turning, AdamW optimizer is used that minimizes the prediction loss and does regularization by weight decay.

C. *Evaluation Metrics*

The trained model is evaluated using F1 measure, Accuracy and AUC-ROC curve metrics.

a) F1 measure:

F1 measure is the harmonic mean of Precision and Recall. This metric is appropriate to use when the dataset has imbalanced data. As the MBTI datasets have non-uniform distribution for each personality traits, it is highly recommended to use this metric to evaluate the model.

Precision: Precision is the measure of the correctly identified positive cases from all the predicted positive cases.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (4)$$

Recall: Recall is the measure of correctly identified positives cases from the actual positive cases.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (5)$$

$$\text{F1 measure} = \frac{2*(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (6)$$

b) Accuracy:

Accuracy is the all correctly identified cases.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} \quad (7)$$

c) *AUC- ROC Curve*

AUC and ROC curve is used as a performance measurement for the classification problem. Area Under the Curve (AUC) is the measure of separability, it tells how much the model is capable of distinguishing between the classes. The Receiver Operating Characteristic (ROC) is the probability curve and provides a graphical representation of the classifier's performance.

D. Experiment

First, the four binary datasets were preprocessed using NLP preprocessing techniques. Next, splitting the datasets into training, validation and testing dataset. The ratio for training dataset is 70% and the 30% of remaining data was split for validation and testing dataset. Second, the SMALL BERT model was chose and loaded from TensorFlow Hub and fine-tuned. The four classifiers' models were experimented with different number of Epochs (50, 7, 5), learning rate (5e-5, 3e-5, 2e-5 and 1e-5), batch size (32, 16, 8) and also experimented with different optimizers like, Adam, AdamW and RectifiedAdam. Finally, the four models were trained with epochs = 7, learning rate = 3e-5, batch size = 32 and with the AdamW optimizer.

The models were evaluated with the results using multiple metrics such as accuracy, precision, recall, F1 score and AUC score, as the metrics suitable for imbalanced dataset. Also, while comparing the results it's important to focus on F1-score as it measures the balance between the precision and recall.

VI. RESULTS AND DISCUSSIONS

A. MBTI PersonalityCafe dataset

Each of the four models were trained and the following results were obtained. Fig.10. shows the training and validation loss and training and validation accuracy of Introvert/Extrovert model. It is obvious that the model learns poorly and is overfitting as the validation loss keeps increasing and the validation accuracy stopped improving and starts declining after 5 epochs. Fig.11. shows the training and validation loss and training and validation accuracy of Sensing/INtuition model. Fig.12. shows the training and validation loss and training and validation accuracy of Feeling/Thinking model. Fig.13. shows the training and validation loss and training and validation accuracy of Judging/Perceiving model. All these four binary class models show that its overfitting.

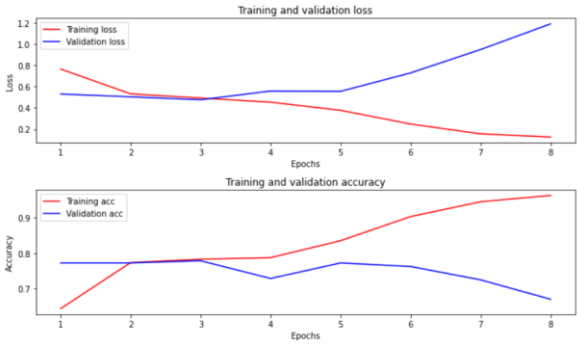

Fig.10. Training and Validation Loss and Training and Validation Accuracy of Introvert/Extrovert model.

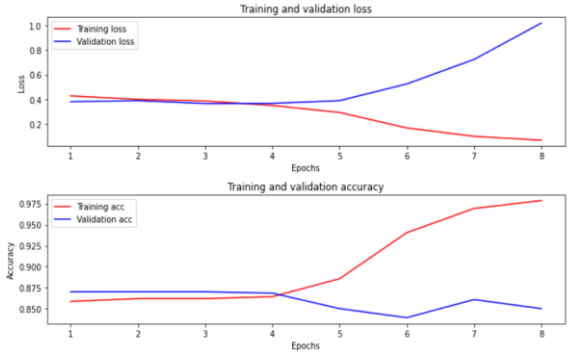

Fig.11. Training and Validation Loss and Training and Validation Accuracy of Sensing/INtuition model.

Table 1. presents the performance measurements of each of the four models. As the PersonalityCafe MBTI dataset is highly imbalanced, it is essential to choose the right metrics. Because of the imbalanced

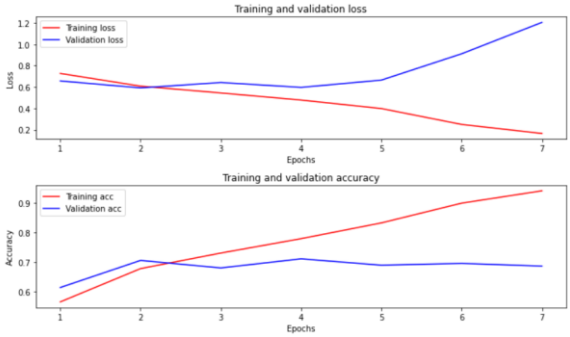

Fig.12. Training and Validation Loss and Taining and Validation Accuracy of Feeling/Thinking model.

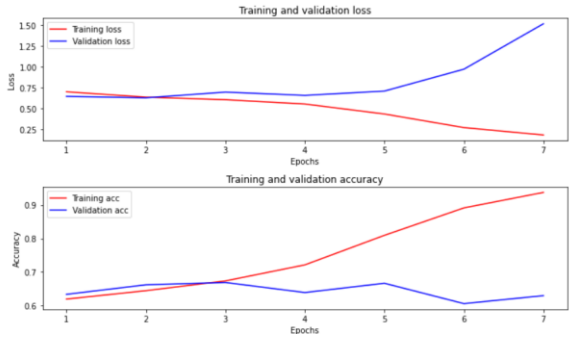

Fig.13. Training and Validation Loss and Training and Validation Accuracy of Judging/Perceiving model.

dataset, using accuracy as a metric is doubtful and misleading (N. Cerkez, B. Vrdoljak et al). This is especially true when performing a binary classification with highly imbalanced datasets. For example, there was a high imbalance in the first two dichotomies. Introverts class have 76.96% of the samples and INtuition class accounts for 86.20% of the samples. Therefore, having high accuracy doesn't validate the model as successful model because high accuracy for the imbalanced data means that the model predicts the majority class and fails to predict the minority class. On the other hand, precision gives the classifier exactness as it gives information on how much to trust the model when it predicts a class as positive. Also, recall measures the completeness of classifiers as it calculates the fraction of true positives and the total number of positively classified class (N. Cerkez, B. Vrdoljak et al). F1-score is more reliable than accuracy because it balances between precision and recall as a weighted average. From the Table 1, the F1-score and the AUC score show that the Introvert/Extrovert model and Feeling/Thinking model have performed good and the other two models have performed fairly well. Fig.14. shows the ROC curve for the four models.

Table 1. Performance Metrics of four binary class models

	Models	Accuracy	Precision	Recall	F1 Score	AUC
0	I-E model	0.765745	0.774834	0.965944	0.859899	0.721344
1	S-N model	0.870968	0.739130	0.178947	0.288136	0.692190
2	F-T model	0.713518	0.694885	0.663300	0.678725	0.780562
3	J-P model	0.651306	0.663395	0.858958	0.748616	0.678020

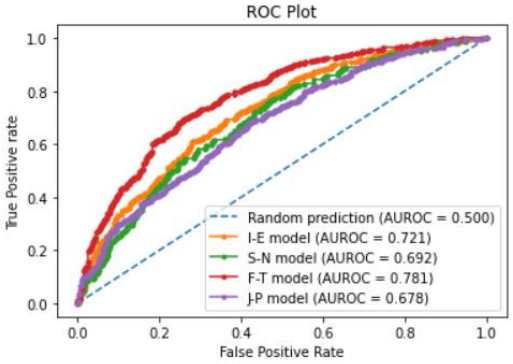

Fig.14. ROC Curve for four binary class models

A. MBTI Reddit data

As this MBTI dataset is also highly imbalanced, the model performed poorly with AUC score of about 50% to 60% for each of the four binary models that can be seen from Fig 15. The models could predict only the majority classes and missed to predict all the minority classes. To enhance the performance measure for this dataset, the data was augmented for the minority class and generated four synthetic datasets. These newly generate datasets were trained and evaluated. The performance metrics of all the four models showed slight improvement. Fig 16 shows the performance of the models after data augmentation.

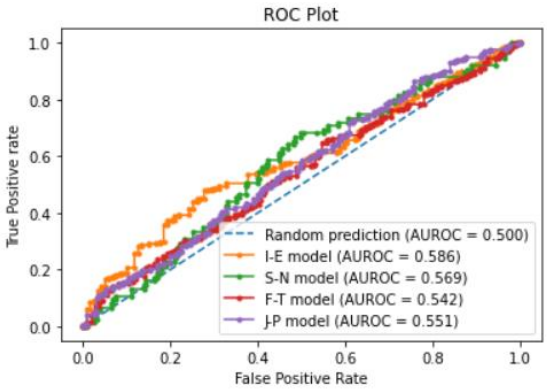

Fig 15. ROC curve of MBTI Reddit data before data augmentation

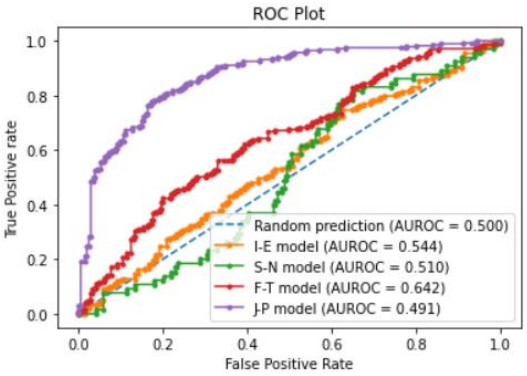

Fig 16. MBTI Reddit data after data augmentation

VII. CONCLUSION

The aim of this project is to build a personality profile using the social media text posted by the user. In this study, the BERT language representation model combined with NLP and deep learning architecture approach is used to predict the MBTI type personality from a user's data and extracting valuable insights from it. Also, the results are evaluated using different metrics. Because of the highly imbalanced of dataset, the BERT could attain only an optimal performance. By having the balanced dataset and increasing the number of samples, the performance of the model can be improved.

VIII. FUTURE WORK

Further work can be done to further improve the proposed personality prediction model by utilizing the larger and balanced datasets. Also, implementing other pre-trained models like ALBERT which is A Lite BERT for Self-supervised Learning of Language Representation and DistilBERT and evaluate the performance of the personality prediction model. Furthermore, deploying the personality prediction tool as a web service and make it available to all the people at free of cost.

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my supervisor, Mr. Bilal Hassan. His professional guidance and constant encouragement helped me in completing the project successfully. His guidance in implementing Deep Learning techniques is highly appreciated.

REFERENCES

MBTI Kaggle dataset. https://www.kaggle.com/datasnaek/mbti-type

MBTI9K dataset. Matej Gjurković and Jan Šnajder (2018). **Reddit: A Gold Mine for Personality Prediction** . Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media.

MTurk MBTI dataset.] Stajner, S., Yenikent, S. 2021. Why Is MBTI Personality Detection from Texts a Difficult Task? In Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 3580-3589.

I. B. Myers and P. B. Myers, Gifts Differing. Palo Alto, CA, USA: Consulting Psychologists Press, 1990.

] C. G. Jung, Psychological Types The Collected Works of C. G. Jung, vol. 6.Princeton, NJ, USA: Princeton Univ. Press, 1976.

M. N. Sahono *et al*., "Extrovert and Introvert Classification based on Myers-Briggs Type Indicator(MBTI) using Support Vector Machine (SVM)," *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2020, pp. 572-577

M. B. Yel, S. Sfenrianto and E. T. Meistiawan, "An Adaptive e-Leaming Model Based on Myers-Briggs Type Indicator(MBTI)," *2018 Third International Conference on Informatics and Computing (ICIC)*, 2018, pp. 1-4,

S. Maniar, K. Patil, B. Rao and R. Shankarmani, "Depression Detection from Tweets Along with Clinical Tests," *2021 International Conference on Intelligent Technologies (CONIT)*, 2021, pp. 1-6, doi: 10.1109/CONIT51480.2021.9498486.

T. Pradhan, R. Bhansali, D. Chandnani and A. Pangaonkar, "Analysis of Personality Traits using Natural Language Processing and Deep Learning," *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2020, pp. 457-461, doi: 10.1109/ICIRCA48905.2020.9183090.

S. Bharadwaj, S. Sridhar, R. Choudhary and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 1076-1082, doi: 10.1109/ICACCI.2018.8554828.

N. Cerkez, B. Vrdoljak and S. Skansi, "A Method for MBTI Classification based on Impact of Class Components," in IEEE Access, doi: 10.1109/ACCESS.2021.3121137.

Christian, H., Suhartono, D., Chowanda, A. *et al.* Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *J Big Data* **8,** 68 (2021). https://doi.org/10.1186/s40537-021-00459-1

Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria and S. Eetemadi, "Bottom-Up and Top-Down: Predicting Personality with Psycholinguistic and Language Model Features," 2020 IEEE International Conference on Data Mining (ICDM), 2020, pp. 1184-1189, doi: 10.1109/ICDM50108.2020.00146.

Bin Xu, Qiaoqiao Zhang, Kening Gao, Ge Yu, Zhaowu Zhang, and Yidi Du. 2021. Recognition Of Learners' Personality Traits For Software Engineering Education. In ACM Turing Award Celebration Conference - China (ACM TURC 2021) (ACM TURC 2021). Association for Computing Machinery, New York, NY, USA, 1–7. DOI:https://doi.org/10.1145/3472634.3472636

R. Moraes, L. L. Pinto, M. Pilankar and P. Rane, "Personality Assessment Using Social Media for Hiring Candidates," 2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA), 2020, pp. 192-197, doi: 10.1109/CSCITA47329.2020.9137818.

The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), https://jalammar.github.io/illustrated-bert/

The Illustrated Transformer, https://jalammar.github.io/illustrated-transformer/

Understanding AUC - ROC Curve-https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Dealing with Data Imbalance in Text Classification-Cristian Padurariua,b, Mihaela Elena Breabana,b*

MSc Project - Reflective Essay

Project Title:	Personality Type Prediction Using Myers-Briggs Type Indicator from Social Media Posts
Student Name:	**Megico Mejora Maria Nayagam**
Student Number:	**200910992**
Supervisor Name:	Mr.Bilal Hassan
Programme of Study:	MSc Big data Science

1. Analysis of strengths / weaknesses

1.1 Strengths

- The main strength of using MBTI is to create awareness for the individuals to own their preference and identify where the other styles could add more value to improve their personality and how to achieve this. Understanding one's preferences increases the flexibility and adaptability that is needed for healthy development depending on the individuals personal and professional life.

- Predicting an individual's personality from the text extracted from their social media post written by the user. Data collected from multiple social media platforms.

- Using Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art language model for Natural Language Processing (NLP).

- Proposed a framework with BERT model combined with NLP features and Deep Learning architecture approach to predict the personality of the author. The method used for personality analysis is Myers-Briggs Type Indicator (MBTI). This approach provides the model more predictive power.

1.2 Weaknesses

- The data collected from different data sources are insufficient.

- The main weakness in this study is that the datasets are highly imbalanced.

- Due to the non-uniform distribution of the data, predicting an outcome became difficult when there was not enough data to learn from. So, the classifiers were biased towards the majority class. Hence, the model is able to produce only optimal performance.

- Gathering more training data process was too difficult and time consuming.

- As the problem was binary classification, four BERT models were built. Running four BERT models took longer time for execution.

2. Presentation of possibilities for future work

It is recommended that to improve the performance of the proposed personality predicting model in the future, it can be experimented in the following directions:

- Training the model with the larger MBTI training data with more balanced dataset.
- If collecting balanced datasets becomes difficult due to the problem of data scarcity, then can be experimented with different data resampling techniques and implement that approach to enhance the performance of the model.
- Extracting different features from the text would be effective.
- Analysing and comparing different Machine Learning and Deep Learning algorithms for text classification NLP problem would help to achieve best model performance and also, the model would be able to predict the classes without any misinterpretation.
- After obtaining the highest performance personality prediction model, it is highly desirable to build a personality prediction application and deploy the model in the cloud service and make it available for everyone to access so that individuals can use this service and predict their personality.

3. Critical analysis of the relationship between theory and practical work produced

From the preliminary research and literature survey, I understood that most of the studies have proposed only Machine Learning approaches to predict MBTI type personality of an individual as a multiclass classification problem. As I was highly interested in learning Deep Learning, I took this opportunity to learn and implement new methods and techniques in this project. I started my study with an idea of developing a BERT model, a state-of-art language model for NLP combined with Deep Learning architecture. I learned the basics of Deep Learning and TensorFlow for BERT text classification method to effectively build the personality prediction model.

Initially, I struggled to gather data, as only one MBTI dataset was publicly available. Later, collected one more dataset on requests from Reddit. The MBTI Reddit dataset had lots of unannotated samples and erroneous lines. It took considerable time to understand the dataset and cleaning the same. Initially, binary classification approach was formulated to classify the text based on the MBTI method. Hence, four binary datasets were created from one dataset by splitting the combination of personality traits present in the MBTI 'type' column. The main goal of this project is to create four binary BERT models that predicts the MBTI traits from a text and scores the probability as either 'Introvert' or 'Extrovert' from the first model, 'Sensing' or 'INtuition' from the second model, 'Feeling' or 'Thinking' from the third model and 'Judging' or 'Perceiving' from the fourth model.

During the period of developing the model, I found it difficult to effectively continue the project because of the long time taken by the BERT model during the phase of training the data. As I was in the beginning stage of learning Deep Learning architecture methods, I was

unaware of the GPU used for training the deep learning models. Later, I leveraged the GPU to train my four BERT models. But still it took a significant time to train data. Furthermore, fine tuning the model by tuning the three main hyperparameters namely epochs, batch size and learning rate was difficult. While fine tuning the hyperparameter epochs for 100 and 50, due to the GPU usage limit, the notebook stopped working and had to restart again which took significant amount of time to complete. Also, experimented by tuning hyperparameters for epochs = 100, 50, 7, 5, learning rate = 1e-5, 2e-5, 3e-5 and 5e-5, batch size = 32, 16 and tried with different optimizer = Adam, AdamW and Rectified Adam.

Due to the highly disproportional number of samples present in the datasets, I faced the problem of overfitting the data. Because of overfitting, the model learned the patterns in the training data well and struggled with the validation data as the validation loss kept increasing. The model predicted only the majority classes and missed to predict the minority classes. To overcome this overfitting, I decided to do data augmentation for the training data using 'nlpaug' library for one of the datasets collected. This library provides Contextual Word Embeddings for BERT language model. It has two actions namely Insertion and Substitution. Insertion is predicted by BERT language model rather than pick one word randomly and Substitution use surrounding words as a feature to predict the target word. I choose Substitution action and generated four augmented binary datasets. It doubled the minority class present in the training data. In the end, even data augmentation took longer execution time to complete than I expected. Finally, I concatenated both the datasets and trained again. Even after using the augmented data, the model performance didn't increase as the dataset was still highly imbalanced.

In hindsight, I clearly understood that there is gap between the theory and practical work.

4. Awareness of Legal, Social Ethical Issues and Sustainability

There were no human experiments as part of the project and so the types of ethical issues are predominantly all indirect.

A possible sustainability issue of the project will be the compute power used to execute models. The recent emerging concern to the Machine Learning community is the carbon footprint of training large neural networks. It has been estimated that training transformers such as BERT (Devlin et al., 2019), large neural networks developed for multi-task learning from natural language text, can consume nearly sixty times more carbon than that of an average human lifetime (Strubell et al., 2019). Experiments like these to test multiple models are consuming energy more to run, and so have a carbon footprint and should be subject to environmental sustainability ethics. This is much truer for deep learning models that take many days to train, but still it's also applicable to the experiments I ran in the project. Compute power should be minimised by training the models using simple and light neural networks.

5. References

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pretraining of deep bidirectional transformers for language understanding. In NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019. ISBN 9781950737130.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://www.aclweb.org/anthology/P1