

ENEL 645 Final Project: Analyzing MBTI Using Traditional Machine Learning and Deep Learning Models

Fatemeh Ghaffarpour

Dept. of Electrical and Software Engineering
fatemeh.ghaffarpour@ucalgary.ca

Mehrnaz Senobari

Dept. of Electrical and Software Engineering
mehrnaz.senobarivayg@ucalgary.ca

Sahar Hajjarzadeh

Dept. of Electrical and Software Engineering
sahar.hajjarzadeh@ucalgary.ca

Alireza Esmaeili

Dept. of Electrical and Software Engineering
alireza.esmaeili1@ucalgary.ca

Zahra Safari

Dept. of Electrical and Software Engineering
zar.safari@ucalgary.ca

Abstract—Understanding different personality types is important in many areas, such as helping people choose careers or helping companies decide where to place their employees. Traditional methods like the Myers-Briggs Type Indicator (MBTI) surveys and interviews can be slow and not always accurate. This has led researchers to look into using machine learning to predict MBTI personality types. Moreover social media provides a platform where users often share their thoughts and feelings more freely and spontaneously. This distinction makes social media a valuable resource for understanding true personality traits. This research employs a combination of machine learning and deep learning models to classify Social media MBTI types, highlighting the use of various word embedding techniques to enhance model performance. For traditional machine learning, Random Forest, Support Vector Machine (SVM), and Logistic Regression are selected. TF-IDF and Word2Vec are employed as traditional word embedding for the inputs of our traditional classification models. Moreover, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been chosen as our deep learning models, with Bidirectional Encoder Representations from Transformers (BERT) to supply contextual representations of sequenced text data. In this study, Logistic Regression proved to be the superior model by achieving an accuracy rate of 71%. To enhance this model further, we incorporated sentiment analysis as an additional feature, examining the sentiment content within social media posts. This approach aims to enrich our model's capability by considering both the information and sentiment dimensions of online expressions.

Index Terms—Personality Type Prediction, MBTI, Traditional Machine Learning, Deep Learning, Sentiment.

I. INTRODUCTION

UNDERSTANDING the various personality types is crucial in numerous domains, particularly in assisting individuals with their professional endeavours and in the way businesses handle their workforce. Typically, questionnaires and interviews are used to measure personality, but in today's fast-paced environment, these methods are not necessarily

accurate or quick enough. This has prompted research into more sophisticated techniques, such as machine learning, to better understand personality.

Psychological research has placed a lot of emphasis on personality assessment, with instruments like the Big Five personality traits and the Myers-Briggs Type Indicator (MBTI) providing insightful analysis.

Based on four preference dichotomies, the MBTI is a psychological evaluation tool that divides people into 16 different personality types. The MBTI personality framework is based on these dichotomies: Extraversion (E) versus Introversion (I), Sensing (S) versus Intuition (N), Thinking (T) versus Feeling (F), and Judging (J) versus Perceiving (P). [2].

- **Extraversion (E) vs. Introversion (I):** Reflects where an individual draws energy from, either from the external world (E) or internal world (I).
- **Sensing (S) vs. Intuition (N):** Indicates whether a person primarily focuses on the present, concrete information (S) or on possibilities and abstract concepts (N).
- **Thinking (T) vs. Feeling (F):** Determines if decision-making is guided more by objective logic (T) or personal values and emotions (F).
- **Judging (J) vs. Perceiving (P):** Describes whether one prefers a structured (J) or a more flexible, spontaneous approach (P) to life.

As shown in Fig. 1, each combination of these preferences results in a unique personality type, such as INTJ or ESFP, with its own set of characteristics and tendencies. The interplay of these dichotomies explains the complexity of human behaviour and provides insights into personal growth and interpersonal relations [3].

Traditional methods of personality assessment, such as questionnaires, often fall short in real-time analysis, leading to gaps in understanding individuals' personalities. Advance-

ESTJ Ambitious Adventurer	ESTP Competitive Doer	ESFP People Entertainer	ESFJ Romantic Adventurer
ISTJ Practical Leader	ISTP Traditional Advisor	ISFP Everyday Artist	ISFJ Friendly Neighbour
INTJ Innovative Visionary	INTP Creative Scientist	INFP Artistic Dreamer	INFJ Sage Mentor
ENTJ Hardworking Visionary	ENTP Inventive Innovator	ENFP Dream Seeker	ENFJ People Visionary

Fig. 1. 16 different personality types under the Myers-Briggs Type Indicator [3]

ments in machine learning, however, offer the possibility of using text-based analysis for personality prediction, providing a more dynamic and comprehensive approach [1].

In this study, different approaches was adopted to build the model, combining traditional machine learning algorithms with advanced deep learning methods. The data preprocessing phase, executed using Python, involved a series of steps to refine the dataset: missing values were removed, textual content was standardized to lowercase, and non-essential elements like URLs were eliminated. Additionally, the dataset underwent balancing through oversampling and undersampling techniques to address the uneven distribution across personality types. For word embedding, which transforms text into a format interpretable by the models, various methods were explored, including Term Frequency - Inverse Document Frequency (TF-IDF), Word2Vec, and Bidirectional Encoder Representations from Transformers (BERT) [9], each offering distinct advantages in capturing linguistic nuances. The motivation behind our approach involves leveraging the strengths of both traditional machine learning models—Random Forest, Support Vector Machine (SVM), and Logistic Regression—and advanced deep learning techniques, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) [13], to analyze MBTI personality types from social media texts. Random Forest is chosen for its robustness in handling complex patterns within dense datasets and preventing overfitting, SVM for its effectiveness in high-dimensional spaces, and Logistic Regression for its quick, probabilistic insights. On the other hand, LSTM and GRU are selected for their ability to capture long-term dependencies

and subtle nuances in sequential text data.

We are looking to answer these questions:

- How can machine learning algorithms analyze textual content to find individuals' personality types?
- How can the accuracy of machine learning models in personality type prediction be improved?
- How can the existing datasets for the task of personality type prediction become richer?

This study advances the MBTI personality prediction by integrating sentiment analysis, marking a significant step forward in the application of machine learning to personality assessment.

The source code for this project is publicly available on GitHub¹ for further reference and collaboration.

The remainder of this report is organized as follows: Section II provides a comprehensive review of the literature, discussing various approaches for MBTI personality classification, including studies that differentiate between classifying the four dichotomies and the comprehensive 16 personality types. Section III describes the methodology, including dataset details, data preprocessing techniques, data balancing methods, embedding strategies, and the machine learning models developed. Section IV presents the results of our models, offering a comparative analysis of their performance and discussing the implications of these findings. The strengths and limitations of the study are also contemplated here. Finally, Section V concludes the paper with a summary of our contributions to the field of personality prediction, reflections on the study's limitations, and suggestions for future research directions.

II. RELATED WORK

The domain of personality prediction using textual data, particularly through the MBTI, has seen varied methodologies targeting either the four fundamental dichotomies (E/I, S/N, T/F, J/P) or the comprehensive 16 personality types.

Jain et al. [10] pioneered the use of Personality BERT, a transformer-based model fine-tuned on the Kaggle MBTI dataset. This study aimed at classifying the 16 personality types, leveraging the nuanced capabilities of BERT for deep textual analysis. The model achieved a notable F1 score of 0.6945, showcasing its effectiveness in personality classification.

Mushtaq, Ashraf, and Sabahat [5] explored the combination of K-Means Clustering and Gradient Boosting on the PersonalityCafe dataset, specifically focusing on the 16 personality types as indicated by the four MBTI dichotomies: Introversion (I)/Extraversion (E), Intuition (N)–Sensing (S), Feeling (F) - Thinking (T), and Judging (J) – Perceiving (P). Their approach demonstrated the potential of machine learning algorithms in discerning complex personality patterns from textual data, with each classifier after hyper-parameter tuning achieving an accuracy within the 85-90% range. The overall average accuracy across all classifiers was 86.3%, signifying a strong model performance in MBTI personality prediction.

¹https://github.com/mehrseno/ENEL645_FinalProject

Cui and Qi [4] conducted a survey of natural language processing (NLP) and machine learning methods for MBTI personality type prediction, reviewing a range of models and approaches. According to their findings, the best-performing model was a deep learning architecture that achieved a training accuracy of 40% and a test accuracy of 38%, outperforming other methods such as Regularized SVM and Naive Bayes. Their analysis underlines the potential of deep learning in the field of personality prediction, particularly when sophisticated text preprocessing and feature selection are applied.

Ontoum and Chan [1] investigated the use of traditional and deep learning models for personality prediction from text posting styles on the MBTI dataset. Their study, focusing on the 16 personality types, revealed that deep learning methods, particularly Recurrent Neural Networks (RNN) [13], were most effective. The RNN model achieved the highest overall accuracy of 49.75%, outperforming Naive Bayes and Support Vector Machines. This underscores the potential of advanced neural network architectures in accurately capturing the nuances of personality from textual data.

Ryan et al. [2] addressed the challenge of data imbalance in personality prediction by applying the SMOTE technique alongside various machine learning models on the Kaggle MBTI dataset, which contains data on the 16 personality types. Their work highlighted the importance of data preprocessing in achieving reliable predictions. The study demonstrated that the use of SMOTE significantly improved model performance, with Logistic Regression emerging as the best-performing model, achieving an F1 score of 83.37%. This result reinforces the effectiveness of addressing class imbalance in enhancing the predictive accuracy of personality classification models.

Amirhosseini and Kazemian [3] applied various machine learning algorithms for MBTI-based personality prediction, using a dataset from a university's psychology department. While their focus was likely on the 16 personality types, the study highlighted the algorithms' capacity to decode personality traits from textual data. The Extreme Gradient Boosting model was identified as the best-performing model, particularly excelling in the Intuition (N)–Sensing (S) dichotomy with an accuracy of 86.06%. This finding emphasizes the effectiveness of gradient boosting methods in the context of MBTI personality classification.

These studies collectively advance the field of personality prediction, demonstrating the diverse applications of NLP and machine learning. From deep learning to algorithmic combinations, the research progresses towards more nuanced personality assessments, with significant implications for personalized content delivery and social dynamics understanding.

III. MATERIALS AND METHODS

This section discusses the dataset, data preprocessing techniques, data balancing techniques, data embedding, and the developed models.

A. Dataset Description

The dataset is sourced from the PersonalityCafe website, which comprises details about individuals' personalities cate-

gorized according to the MBTI system. This dataset includes 8675 rows and two columns. The first column, labelled "type", is filled across 4 axes (Introversion (I) – Extroversion (E), Intuition (N) – Sensing (S), Thinking (T) – Feeling (F) and Judging (J) – Perceiving (P)). The second column, named "posts", contains texts extracted from people's posts on the PersonalityCafe forum, divided by " ||| " symbol. This dataset is publicly available on Kaggle [6]. Fig. 2 shows an overview of the dataset.

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krwIII...
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.IIIThat's another silly misconce...
...
8670	ISFP	'https://www.youtube.com/watch?v=t8edHB_h908II...
8671	ENFP	'So...if this thread already exists someplace ...
8672	INTP	'So many questions when i do these things. I ...
8673	INFP	'I am very conflicted right now when it comes ...
8674	INFP	'It has been too long since I have been on per...
8675 rows x 2 columns		

Fig. 2. Schematic of the dataset for personality prediction based on MBTI.

B. Exploratory Data Analysis

Before applying the machine learning models for solving the MBTI personality prediction, understanding the distribution of data is important. In Fig. 3, the diagram illustrates the distribution of each personality type in the Kaggle dataset, while Table I presents the number of samples for each personality class. This information indicates the imbalance in the data for each personality type. To address this problem, several methods, such as data augmentation, oversampling and undersampling, can be employed [7]. In the next subsection, the methods for handling this problem will be discussed.

C. Data Preprocessing

Before using the text data to build a machine learning model, it is essential to conduct preprocessing steps in NLP problems. In the first step, the removal of missing values from the dataset should be done. Upon observing the data, no missing values were found. The second step involves converting words to lowercase to ensure consistent and uniform word representation. The third step involves removing irrelevant information from the posts that are not useful for the machine learning model. Such as removing HTTP/HTTPS URLs and URLs starting with 'www.' which can be found in posts. The

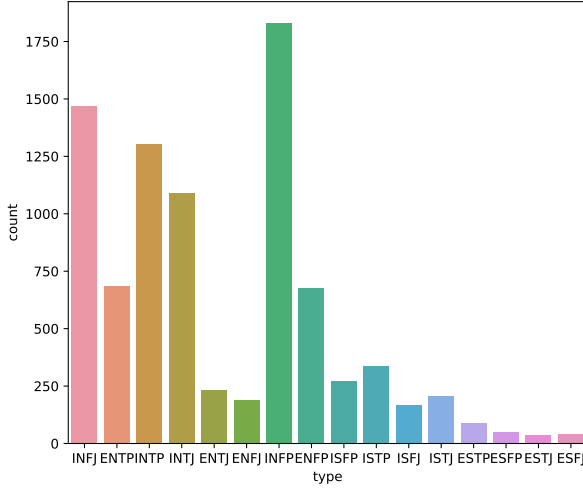


Fig. 3. Plot of distribution of each personality type in the dataset.

TABLE I
NUMBER OF SAMPLES OF DATA FOR EACH PERSONALITY TYPE IN THE DATASET.

Number	Personality Type	Number of Samples
1	INFP	1832
2	INFJ	1470
3	INTP	1304
4	INTJ	1091
5	ENTP	685
6	ENFP	675
7	ISTP	337
8	ISFP	271
9	ENTJ	231
10	ISTJ	205
11	ENFJ	190
12	ISFJ	166
13	ESTP	89
14	ESFP	48
15	ESFJ	42
16	ESTJ	39

fourth step involves replacing any symbols, except 0-9 and a-z, with spaces.

After cleaning the data, methods for balancing the number of samples in each personality type are applied. For personality types with a high number of samples in the dataset, undersampling is applied. For personality types with a low number of samples, oversampling is employed, along with two different types of data augmentation techniques.

In this study, two types of data augmentation are utilized. Firstly, back translation is employed to generate additional samples for MBTI types with a low number of samples. This technique involves translating each post into French and then back into English using the Google Translate API. Secondly, synonym generation techniques are used to create new data by replacing words with their synonyms in posts. This process utilizes the NLTK (Natural Language Toolkit) library, which

incorporates WordNet, a database of the English language that organizes words into sets of synonyms, known as synsets, and delineates semantic relationships between them.

Finally, two copies of the balanced dataset are generated: one for use with traditional machine learning models, and the other for use with deep learning models. For the traditional machine learning model, stop words are removed, and lemmatization is applied. The dataset for the deep learning model will be used with models that are sensitive to context; thus, removing stop words and lemmatization can destroy semantic relations.

D. Data Embedding

One of the important steps in NLP problems is text vectorization, which involves converting text data into meaningful numerical vectors [8]. Based on the models, different types of word embedding techniques are used:

- Word embedding for traditional machine learning models: There are several approaches for text vectorization, including Bag of Words, Word2Vec, and TF-IDF. While Bag of Words counts word occurrences without considering word meaning, Word2Vec and TF-IDF assess word relevance and capture semantic relationships differently. In this research, Word2Vec and TF-IDF vectorization are utilized. For Word2Vec word embedding, Google's pre-trained Word2Vec model using gensim with 300 features is employed. For TF-IDF we used Zipf's law for text analysis, posits that the frequency of a word is inversely proportional to its rank in the frequency table. Essentially, a few words are very common, and many are rare. Guided by this principle, 6800 is chosen as our feature max length because it allowed us to include the most semantically significant words while excluding those that are less common and potentially less informative. This cutoff is strategically selected to ensure that our model remains robust and generalizable, avoiding the noise of less frequent words that could lead. Then Singular Value Decomposition (SVD) is used to narrow down the vast amount of information into more manageable, essential elements. This not only speeds up the computing process but also ensures that our model focuses on the most impactful aspects of the data, leading to more accurate and efficient outcomes.
- Word embedding for deep learning models: Models such as LSTM, GRU, and BERT are all sensitive to the context of words and sequential model. Therefore, contextualized word representations should be used for word vectorization with these models. Thus, the pre-trained Bert embedding with parameters `max_length = 512` and `padding='max_length'` is utilized. The maximum sequence length for Bert Embedding is 512, and this value is used as the post length in the dataset is too long. Additionally, sentences with a length of less than 512 should be padded to 512.

TABLE II
ACCURACY OF DIFFERENT DEEP LEARNING MODEL WITH DIFFERENT
WORD EMBEDDING.

Name of Model/Word Embedding	Bert Embedding
LSTM	22.27%
GRU	23.92%

E. Models

This work aims to perform multi-classification (16 personality types) on text data. In this research, traditional machine learning models are used because they are still efficient for text classification tasks, especially with limited computational resources. Three models using two different word embedding (Word2Vec and TF-IDF) are trained on the dataset: Random Forest [11], Logistic Regression [12], and Support Vector Machine [12].

In addition, three deep learning models using Bert embedding are trained on the dataset: BERT, LSTM, and GRU. These three deep learning models are selected because they are sensitive to the context of text.

We use Gridsearch to find best parameters for traditional machine learning classifier. Our Logistic Regression classifier has 'max_iter'= 100. Also, The random forest classifier has max_depth=10. For the SVM classifier we have Best Parameters = 'C': 10, 'gamma': 'scale', 'kernel': 'linear'.

For Deep learning models, Different experimental setup such as learning rate, optimizer and the number of neurons are tried and selected the best combination. we utilizes a GRU and lstm neural network with specific parameters chosen to optimize the classification of text data into distinct categories. Both our models share the same configuration: a GRU/LSTM layer with 64 units to balance learning capacity and computational efficiency, L2 regularization (0.01) and dropout (0.3) to prevent overfitting, and a dense layer with 16 units and softmax activation for multi-class classification. We employ the Adam optimizer with a learning rate of 0.001 for its adaptive learning capabilities, and sparse_categorical_crossentropy as our loss function, suitable for integer classes.

IV. RESULTS AND DISCUSSION

A. Models Results

This section discusses the accuracy of each model with different word embedding methods.

- Table II shows the accuracy of each deep learning model with different word embedding methods.
- Table III shows the accuracy of each traditional machine learning model with different word embedding methods.

As shown in Table III, the Logistic Regression model achieves an accuracy of 71.11%, performing better than other models for the personality type classification task using traditional machine learning methods. Fig. 4 displays the confusion matrix of the Logistic Regression model using TF-IDF on the test dataset and Table IV shows the classification report of test data for this model.

TABLE III
ACCURACY OF DIFFERENT TRADITIONAL MACHINE LEARNING MODEL
WITH DIFFERENT WORD EMBEDDING.

Name of Model/Word Embedding	TF-IDF	Word2Vec
Random Forest	61.13%	29.3%
Support Vector Machine	67%	34%
Logistic Regression	71.11%	31.15%

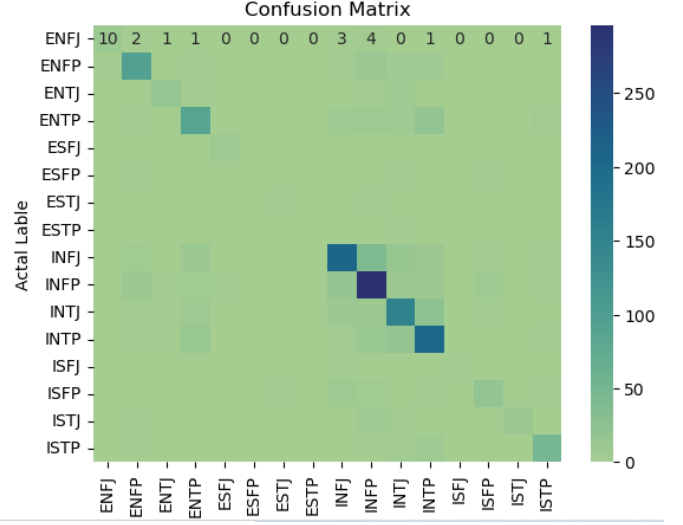


Fig. 4. Confusion matrix of the Logistic Regression with TF-IDF model on the test data.

B. Models Evaluation

To evaluate the performance of the models, metrics such as accuracy, classification report, and confusion matrix are used.

C. Models Discussions

The accuracy of the Logistic Regression model is the highest among both traditional machine learning models and deep learning models. Therefore, traditional machine learning

TABLE IV
CLASSIFICATION REPORT OF TEST DATA FOR LOGISTIC REGRESSION
MODEL WITH TF-IDF WORD EMBEDDING.

Personality Type	precision	recall	f1-score	support
ENFJ	0.67	0.43	0.53	23
ENFP	0.74	0.73	0.73	131
ENTJ	0.61	0.55	0.58	31
ENTP	0.68	0.64	0.66	135
ESFJ	0.5	0.88	0.64	8
ESFP	0.10	0.10	0.10	10
ESTJ	0.43	0.38	0.40	8
ESTP	0.33	0.17	0.22	6
INFJ	0.78	0.70	0.74	294
INFP	0.74	0.81	0.77	366
INTJ	0.67	0.71	0.69	218
INTP	0.72	0.78	0.75	261
ISFJ	0.75	0.33	0.46	9
ISFP	0.53	0.46	0.49	41
ISTJ	0.69	0.41	0.51	27
ISTP	0.72	0.74	0.73	66

models may outperform deep learning models in this problem. However, the Logistic Regression model does not perform well on the ESNP, ESTJ, and ESTP personality types.

To address the identified challenges, this study adopts an innovative approach by leveraging the *TextBlob* library to enrich the feature set of the machine learning model with sentiment analysis of the textual data. Specifically, two distinct sentiment attributes, polarity and subjectivity, derived from the *TextBlob* *sentiment* analysis, are incorporated as individual features into the model.

Polarity provides a measure of the emotional tone of the text, ranging from -1, indicating a highly negative sentiment, to +1, signifying a highly positive sentiment. This allows the model to gauge the emotional valence of the text more effectively.

On the other hand, *Subjectivity* quantifies the extent to which the text reflects personal opinions or factual information, with values ranging from 0, denoting highly objective and fact-based content, to 1, indicating highly subjective and opinion-laden content. This distinction between subjective and objective content enriches the model by offering insights into whether the text is influenced by personal feelings and biases or grounded in impartial facts. By integrating these nuanced sentiment features, the model gains a deeper understanding of the underlying emotional and factual dynamics of the text, enhancing its classification capabilities.

However, despite the addition of sentiment features such as polarity and subjectivity, the anticipated improvement in accuracy was not observed. This suggests that while sentiment analysis adds depth to the model's comprehension of textual content, its direct impact on classification performance for this specific application may require further investigation and optimization.

D. Strengths and Limitations

The strengths of this research lie in firstly comparing different types of traditional machine learning models (Logistic Regression, SVM, and Random Forest) and deep learning models (BERT, LSTM, and GRU). Additionally, various word embedding techniques are employed, including Word2Vec, TF-IDF, and BERT embedding. Furthermore, the novelty of this study lies in the utilization of data augmentation techniques such as synonym replacement and back translation. Moreover, new features are added to the dataset to enrich the data, including sentiment analysis.

The limitation of this work primarily lies in the model's accuracy, a challenge we plan to address in future efforts through several strategies. Firstly, hyper-parameter tuning will be employed to optimize the model's performance. Secondly, we aim to leverage strong deep learning models, specifically fine-tuning techniques, to enhance accuracy further. Notably, we had intended to fine-tune a BERT model to improve our results significantly. However, this was not feasible within the scope of the current study due to limitations in computational resources. Additionally, we plan to use generative models such as (GPT3) for data augmentation, which will help mitigate the

issue of data overfitting. Another limitation encountered was the usage cap of the Google Translate API, which restricts back translation to 50,000 tokens, posing a challenge for extensive language processing tasks. These areas will be focal points for overcoming the current study's limitations in future research.

V. CONCLUSIONS

In conclusion, this study has demonstrated the effectiveness of employing machine learning and deep learning techniques to predict MBTI personality types using social media data. Our investigation into various machine learning classifier models, including Random Forest, SVM, and Logistic Regression, alongside deep learning methods like LSTM, GRU, and the contextual capabilities of BERT, has revealed the significant potential of these technologies in extracting and interpreting personality indicators from the vast and rich data available on social media platforms.

Particularly, Logistic Regression emerged as the most effective model, achieving an accuracy rate of 71%. This success underscores the model's capability in handling the complexities of language used in social media posts and its efficiency in classifying personality types accurately. Further we added sentiment analysis as new feature to enhance the Logistic Regression model accuracy. Nonetheless, the integration of sentiment attributes, including polarity and subjectivity, did not lead to the expected increase in accuracy. This implies that while incorporating sentiment analysis enriches the model's understanding of text, its straightforward influence on the classification efficacy for this particular task might necessitate additional scrutiny and fine-tuning.

This research has laid a foundation for future studies to explore and refine these models further. Future work could involve the exploration of more sophisticated NLP techniques, the integration of additional linguistic features, and the application of these models to broader datasets to validate and enhance their predictive accuracy. The ultimate goal would be to develop robust systems that can assist individuals and organizations in understanding personality dynamics at scale, fostering better communication, career development, and team building based on a deeper insight into human personality traits.

REFERENCES

- [1] S. Ontoum and J. H. Chan, "Personality Type Based on Myers-Briggs Type Indicator with Text Posting Style by using Traditional and Deep Learning," *arXiv*, 2022.
- [2] G. Ryan, P. Katarina, and D. Suhartono, "MBTI Personality Prediction Using Machine Learning and SMOTE for Balancing Data Based on Statement Sentences," *Information*, vol. 14, no. 4, pp. 217, 2023.
- [3] M. H. Amirhosseini and H. Kazemian, "Machine Learning Approach to Personality Type Prediction Based on the Myers-Briggs Type Indicator," *School of Computing and Digital Media, London Metropolitan University*, 2020, London, UK.
- [4] B. Cui and C. Qi, "Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction," 2017.

- [5] Z. Mushtaq, S. Ashraf, and N. Sabahat, "Predicting MBTI Personality type with K-means Clustering and Gradient Boosting," in *2020 IEEE 23rd International Multitopic Conference (INMIC)*, Bahawalpur, Pakistan, pp. 1–5, Nov. 2020.
- [6] "MBTI Personality Type Dataset," Kaggle, Accessed: Apr. 7, 2024. [Online]. Available: <https://www.kaggle.com/datasnaek/mbti-type>.
- [7] V. S. Spelman and R. Porkodi, "A Review on Handling Imbalanced Data," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pp. 1–11, Mar. 2018.
- [8] V. A. Kozhevnikov, E. S. Pankratova, "Research of the Text Data Vectorization and Classification Algorithms of Machine Learning," in *Theoretical & Applied Science*, vol. 85, no. 05, pp. 574–585, May 2020.
- [9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv*, 2019. [Online]. Available: <https://arxiv.org/abs/1810.04805>.
- [10] D. Jain, A. Kumar, and R. Beniwal, "Personality BERT: A Transformer-Based Model for Personality Detection from Textual Data," in *Proceedings of the International Conference on Computing and Communication Networks*, 2022.
- [11] A. Khvostikov, K. Aderghal, J. Benois-Pineau, G. Catheline, M. K. A. Gwenaëlle, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review," *Frontiers in Aging Neuroscience*, vol. 10, p. 329, 2018.
- [12] B. E. Boser, I. M. Guyon, V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, (COLT '92), New York, NY, USA, pp. 144–152, 1992.
- [13] A. Upreti, "Convolutional Neural Network (CNN): A comprehensive overview," *IJMRGE*, pp. 488–493, Aug. 2022.