Final Report: MBTI Personality Prediction

Vibhu Krovvidi (vk2500), Meghan Shah (ms5767), Vrinda Bhat (vgb2113), Dhivyadarsan Gomathi Muthamilselvan (dg3233), Varalika Mahajan (vm2695)

Abstract: The Myers-Briggs Personality Type Index (MBTI) has been influential to the field of psychology since its publication in 1962 and remains one of the most widely used and recognized personality tools. The MBTI test provides people with one of 16 personality type profiles, from which they can assess where they fall on scales such as Extraversion vs. Introversion, Thinking vs. Feeling, etc. Given both the popularity and longevity of this scale, along with the increasing ubiquity of social media, it can be valuable to successfully predict a person's MBTI personality type from their social media posts. Utilizing text pruning and cleaning of posts along with word embeddings, this project increased the performance of model predictions by choosing to divide the dataset into 4 equal datasets with each potential personality pair considered separately (ie. E/I, S/N). After trying a range of models, different models seemed to perform better at predicting different personality pairs as assessed through micro-F1 scores.

Introduction: MBTI remains popular and outperforms the Big Five openness in specific domains, since it consists of four pairs of opposing dichotomies, namely: Extraversion-Introversion(E/I), Sensing-Intuition(S/N), Thinking-Feeling(T/F) and Judging-Perceiving (J/P). This project focuses on utilizing machine learning to create classifiers that can categorize individuals based on text samples from their social media postings into their Myers-Briggs Type Index (MBTI) personality type. Firstly, because of social media's widespread use, a classifier of this kind would have access to a large amount of data on which to conduct personality tests, giving more individuals access to their MBTI personality type accurately and rapidly. Accurate inference of users' personalities is substantial to the performance of downstream applications such as personalized advertisements on social media. Since social media personalization positively impacts consumer brand engagement and brand attachment, but self-reported personality assessments are not common across social media platforms, a more scalable and sustainable way to infer users' personalities can be through the analysis of users' posts on social media.

Approach: Using Natural Language Processing and common machine learning techniques, this project tries to predict a user's MBTI type by combining predictions across the 4 dichotomies. Treating each dichotomy as independent allows for stronger predictive power than a consolidated classification task would have created.

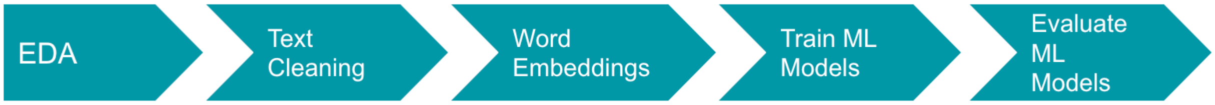

Fig 1. Project Approach Overview

Pre-Processing: Since most ML models cannot directly process strings, the posts must be converted into numerical form. To do this, we pre-process the posts to remove irrelevant words (stop words), reduce words to their roots (lemmatization), prune text (only keeping adjectives, nouns, and prepositions), and remove special characters. Next, we used the GloVE pre-trained word embeddings and mapped words to 100-dimensional vectors which project words into a 100-dimensional space such that similar words are similarly located in the space. At this point, we had a series of posts in vector array form and their corresponding MBTI types.

MBTI types operate in mutually exclusive characteristic pairs like Extroverted/Introverted and Feeling/Thinking. It thus makes more sense to divide the dataset into 4 equal datasets with each pair considered separately. Thus, we have 4 datasets with labels encoded as binary values (eg E = 1, I = 0). Since an array of word vectors is not easily processed by most classification algorithms, we decided to take the average vector of each post. This is justified since each word's meaning is encapsulated by the vector space, and thus the average vector would correspond to the average 'meaning' of the post. This also has the added advantage of fixing the number of independent variables to 100 columns, with each representing a dimension of the space.

Training ML Models: We began our training of ML models by splitting the datasets into development and test sets in a 75:25 ratio with stratified random sampling to preserve class ratios in both sets. By having 4 datasets from the original dataset, a challenge is the unequal ratio of classes. To counter this, we used SMOTE, a method of data augmentation that synthetically creates points to balance the class ratio. We then ran a series of Machine Learning models: Support Vector, Random Forest, Histogram Gradient Boosting, K-Nearest Neighbors, Logistic Regression, Naive Bayes, and a Long-Short Term Memory Recurrent Neural Network. To ensure that we are not overfitting these models and that we are selecting the best hyperparameters for each model, we used Grid Search cross-validation with 5 folds. We encapsulated this search within a pipeline to ensure that the SMOTE processing did not leak between training and validation folds. Due to the imbalanced nature of the datasets, we opted to use the micro F1-Score (which balances precision and recall) as a metric of success. Our objective was thus to find the model that produced the best micro F1-Scores for each MBTI characteristic pair.

Table 1. Micro F1-Score of Models

Pair	SVC	RF	HistGB	Calib-HistGB	KNN	LogR	NB	LSTM
I/E	0.585	0.231	0.637	**0.678**	0.569	0.583	0.432	0.231
S/N	0.571	0.138	0.757	0.807	0.643	0.587	0.634	**0.862**
T/F	0.683	0.459	0.646	0.642	0.554	**0.692**	0.544	0.541
J/P	0.598	0.396	0.546	0.547	0.504	**0.633**	0.515	0.396

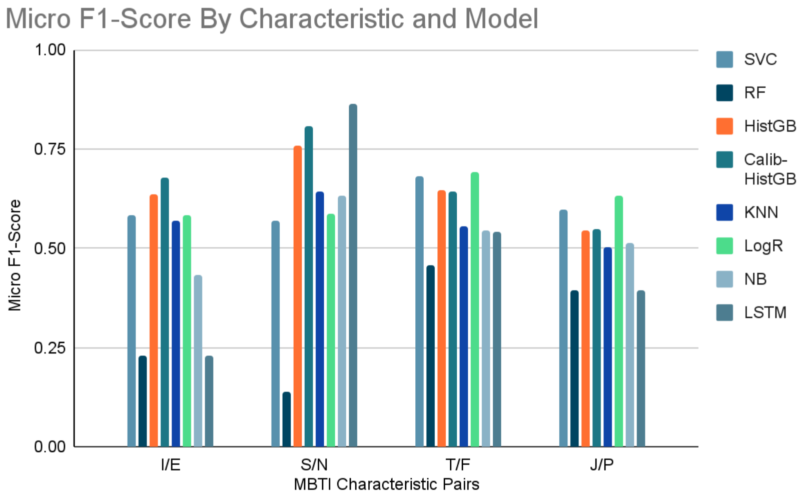

Fig 2. Comparison of Model Performance

Analysis: The prediction performance of the eight proposed classifiers was evaluated based on the F1-Micro score. Since the accuracy metric is more sensitive to the distribution of the target variable, the F1-Micro score is evaluated on top of accuracy since it is more important to capture the sensitivity and specificity performance of a classifier on an imbalanced dataset. The F1-Micro score for the classifiers on the dataset is tabulated in Table 1. According to Table 1 above, Calibrated HistGB performed best for I/E, LSTM performed best for S/N, and Logistic Regression performed best for T/F and J/P.

Conclusion: This project demonstrates that the MBTI profiles should be viewed as a composite of dichotomies as opposed to 16 independent classes. Eight classifiers were compared in the task of MBTI dichotomy prediction. Different classifiers worked best with different pairs suggesting that the dependence amongst certain MBTI types makes direct personality predictions ineffective. The diversity of micro F1 Scores across models and dichotomies is interesting as it suggests that the dichotomy predictions might rely heavily on linguistic semantics rather than a statistical approach to the terms like TF and TF-IDF. A semantic-based approach would thus be necessary for tackling the prediction of each MBTI dichotomy separately.

Future Work: We propose using LSTM to perform sequence classification of each pair. Using each word vector as an element of the sequence as opposed to the average vector of a post coupled with the gated structure of the LSTM can mean that linguistic patterns within posts as well as across posts can be learned to predict MBTI personalities more effectively. Alternatively, we can employ the Generative Pre-trained Transformer 3 (GPT-3) to boost performance. This project can therefore serve as a solid foundation for more involved deep learning modeling on the topic. By including more social media postings for each target class, we can also increase the dataset's quality. We may include LinkedIn posts as well, making the categorization more applicable to employers.