Predicting MBTI Type Through Machine Learning Text Classification

John Clark , Shrinivas Joshi , Courtney Kennedy

University of California, Berkeley

Abstract

The use of machine learning to predict personality type has gained increasing attention over the past several years. Initial efforts were largely focused on the Five Factor Model of Personality, also referred to as the Big 5. More recently, efforts have been undertaken by several researchers to utilize machine learning to predict the Myers-Briggs Type Indicator (MBTI) through the analysis of text in social media posts. We introduce four significant items in our analysis: 1) A model utilizing modern transformer architectures to predict MBTI type, 2) A new and large database of 3.3 million posts from a popular online personality type forum called Typology Central upon which we train our model on up to 1.2 million rows, 3) Corrections for dataset imbalance through resampling, and 4) Introduction of linguistic features to enhance prediction accuracy. We compare both 16 multi-class classification approaches with 4-fold binary classification. Our goal in this effort is to improve upon the prediction accuracy of previous studies using automated machine learning prediction while taking care to prevent any potential data leakage. We also seek comparable results to those achieved by manual MBTI questionnaires.

1.0 Introduction

Personality assessments are typically conducted through questionnaires, at times also combined with interviews. The most common personality assessment in use today is the Myers-Briggs Type Indicator. There are three versions of the official assessment which contain 93, 144, and 222 questions respectively. The MBTI identifies consistent and habitual patterns in an individual's thinking. A small number of studies have used machine learning text classification techniques, with varying levels of success, to predict MBTI type based on social media posts (see Appendix for a summary). We attempted to consider lessons learned from each of these studies to develop an improved model using transformer architectures. Our efforts were conducted as a series of incremental experiments using a variety of classifiers and features as well as resampling techniques to balance the data.

In this project, we used a novel dataset that has not been used in previous studies from a personality-type Internet Forum called Typology Central. This site is 15 years old and contains 3.3 million posts from approximately 40,000 users. Users include their own MBTI types in their profiles so that it is possible to have an MBTI label associated with each post.

The MBTI system defines 16 types, each with four attributes, designated by letters. The four attributes include

1. Introversion (I) versus Extraversion (E),
2. Sensing (S) versus Intuition (N),
3. Thinking (T) versus Feeling (F), and
4. Judging (J) versus Perceiving (P)

No person is an absolute introvert, extravert, thinker or feeler, etc. Rather, an individual's type is determined by that person's ***preference*** for each of these four attributes that characterize habitual patterns of thinking. *Introverts* direct their energy inward to their own feelings and thoughts whereas *Extraverts* are more interested in the outer world of people and things. Introverts and Extraverts gain energy by engaging their preferred orientation. There are two types of attributes that describe cognitive processes of how we ***perceive*** information. *Sensing* refers to perceiving information through the five senses. *Intuition* is an unconscious process where ideas or associations are incorporated into perceptions that come from the inside or outside. These two processes compete for attention but individuals will tend to prefer one over the other. There are two types of attributes that describe cognitive processes we use to ***judge*** or come to conclusions. *Thinking* is a logical and impersonal process to come to decisions. *Feeling* applies personal and subjective values to decisions. Finally, a person may have a preference for *Judging* or for *Perceiving*, as an orientation and overall way of life. We all perceive and judge but it is difficult to do both at the same time. Those who prefer Perceiving will tend to be more spontaneous and prefer to keep their options open longer whereas those who prefer Judging prefer to plan and organize and may make decisions more quickly.

In this paper, we first analyze previous work where machine learning algorithms were used to predict personality type — primarily the MBTI. We compare the use of a multi-class model of the 16 types against the use of a 4-fold binary classification. Our initial baseline analysis was conducted on a sample of social media posts, making predictions of MBTI type based on those posts using the multi-class model. In this baseline, we evaluated Logistic Regression, Support Vector Machine (SVM) as well as Bidirectional Encoder Representations from Transformers (BERT). After this baseline analysis, we considered what we learned as well as further data points from the second review of past literature to make a series of recommendations on further experiments and steps.

Subsequent experiments included the addition of a Convolutional Neural Network layer, binary classification layers, the addition of multi-headed attention to BERT, analysis of users vs. posts, and additional features including user profile information and a novel approach for incorporating three different types of linguistic features (Empath, NRC Emotion Lexicon, and NLTK Part of Speech Tags) which were fed into BERT. Finally, we evaluated additional classifiers including SetFit and T5. We assessed the success of these experiments by comparing accuracy scores against models developed in previous similar studies since all of those studies consistently reported on that metric.

2.0 Prior Work

Khan, et al. (2020) highlighted that the skewness of the datasets was the main issue with prior work, which they addressed through the use of oversampling. Hernandez and Knight (2017) compared an analysis of individual posts vs. an analysis of a collection of posts for particular users. They found that user classification using a corpus of posts from a user achieved far higher accuracy, indicating that a sufficient amount of text may be needed to support accurate classification. Celli and Lepri (2018) and Stajner and Yenikent (2020 and 2021) came to the conclusion that tweets might not contain sufficient amounts of MBTI signals, even after concatenating up to 150 - 200 tweets per user, due to the nature of Twitter posts. Stajner and Yenikent (2021) used 2000 Twitter posts per user in their classifier and they indicated, “In most of the languages, the best classifiers outperformed the majority-class baselines for only the E/I and T/F dimensions."

As one example of a personality type and correlation to language, studies by Maitresse et al. (2006), noted that extraversion is marked by greater use of verbs, adverbs, and pronouns. They also noted that sentences of extraverted types tended to be simpler with fewer words, lower linguistic diversity, and fewer negations. They used statistical regression combined with the LIWC utility and MRC database to predict Big 5 personality types. They described that introverts generally use more negating words and use a wider vocabulary. Gjurkovic, et al. (2017) utilized a number of additional features in addition to the text itself to predict personality type. This included linguistic features, using the LIWC database and the MRC Psycholinguistic Database. Features included standard word count statistics (word count, words longer than six letters, number of prepositions, etc), psychological processes (emotional, cognitive, sensory, and social processes), relativity (words about time, the past, the future), personal concerns (such as occupation, financial issues, health) and other dimensions such as counts of various punctuation and profanity. Additionally, they utilized user activity features such as the number of posts, time intervals between comment timestamps, as well as daily, weekly, and monthly distribution of posts.

The significant majority of prior work involved the use of sklearn classifiers such as Naive Bayes, SVC, Logistic Regression, and XGBoost. Hernandez and Knight (2017) used transformers. They utilized GloVe to create embeddings and tried a number of Recurrent Neural Network (RNN) classifiers including SimpleRNN, GRU, LSTM, and Bidirectional LSTM. LSTM provided the highest accuracy. Adam and binary cross entropy were utilized for the loss function and they used four different binary classifiers to predict personality type. To analyze users as opposed to individual posts, they took the mean of the class probability for all of the posts in a user's corpus and rounded it to either 0 or 1 to predict the letter type. Gjurkovic et al. (2021) used both sklearn and neural networks, including BERT, in an attempt to predict Big 5 type and also analyzed correlations between Big 5 and MBTI as well as Enneagram though they reported that BERT did not perform well. They did not run models to predict MBTI type.

3.0 Methods

3.1 General Description of Data

Each user in the Typology Central data set has a unique username and optional information associated with their profile. This information includes the following: age, occupation, number of posts, and MBTI type. Each post is made by a user, who is identified by their username. Therefore, we can use the MBTI type information in a user’s profile to define a label for each post. This information provides us with a means to perform supervised training of various models on the text of posts. The posts vary in length and style and use BBCode Markup for quoted text, links to websites, images, video, and text styling. Within the text of the posts on Typology Central, users discuss a wide variety of topics including personality type.

3.2 Data Cleaning

After analyzing the data, we performed several different types of data cleaning. Our goal in cleaning the data was to remove any invalid samples and sources of bias. One potential concern was the existence of MBTI type information in the posts themselves creating a risk of data leakage. As Clark (2015) noted, there is a relationship between Enneagram type and MBTI type and for that reason, Enneagram type information present in posts could also result in data leakage.

For each user entry, we performed the following steps:

- Removed users with invalid MBTI type
- Removed users with invalid Age
- Removed users with an invalid number of posts

For each post, we performed the following steps:

- Removed tabs and newlines
- Made everything lowercase
- Removed BBCode markup, including quoted text
- Removed links to images and videos
- Removed HTML links and tokens ending in .com
- Removed special characters
- Masked references to MBTI types using XXXX
- Masked references to Enneagram types using YYY
- After everything above, we then removed any posts that had zero length

After all of these efforts, our cleaned database had approximately 1.3 million posts and 7000 users.

3.2.2. Resampling

In evaluating the data, we noticed that the distribution of posts across MBTI types and subtypes was not uniform and did not match the typical MBTI distribution found in the general population. This presented a challenge because algorithms applied on unbalanced classified data can result in the outcomes diverging towards the classes with a larger amount of data and bypassing the smaller ones.

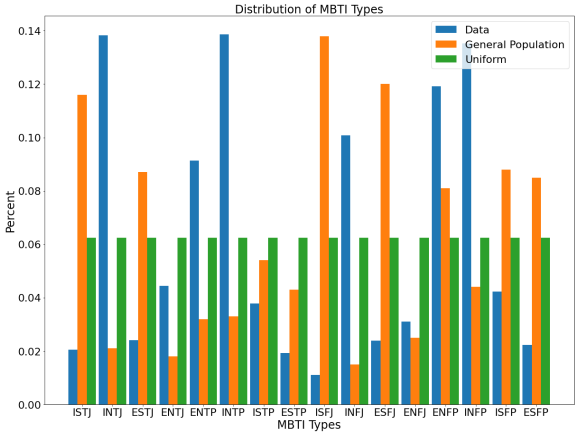

Figure 1: Distribution of MBTI Types in the original Typology Central data, compared to the general population and uniform distribution.

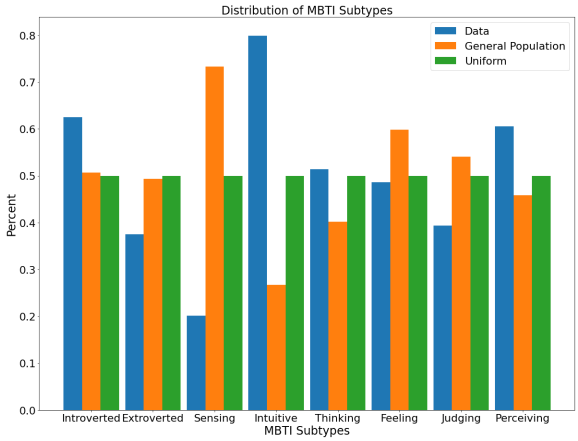

Figure 2: Distribution of MBTI subtypes in the original Typology Central data, compared to the general population and uniform distribution.

To adjust for this imbalance in the input data, we resampled the training data, using the sklearn resample function, in two different ways: 1) to match the distribution found in the general population and 2) to be uniformly distributed. We shuffled the data and then split it into training, validation and test datasets, and then applied resampling to the training set. We did not resample validation and test data.

3.3 Baseline

We began by using a sample of our dataset with a very basic multi-class model classifying text only for each of the 16 MBTI types using Logistic Regression, Linear SVC, and BERT, based on individual post data. Tensorflow and Keras were used to implement BERT. We used a grid search to find the best-performing parameters for the Logistical Regression and Linear SVC models. The three inputs to the model were each tested:

- Raw unbalanced data found in the original dataset,
- Resampled with a uniform distribution across the types,
- Resampled to a distribution across each MBTI type that matches the MBTI distribution found in the general population.

Resampled datasets performed better than unbalanced data. Even with resampling, however, the multi-class models running against individual posts and no additional features performed very poorly as shown in the following table. It caused us to question whether our models were learning.

Classifier	Key Parameters	Best Accuracy Score
Logistic Regression	C = 10	0.0502
Linear SVC	C = 5	0.0544
BERT	max_length = 128 learning_rate = 0.00005	0.01250

Table 1: Baseline accuracy scores

Through test runs of various iterations of the model and further consideration of the literature, we identified several areas of focus upon which to base our experiments.

- **More robust transformer models** - We determined that we should evaluate more complex and finely tuned transformer models.
- **MBTI Multi-class granularity** - We considered that the multi-class model may not be granular enough for us to determine where our models were making mistakes and decided to add models using four binary classes.
- **Testing a corpus of posts vs. individual posts** - Some previous studies appeared to show that testing against a corpus of posts for an individual user, vs. individual posts, provided more accurate results.
- **Additional features** - We considered the value of adding features to support the classification, including linguistic features and user activity features. Several linguistic feature databases were mentioned in the literature, including LIWC, MRC, and the NRC Emotion Lexicon.

3.4. The Experiment

3.4.1 Feature Engineering

Based on the observations in our baseline, we evaluated a number of additional features to be candidates for subsequent experiments.

3.4.1.1 MBTI Subtype Pair Classification

First, we added categorical labels for each of the MBTI subtype pairs. These were expressed as booleans, with the following meanings:

Feature	True Value	False Value
is_I	Post author was Introverted.	The post author was Extraverted.
is_S	Post author was Sensing.	The post author was Intuitive.
is_J	Post author was Judging.	The post author was Perceiving.
is_T	Post author was Thinking.	The post author was Feeling.

Table 2: 4-fold binary features

3.4.1.2 - Linguistic Features

Our first choice for Linguistic Features was the LIWC database. Unfortunately, the LIWC database, referenced in comparative studies, was no longer available for use. After considering several alternatives, we chose three options to provide additional features.

- Empath (Ethan Fast et al. 2016), analyzes text across 200 topics and emotions and generates and validates lexical categories, and draws connotations between words and phrases. The product creators consider this module to be comparable to the LIWC product, but with a much larger database of 1.8 billion words.
- NRC Emotion Lexicon, contains a list of English words and their associations with eight human emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (positive and negative). It has been successfully used for word, sentence, tweet, and document-level sentiment and emotion analysis for personality trait identification.
- POSTAG, denotes 17 NLTK part-of-speech categories, such as nouns, verbs, adjectives, adverbs, etc. Previous studies (T.A. Litinova and P.V. Serin, 2015) have shown a correlation between several Big-5 personality dimensions and the frequency of different parts of speech. Since MBTI has a strong correlation with Big-5, we considered that this may be an effective feature set.

3.4.2.3 - User Profile Information

We analyzed several additional features at the user profile level including:

- Age
- Number of posts
- Username
- Occupation (optional field)

Username was the most important feature of these because it provided an effective way to “group” posts at the user level rather than the individual post level. Other studies appeared to concatenate post data at the user level and feed those into a classifier. That was not an ideal option in our study due to token limits in the transformers that we utilized. Since our project goal is only to predict MBTI type, we did not consider age bias to be a significant consideration.

3.4.2 BERT + CNN Against Individual Posts and Multi-class Prediction

Our first experiment involved the use of BERT with a CNN, using multi-class prediction directly on post data without any additional features. A data generator allowed us to process the data in smaller chunks and was used to enable us to process the entire dataset of 1.3 million rows. The accuracy of this model slightly improved from the base case yet was still poor. Overall accuracy was still only at the level of .07 which is not much higher than random chance (for 1 out of 16) or .0625. The confusion matrix highlighted where our model was working incorrectly. Our model consistently overpredicted the following types: INFJ, ISTP, ESFJ, ESFP, and ISFP.

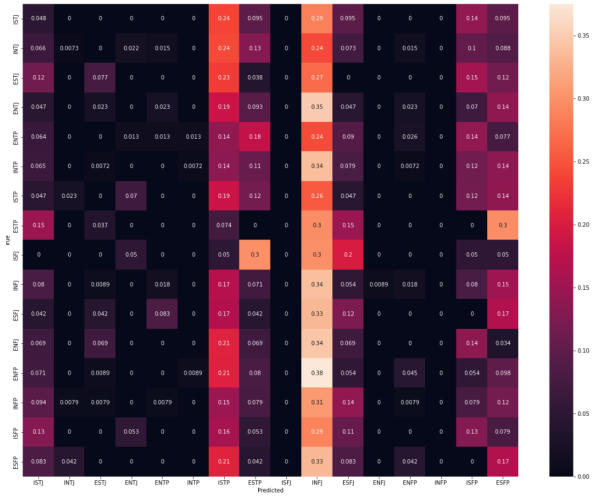

Figure 3 - BERT + CNN confusion matrix

3.4.3 BERT With Multi-Headed Attention Based on Users & Posts with Linguistic Features, User Profile Information, and Multi-class Prediction

Our next major experiment included several enhancements using a BERT-only model — specifically, the addition of more features. These features included:

- Username
- Number of posts
- Age
- Empath (variable number of tokens)
- NRC emotion lexicon (12 tokens)
- NLTK part-of-speech categories (17 tokens)

We consolidated all of these features as well as individual post data as input into BERT. Since BERT is designed to evaluate text vs. individual features, we extracted these features as additional words/tokens and forced an SEP token between each feature (or feature set) to signal that BERT should treat them as separate sentences. The intent was to leverage BERT's use of attention score vectors for sentences and to allow BERT to do the work for us to classify data at the User vs. Post level. Additional multi-headed attention layers were added to evaluate across sentences.

After making these changes, we evaluated the model with a sample of 5000 posts. Results using this method were substantially improved from previous models:

	precision	recall	f1-score	support
ISTJ	0.96	0.80	0.88	92
INTJ	0.93	0.72	0.81	691
ESTJ	0.98	0.97	0.97	125
ENTJ	0.93	0.89	0.91	215
ENTP	0.75	0.80	0.78	493
INTP	0.99	0.45	0.62	706
ISTP	0.58	0.88	0.70	208
ESTP	0.94	0.91	0.92	79
ISFJ	0.78	0.88	0.83	65
INFJ	0.96	0.57	0.71	477
ESFJ	0.41	0.99	0.58	120
ENFJ	0.54	0.91	0.68	147
ENFP	0.80	0.79	0.79	598
INFP	0.59	0.82	0.68	675
ISFP	0.67	0.92	0.77	201
ESFP	0.79	0.84	0.82	108
accuracy			0.75	5000
macro avg	0.79	0.82	0.78	5000
weighted avg	0.81	0.75	0.75	5000

Table 3 - Classification report for BERT with multi-headed attention based on users with linguistic features, user profile info, and multi-class prediction

Our confusion matrix also showed very significant improvements. The model had more difficulty predicting INTP, INFJ, INTJ and ENTP types.

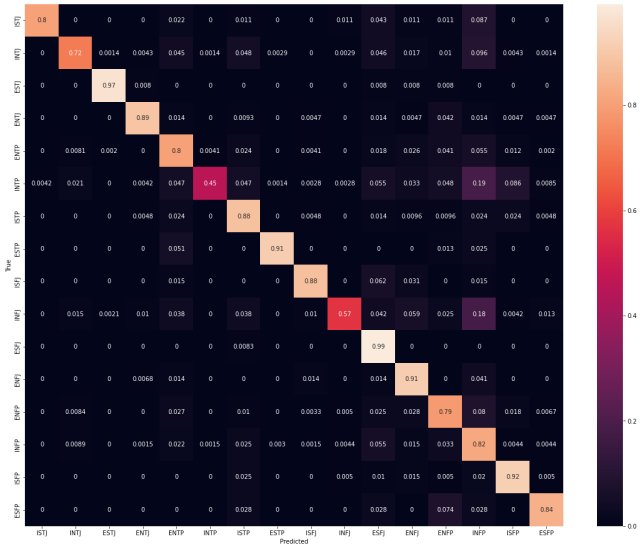

Figure 4 - Confusion matrix for BERT with multi-headed attention based on users with linguistic features, user profile info, and multi-class prediction

3.4.4 BERT With Multi-Headed Attention Based on Users & Posts with Linguistic Features, User Profile Information, and Binary Class Prediction

Our next experiment was to add binary classes for each of the MBTI type attributes and was tested using a sample of 5000 posts. The overall performance of the model on the individual dimensions was strong with accuracy results ranging from .70 to .89.

	precision	recall	f1-score	support
Introvert	0.62	0.98	0.76	1885
Extrovert	0.98	0.64	0.77	3115
accuracy			0.77	5000
macro avg	0.80	0.81	0.77	5000
weighted avg	0.85	0.77	0.77	5000

	precision	recall	f1-score	support
Sensing	0.99	0.64	0.77	4002
Intution	0.40	0.97	0.57	998
accuracy			0.70	5000
macro avg	0.70	0.81	0.67	5000
weighted avg	0.87	0.70	0.73	5000

	precision	recall	f1-score	support
Thinking	0.78	0.96	0.86	2391
Feeling	0.95	0.75	0.84	2609
accuracy			0.85	5000
macro avg	0.86	0.85	0.85	5000
weighted avg	0.87	0.85	0.85	5000

	precision	recall	f1-score	support
Judging	0.88	0.94	0.91	3068
Perceiving	0.89	0.80	0.84	1932
accuracy			0.89	5000
macro avg	0.89	0.87	0.88	5000
weighted avg	0.89	0.89	0.88	5000

Table 4 - Classification report for BERT with multi-headed attention based on users with linguistic features, user profile info, and binary class prediction

This model was very good at predicting introversion (.98), intuition (.97), thinking (.96) and judging (.94). It was less successful in predicting extraversion (.64), sensing (.64), feeling (.75), and perceiving (.80).

3.4.5 BERT Test Run With Full Dataset

In an attempt to address errors in the smaller samples, we next trained our BERT classifier using the full training dataset of 1.2 million posts. This was tested using 93,000 test posts. The larger training set significantly improved accuracy scores for the multi-class model to .81. T/F (.90), J/P (.92), I/E (.91) and N/S (.94) all demonstrated significant improvements.

3.4.6 SetFit

For our next experiment, we considered the fact that our model's long training time may potentially make it difficult to productize. Using a premium, high ram GPU in Google Colab, training BERT using 1.2 million posts took 7 ½ hours for one epoch. To address this issue, we considered SetFit as a potential alternative (Sentence Transformer). During our testing, after tuning hyperparameters for learning rate, number of epochs, batch size, number of iterations, seed, max iterations, and solver (lbfgs), we were able to achieve .77 accuracy on the multi-class model. This result was similar to the accuracy of our BERT model, but was achieved with a small fraction of the training time that our BERT model required.

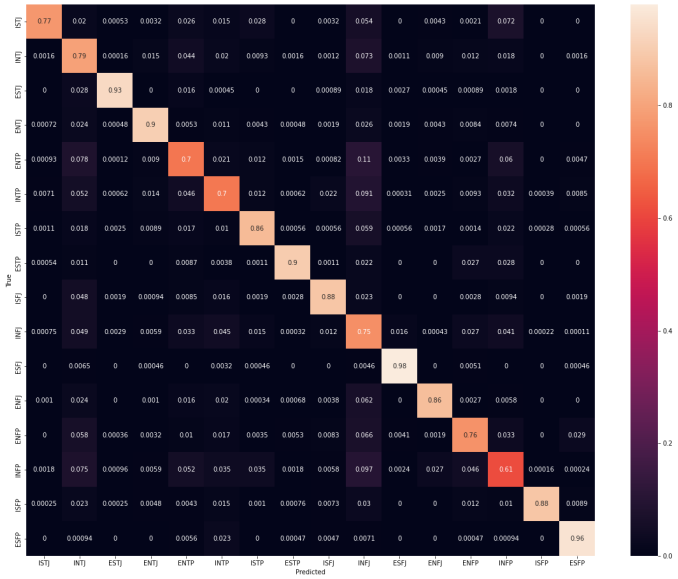

Figure 5 - Confusion matrix for SetFit with full training dataset

3.4.7 T5

Our final experiment was to use T5 to explore the effectiveness of a text-to-text transformer for our classification task. We focused on using the T5-small model, and trained it with a subset of our training dataset that included 120,000 samples. We used the same set of features as with the final BERT model though we utilized PyTorch instead of TensorFlow/Keras. The accuracy of our T5 model far exceeded our other models, as shown in the following chart.

	precision	recall	f1-score	support
ENFJ	0.95	0.95	0.95	2920
ENFP	0.98	0.92	0.95	11048
ENTJ	0.93	0.96	0.94	4186
ENTP	0.92	0.94	0.93	8567
ESFJ	1.00	0.99	0.99	2169
ESFP	0.98	0.98	0.98	2126
ESTJ	0.99	0.99	0.99	2243
ESTP	0.94	0.98	0.96	1840
INFJ	0.93	0.89	0.91	9275
INFP	0.95	0.89	0.92	12497
INTJ	0.98	0.92	0.95	12850
INTP	0.95	0.90	0.93	12925
ISFJ	0.82	0.95	0.88	1061
ISFP	0.97	0.96	0.96	3946
ISTJ	0.62	0.98	0.76	1877
ISTP	0.66	0.97	0.78	3576
accuracy			0.93	93106
macro avg	0.91	0.95	0.92	93106
weighted avg	0.94	0.93	0.93	93106

Table 5 - Classification report for T5 on users with linguistic features, user profile info, and multi-class prediction

With one epoch, we noted a small number of out-of-class inferences. We eliminated that issue and substantially reduced both Training and Validation Loss by running 4 epochs.

Epoch	Train Loss	Val Loss	Val PPL.
1	0.481	0.139	1.149
2	0.136	0.078	1.081
3	0.089	0.060	1.062
4	0.074	0.055	1.057

Table 6 - T5 loss patterns with increased epochs

Our T5 Model also had binary classification results that were exceptional, with accuracy scores from .98 to .99. These are shown in the Appendix.

3.4.8 Error Analysis

We evaluated the confusion matrix for each of our models and also analyzed erroneous classification data to determine where the models may be making mistakes. Looking at patterns for mispredicted types, our models had relatively consistent challenges predicting the INTP and INFJ personality types, as shown in Figure 6. The main issue with the BERT multi-class model was overprediction of the INFP and ENFP personality type. All of our models therefore had issues with the INFP type in one manner or another. BERT performed most poorly in predicting INTP, with an overall .60 accuracy and INFJ with an overall accuracy of .63.

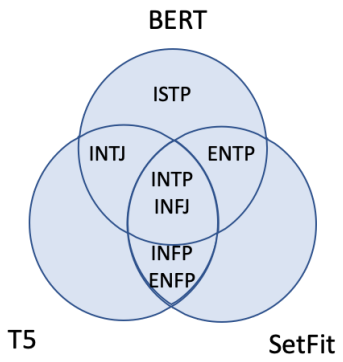

Figure 6 - Top 5 Incorrectly predicted MBTI types for each model

The SetFit model provided .61 accuracy for INFP prediction, .70 accuracy for INTP and ENTP prediction, and .75 accuracy for INFJ. It appeared to have similar issues as BERT, but those problems were more dispersed. Though the T5 model was highly successful overall, we did notice the model was less successful at predicting INFJ (.89), INFP (.89) and INTP (.90). The specific misclassification patterns for this model were less clear than in the BERT and SetFit cases. We hypothesized overall that the linguistic features were the most significant contributor to errors across all our models and therefore tuning those features may have the most significant impact on outcomes.

4. Results and Discussion

Our goal was to predict MBTI type through an evaluation of transformer-based architectures and to increase the performance as compared to previous models. In prior work, we noted several apparent limitations. A core weakness of all but one of the studies we evaluated was a failure to compensate for imbalanced input data. We also noted the risk of data leakage in other studies, several of which did not remove MBTI information from post text and none of which removed enneagram information. Only two had attempted transformer architectures. Our results are compared to those of other studies in the Appendix.

Because our results were so much better with the additional features, we analyzed how those features contributed to accuracy. To evaluate this, we ran a comparison removing one feature at a time, illustrated in the following table.

Item Removed	I/E	S/N	T/F	J/P	Multi-class
NRC Emotion Lexicon	.87	.86	.86	.85	.73
Empath	**.91**	**.95**	**.87**	**.88**	**.79**
Part of Speech	.86	.85	**.87**	.87	.77
Username	.86	.77	.85	.84	.73
No Features Removed	.84	.88	.86	.86	.77

Table 7 - Comparative evaluation, removing various features from the BERT model shows that Empath was the least effective feature set. Removing it improved accuracy. Bold are the best scores, red are the worst.

The results indicate, as expected, that username was the most important feature we added because this increased the number of posts and the corresponding volume of text that could be analyzed per user. The least effective feature set was Empath, which appeared to detract from accuracy. The NRC Emotion Lexicon and NLTK part of speech tags contributed positively to the overall accuracy and indeed, the best results combined NRC Emotion Lexicon, Part of Speech tags, and Username.

The question arises as to why T5 was the best performing model vs BERT and SetFit. T5 has both encoder-decoder blocks and is not “encoder-only”, enabling more model parameters. We used the T5-small model, which provides eight-headed attention across the encoder and decoder, resulting in approximately 60 million parameters. Our belief is that connected encoder-decoder blocks with additional attention heads and a common loss function are highly effective in improving classification accuracy.

5. Conclusion

5.1 Summary

We sought to make improvements in the automated prediction of MBTI type based on an analysis of the text in social media posts. T5 multi-class prediction accuracy of .93 and binary classification test results of .98 - .99 exceeded our expectations. Our classifier compares favorably with published MBTI test reliability statistics that showed an overall average of .815 on each of the four binary attributes (Capraro et al., 2002). It also compares favorably with other studies that have applied machine learning to this task as noted in Table 8 in the Appendix.

Our overall conclusion is that no transformer-classifier performed well without additional linguistic features. In our final test runs, we had added 32 tokens derived from our additional feature set. Consistent with other studies, we noted that the volume of text analyzed significantly affects accuracy, and that analysis of users vs. individual posts results in substantial performance improvements. Figure 6 in the Appendix clearly shows how accuracy improves as the number of posts increases.

We observed that a larger training dataset significantly increased accuracy. We saw that the Empath database impaired the effectiveness of our classifier. We noted adding multi-headed attention increased our model's effectiveness with BERT.

5.2 Recommended Future Work

Given the effectiveness of linguistic features shown in our study, we recommend future work focus on evaluating and comparing various options to determine an optimal mix of these types of features. In particular, it would be helpful to identify a set of options that provide features similar to the LIWC database. Given the token limit for our classifiers (384 for SetFit and 512 for BERT and T5), with the method we used to add features, there is a tradeoff between the number of features added and the amount of tokenized text captured in post content. It should be noted that the linguistic features were derived using the entire post content with no truncation as employed by model tokenizers. Future studies should also consider the use of PyTorch instead of Tensorflow because at the time of this study better model interpretability tools appear to be available (e.g., Captum) that supports troubleshooting and feature analysis to enable improvements to the transformer models.

We believe that further opportunities may be available to improve performance of the multi-class model through hyperparameter tuning. If our classifier were to be productized, the volume of text available per user would be a key design consideration. Thresholds for an acceptable volume of text would need to be considered. Finally, other social media platforms, such as Twitter, Tumblr, or Reddit may provide a different mix of post content, that in combination could be used to support more effective training of the classifier.

Appendix

Figure 7 - Classification errors vs. number of posts for a user with the T5 classifier (top 5 mispredicted types)

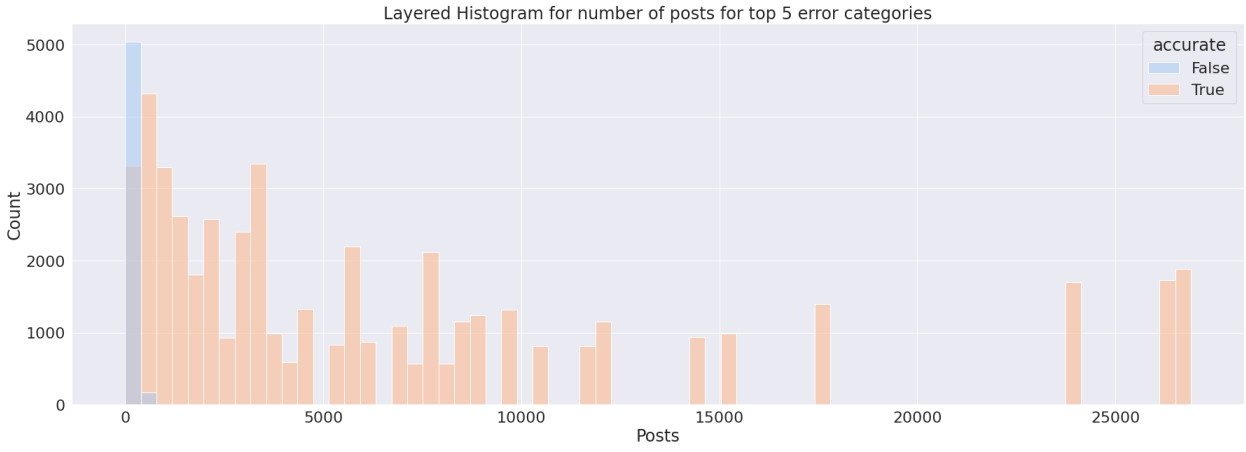

Table 8 - Summary of related research compared to this study

Research	Objective	Methods	Accuracy	Limitations
(15) Capraro et al. 2002	Assess the reliability of official MBTI assessment instrument	Review of reliability estimates, from 70 articles, calculating mean, standard deviation, min, and max.	I/E - .838 N/S - .843 T/F - .764 J/P - .822 Overall - .815	No actual testing or regression analysis was conducted by the authors as they relied on other studies.
(4) Rayne Hernandez, et al. 2017	Automated prediction of MBTI type based on a Kaggle dataset of 8675 social media posts	• RNN, GRU, LSTM, and Bidirectional LSTM • Post classification and User classification (using the mean of the class probability of posts in a user's corpus)	Post Classification I/E - .54 N/S - .529 F/T - .578 P/J - .529 User Classification I/E - .676 N/S - .620 F/T - .778 P/J - .637	Relatively small dataset.
(5) Mohammad Hossein Amitrhosseni, et al. 2020	Automated prediction of MBTI type based on a Kaggle dataset of 8675 social media posts	• Extreme gradient boosting • MBTI type removed from posts • 4 binary classes	I/E - .7901 N/S - .8606 F/T - .7178 J/P - .6570	Relatively small dataset. Authors discussed data imbalance but didn't seem to address it.
(6) Matej Gjurkovic, et al. 2018	Automated prediction of MBTI type based on Reddit dataset of 9600 users	• Support vector machine (SVM), Logistic Regression and three-layer Multilayer Perceptron (MLP) • Linguistic (LIWC), MRC Psycholinguistic Database, and user activity features • 4 binary classes	I/E - .818 N/S - .792 T/F - .672 J/P - .748	Methods to identify a person's type appear potentially error-prone.
(7) Mehul Nagpurkar, et al. 2018	Automated prediction of MBTI type based on Kaggle dataset of 8600 posts, Reddit dataset of 9600 posts, and Linkedin posts from 12,000 employees	• Logistic regression, SVM, Random Forest and Naive Bayes • SpaCy parts of speech tagger • Number of unique words divided by the total number of words • Average words per sentence • Average word length • NRC emotion lexicon	I/E - .5385 N/S - .8461 T/F - .3846 J/P - .5385	Did not remove Enneagram type from text and data leakage from this information was seen in results.
(9) Almar Sher Khan, et al. 2020	Automated prediction of MBTI type based on Kaggle dataset of 8600 posts	• XGBoost & comparison with other classifiers • Dataset resampling to correct for imbalance • Linguistic (LIWC) features • Age and gender features	I/E - .9937 N/S - .9992 T/F - .9455 J/P - .9553	No data cleaning or removal of MBTI or Enneagram information. Potential data leakage. Relatively small dataset.
This Study	Automated prediction of MBTI type for a final cleaned dataset of 1.293 million social media posts	• T5, BERT, CNN, SetFit, Linear Regression, SVC • Dataset resampling • Multi-class as well as 4 binary classes • Linguistic features through Empath, POSTAG, and NRC Emotion Lexicon • Removed MBTI and Enneagram type from posts	I/E - .98 N/S - .99 T/F - .98 J/P - .98 Multi-class - .93	The dataset is from a personality forum which may bias the post content. Did not use gender as a feature which contributes to lower results for T/F prediction.

References

1. Isabel Briggs Myers and Peter B. Myers, “Gifts Differing, Understanding Personality Type,” 1980

2. Clemens Stach, Florian Pargent, Sven Hilbert, Gabriella M Harar, Ramona Schoedel, Sumer Vaid, Samuel Gosling, and Markus Buhner, “Personality Research and Assessment in the Era of Machine Learning,” European Journal of Personality, pages 613 - 631, 2020

3. Sanja Stajner and Seren Yenikent, “Why is MBTI Personality Detection from Texts a Difficult Task,” Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3580 - 3589 April 19 - 23, 2021

4. Rayne Hernandez and Ian Scott Knight, “Predicting Myers-Briggs Type Indicator With Text Classification,” 31st Conference on Neural Information Processing Systems (NIPS 2017)

5. Mohammad Hossein Amitrhosseni and Hassen Kazemian, “Machine Learning Approach to Personality Type Prediction Based on the Myers-Briggs Type Indicator,” Multimodal Technologies and Interaction Journal, 2020

6. Mehul Nagpurkar, Text Analytics: What does your LinkedIn profile summary say about your personality? Using Natural Language Processing techniques to predict your personality, March 2, 2020

7. Matej Gjurkovic and Jan Snajder, “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, pages 87 - 97 New Orleans, Louisiana, June 6, 2018

8. Matej Gjurkovic, Mladen Karan, Iva Vukojevic, Milaela Bosnjak, and Jan Snajder, “PANDORA Talks: Personality and Demographics on Reddit,” Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, pages 138 - 152 Online Workshop, June 10, 2021

9. Almar Sher Khan, Hussain Ahmad, Muhammad Zubiar Ashghar, Furqan Khan Saddozai, Arreba Arif, Hassan Ali Khalid, “Personality Classification from Online Text using Machine Learning Approach,” International Journal of Advanced Computer Science and Applications, vol. 11 no. 3, 2020, pages 460 - 476

10. Sanja Stajner and Serin Yenikent, “How to Obtain Reliable Labels for MBTI Classification from Texts?” Proceedings of Recent Advances in Natural Language Processing pages 1360 - 1368 September 1 - 3, 2021

11. Francois Mairesse and Marilyn Walker, “Words Mark the Nerds: Computational Models of Personality Recognition through Language,” Proceedings of the Annual meeting of the Cognitive Science Society, 2006

12. John Clark, “Enneagram and MBTI Correlation,” Typology Central Wiki, November 30, 2015

13. T.A. Litinova and P.V. Serin, “Using Part-of-Speech Sequences Frequencies in a Text to Predict Author Personality: a Corpus Study”, Indian Journal of Science and Technology, Vol 8(S9), 93–97 May 2015

14. Ethan Fast, Binbin Chen, Michael S. Bernstein, “Empath: Understanding Topic Signals in Large-Scale Text”, Stanford University, 2016

15. Robert M. Capraro and Mary Margaret Capraro, “Myers-Briggs Type Indicator Score Reliability Across: Studies a Meta-Analytic Reliability Generalization Study”, Educational and Psychological Measurement, 2002

Git Repository

Our work can be found in the following git repository: https://github.com/cakennedy/266-mbti-project .

Summary of Team Responsibilities

Activity	**Responsible**
Original research, project proposal and overall design	J. Clark
Initial data ingestion and baseline	J. Clark
Data preprocessing and resampling	C. Kennedy
Feature Engineering, Experiments and model development	S. Joshi (BERT & SetFit), C. Kennedy (T5)
Error analysis & Benchmarking	S. Joshi, J. Clark
Report and presentation development	J. Clark, C. Kennedy, S. Joshi