Predicting Myers-Briggs Types with Cognitive Functions and Text Input

Serena Warner

Contents

1 Introduction

From sports betting to stock prices, prediction modeling with machine learning algorithms has been a hot topic in the data science and computer science industries for years. Due to machine learning's youth, many unexplored territories of prediction modeling that are currently being uncovered. One of those territories is personality type prediction.

While there has been a decent amount of research on personality type tests and theory validity, prediction modeling has yet to become a popular approach. In fact, the modeling that has been done has only touched the surface of most theories.

Personality theory can be utilized in countless ways [7]. For example, companies can use personality tests to both accommodate their employees and to decide whether an applicant would fit well in their environment. Additionally, personality theory can be used in academic settings; Some examples would be using a test as an icebreaker, which allows the instructor to get to know their students, as well as among their peers, and how they learn, or advisors having their advisee take a test to help them decide what career would fit them best. These examples illustrate a fraction of potential applications, as there are many possibilities to personality theory assisting everyday life [7].

These important practical uses for personality theory are what drive computer and data scientists to explore personality prediction using machine learning algorithms. The current research on one personality theory, the Myers-Briggs Type Indicator (MBTI), is stuck on a theory from about 80 years ago. The theory that Myers and Briggs brought to life as modern society known as the MBTI has been analyzed and expanded over the years, and new information deserves to be explored using machine learning.

This paper will use information about the MBTI and Jung's theories behind it to create multiple machine learning models which attempt to predict an individual's MBTI type using text input from social media. The results of the models will be compared as well as discussed comparatively to previous studies investigating MBTI prediction. This research aims to determine whether incorporating Jungian cognitive functions into the modeling process improves the performance metric of MBTI type prediction with text input compared to models that rely solely on the four-letter MBTI framework.

2 Myers-Briggs Type Indicator

2.1 Carl Jung's Theory

Carl Jung was a well-known and well-respected psychiatrist, psychotherapist, and psychologist who founded the school of analytical psychology. While this is one of his most popular achievements, he also created the theory behind one of the most popular personality theories in Volume 6: Psychological Types of his collected works. Myers and Briggs were inspired by the theory of Jung and created the MBTI.

Jung wrote that there are two basic types: extroverted and introverted [6]. These two types are split into two categories of functions: rational and irrational. Rational functions involve decision-making and value assignment without the collection of data, while irrational functions involve experiencing and gathering information without evaluation. The irrationals provide raw data, while the rationals evaluate that provided data. Jung identified the rational functions as feeling (F) and thinking (T) and the irrational functions as sensing (S) and intuition (N) [6].

2.2 Myers and Briggs' Work

Katherine Briggs and her daughter, Isabel Myers, were fascinated by personality theory, specifically in the realm of analyzing characters in literature [7]. As a result, in the early 1940s, they began to develop the Myers-Briggs Type Indicator, or MBTI, to test for personality type selection, heavily influenced by Carl Jung's theories. The Educational Testing Service evaluated the MBTI test and chose not to pursue the development due to multiple different issues with the test itself [7].

The sixteen MBTI are made up of four letters which have two dimensions each: extroverted(E)/introverted(I), intuition(N)/sensing(S), feeling(F)/thinking(T), and perceiving(P)/judging(J) [7]. The majority of these dimensions find their origin in Jungian theory. Extroversion, introversion, intuition, sensing, feeling, and thinking are all types that Jung spoke of in his book [6]. However, the last dimension, judging and perceiving, was created to determine if their irrational functions were dominant(P) or their rational functions were

dominant (J). These letters directly translate to functions from a model created by John Beebe [7].

2.3 John Beebe's Work

Formally published in 2004, John Beebe, a Jungian analyst, updated Carl Jung's work by proposing the 8-Function Model [1]. Beebe translated Jung's observed cognitive patterns into an analytical model. The 8 Function Model explicitly states that each of the functions, rational and irrational, can be internally focused, introverted, or externally focused, extroverted, as seen in Table 1. In total, as the name suggests, there are eight cognitive functions [1].

	Extroverted	Introverted
Feeling	Fe	Fi
Thinking	Te	Ti
Sensing	Se	Si
Intuition	Ne	Ni

Table 1: Jungian Functions

John Beebe popularized the idea of function stacks using his 8 Function Model. The first four functions in an individual's stack are deemed as the conscious functions [1, 6, 11]. To determine a stack, one must first look at the strongest and most conscious function, the dominant function. This is the lens the individual views the world. Next is the auxiliary function, the function that supports the dominant function. The tertiary function follows, which is a lesser developed function that emerges later in life and provides balance. Lastly is the inferior function. This is the weakest and least conscious function that is most associated with personal development [1, 6, 11].

The functions in a stack rely on each other [1]. If an individual's dominant function is an extroverted rational function, then their inferior function is going to be the opposite rational function with an introverted focus. This applies to irrational functions, as well. The auxiliary function and the tertiary function have the same property, however, if the dominant function is extroverted then the auxiliary function must be introverted and vice versa [1].

The reason for the name of the 8 Function Model is due to the conscious functions having the subconscious "shadow" functions, or the functions with the opposite focus [1]. These functions are not in an individual's stack.

For an example of a stack, an individual with a dominant function of extroverted intuition

(Ne), thus has an inferior function of introverted sensing (Si), and if their auxiliary function is introverted feeling (Fi), then their tertiary function is extroverted thinking (Te). This individual has a stack of NeFiTeSi. Notice the switching focuses and placement of irrational and rational functions [1]. Figure 1 shows all of the personality type stacks. The building process of a stack is incredibly important, as it will determine an individual's correlating MBTI [1].

2.4 Combined Theory

As seen in Figure 1, each MBTI has a specific function stack. There is a correlation between the letter representation and the functions behind the type.

To explain the types and their alignment, we will use the stack TeSiNeFi. The first letter of a type depends on the focus of the dominant function [1]. The dominant function is extroverted (Te) so the first letter will be E. Following, the second letter is determined by the stronger irrational function. The strongest irrational function is sensing, as it is the

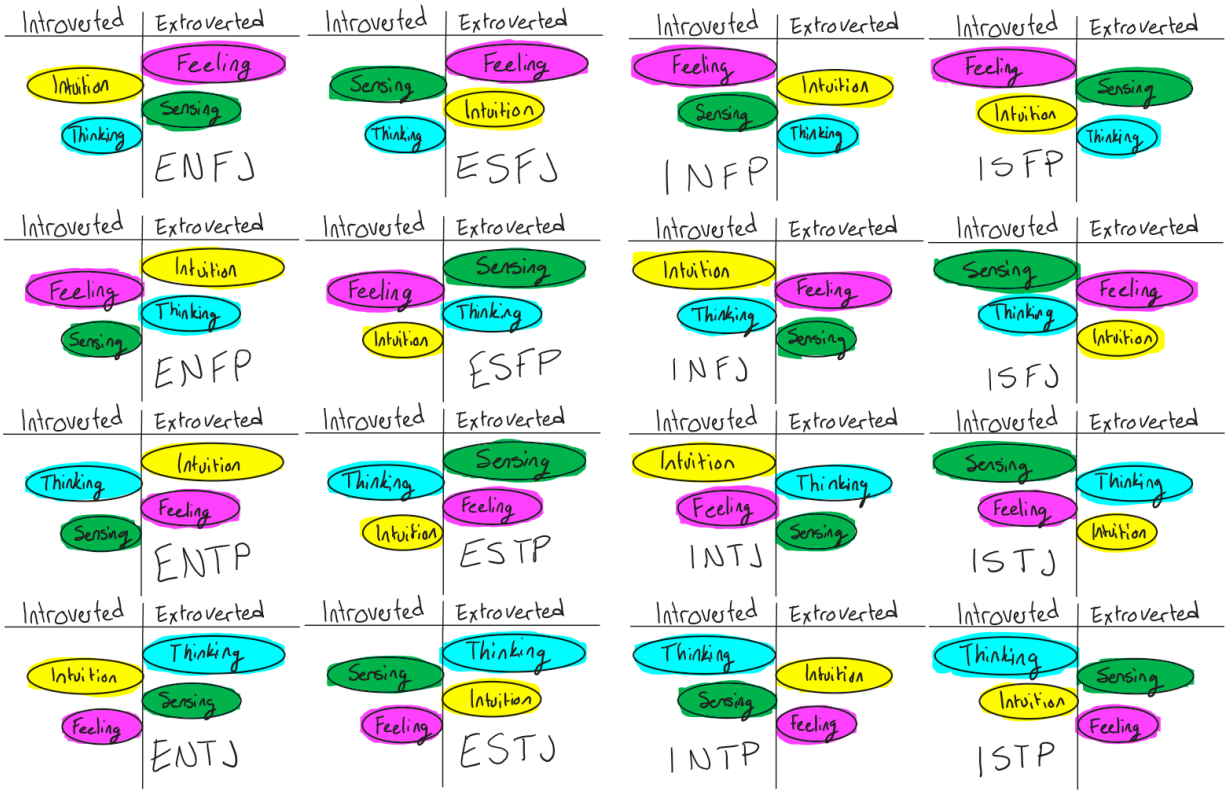

Figure 1: MBTI Types and Jungian Function Stacks

auxiliary function, so the next letter would be S. The third letter is similar to the second, using the rational function instead. The dominant function is thinking and therefore would be the next letter, T. The last letter gets a bit complicated. The previous letters cleanly align with some sort of function in the stack. Judging and perceiving are not functions in the stack. This letter is determined by the extroverted dominant or auxiliary function. Whichever function is extroverted, the last letter will align with either the irrational function (P) or the rational function (J). All together, the type that aligns with TeSiNeFi is ESTJ [1].

Many individuals believe that the theory of Carl Jung is best utilized in combination with both the MBTI and the function stacks.

2.5 Criticisms and Support

There have been a lot of conversations by personality psychologists on whether the MBTI should be taken as a scientific theory. Notably, individuals against the MBTI claim that since Jung's original work was philosophical and founded on anecdotal evidence rather than scientific evidence, it should not even be considered in scientific personality theory [7, 10]. Others argue that while Jung's theory was not scientific in basis, Myers and Briggs created a framework for identifying the types, however, the Myers-Briggs Type Indicator lacks important aspects of Jung's theory and over-simplifies it [14].

While there are many criticisms against the MBTI, the 8 Function Model's validity remains a topic of debate. While some question it's lack of empirical evidence, the model is a valuable tool, highlighting the ongoing tension between scientific rigor and practical utility in personality theory [2, 7, 10].

The advocates for MBTI propose using the cognitive functions in collaboration with the MBTI. This provides both a simple way to communicate the types, as well as including the complexity from Carl Jung's theories [2, 7].

3 Related Works

Previous studies have used a multitude of different algorithms to explore prediction modeling for the MBTI. Specifically, most research includes one major dataset that is available on Kaggle, (MBTI) Myers-Briggs Personality Type Dataset, which is a collection of social media posts from a website called personalitycafe.com. This website serves as a blog site for individuals interested in any personality theory, in which they can interact with others in the community. The dataset collects about nine thousand records of posts, and each record has about 50 posts written by the corresponding user. While the dataset has many instances, these are not equivalent across the 16 types.

Most studies focused on MBTI prediction chose to split each type into four dimensions, each representing each letter of the MBTI. This means for most studies, there were four models for each method used. Most studies looked at the performance metrics of each dimension, and some additionally added an average for the method used across all dimensions.

Table 2 shows a few previous studies that showed high promise in performance metrics. These will be expanded on.

Study	Method(s)
Vaddem and Agarwal (2020)	Logistic Regression, Support Vector Machine, and XGBoost
Ryan and et al. (2023)	Logistic Regression, Linear Support Vector Classification, Stochastic Gradient Descent, Random Forest, XGBoost, and CatBoost
Sánchez-Fernández and et al. (2023)	K-Means Clustering, BIRCH, ANN, FFNN, Logistic Regression, LSTM, Naïve Bayes, and RNN

Table 2: Previous Studies and Methods Used

Vaddem and Agarwal (2020) explore three algorithms: logistic regression, support vector machine, and XGBoost [13]. They split their data on the four MBTI dimensions. This allowed for binary modeling for each dimension. With four models for each algorithm and no

averages, their best-performing method was logistic regression with an accuracy percentage of 79.07% in the I/E dimension, 86.68% in the N/S dimension, 77.87% in the F/T dimension, and 67.17% in the J/P dimension. These metrics show high promise with logistic regression. However, across all three models, the J/P dimension was lower than the other three dimensions by at least 8%. This is interesting to consider when reflecting on the history of the theory, especially how the fourth dimension was created and ties to the original theory [13].

Ryan et al. (2023) explore various supervised algorithms [8]. This study also chose to use the four dimensions of the MBTI. Additionally, the study compares F1 scores with and without utilizing the Synthetic Minority Oversampling Technique, or SMOTE. This technique allowed for the unbalanced data to be synthetically balanced, which prevents a biased prediction towards the more common types in the data. Their best-performing method was logistic regression with an F1 score without SMOTE of 0.8282 and an F1 score with SMOTE of 0.8337. This suggests utilization of SMOTE may benefit analyzation, however, not very much, as well as confirming once again that logistic regression may be a high-performing method [8].

Sánchez-Fernández et al. (2023) explore multiple different neural networks, unsupervised algorithms, logistic regression, and Naïve Bayes [12]. Their models aim to predict the exact personality type for each post, with 16 dimensions, rather than the four-letter dimensions. Their best modeling results were an accuracy of 0.9274 with Naïve Bayes and 0.9100 with logistic regression. Once again, this furthers that logistic regression performs well in MBTI prediction. An interesting aspect of this research is the surprisingly high prediction performance metric compared to the four dimension studies. This suggests there may be a benefit in exploring different breakdowns of the personality types [12].

Myers and Briggs' contributions to the theory of Jung have dominated the MBTI prediction modeling research done thus far. Therefore, the tie to John Beebe's expansion of Jungian theory has not yet been investigated. The research mentioned above suggests that there may be a benefit in exploring the functions from the 8-Function Model, rather than the four-letter dimensions of Myers and Briggs.

4 Data

The data that is being used for this research is sourced from Kaggle, titled "MBTI Personality Types 500 Dataset." This dataset was created by Zeyad Khalid, and last updated in 2022. Each instance is collected from a website called personalitycafe.com, an online forum website for people interested in personality theory to come together and discuss theories. Each piece of data is a collection of words from posts by a single individual alongside their MBTI type. This dataset comes pre-processed, however, does not exclude MBTI-related terms such as the MBTI types or functions.

4.1 Data Manipulation

In order to use the cognitive functions, we must first add new classifications for each instance. In total, there are six new classifications, which will create six models for each algorithm. We use RStudio, with the programming language R, to achieve this goal.

The first category of classifications is the dominant and auxiliary functions, with eight classes for each model. Each type has a unique dominant function and auxiliary function combination. It is unnecessary to add models for the tertiary and inferior functions, as whatever dominant and auxiliary function determines that for each type. As Table 3 demonstrates, some types share dominant functions that would not be able to be solely represented by a letter breakdown. For example, ENFP and ENTP share a dominant Ne function, which means that the similarity is based on the shared three letters, "ENxP," rather than on one letter. This insinuates that we may have a benefit in using this breakdown.

Type	Dominant	Auxiliary	Type	Dominant	Auxiliary
ENFP	Ne	Fi	INFP	Fi	Ne
ENTP	Ne	Ti	INTP	Ti	Ne
ESFP	Se	Fi	ISFP	Fi	Se
ESTP	Se	Ti	ISTP	Ti	Se
ENFJ	Fe	Ni	INFJ	Ni	Fe
ENTJ	Te	Ni	INTJ	Ni	Te
ESFJ	Fe	Si	ISFJ	Si	Fe
ESTJ	Te	Si	ISTJ	Si	Te

Table 3: Dominant and Auxiliary Functions

The second category of classifications is the rational and irrational dominant functions,

with eight classes for each model. This is similar to the dominant and auxiliary functions, however, extroverted and introverted functions will share the same classifications. The same idea with the dominant and auxiliary functions, there is no need to include the inferior rational and irrational functions, as it would be predetermined by the dominants. Table 4 shows the breakdown for each type. One can see that a group of types, such as ESFP, ISFP, ESTP, and ISTP share the same irrational dominant function. This will investigate whether the location of the dominant rational or irrational functions fundamentally changes the type's textual expression.

Types	Rational	Irrational
ENFP & INFP	Fi	Ne
ENTP & INTP	Ti	Ne
ESFP & ISFP	Fi	Se
ESTP & ISTP	Ti	Se
ENFJ & INFJ	Fe	Ni
ENTJ & INTJ	Te	Ni
ESFJ & ISFJ	Fe	Si
ESTJ & ISTJ	Te	Si

Table 4: Dominant Rational and Irrational Functions

The third and final category are the rational combinations and irrational combinations, with binary classifications for each model. There are two different rational combinations, FeTi and TeFi, and two different irrational combinations, NeSi and SeNi, shown breakdown is shown in Table 5. As one can see, types such as ENTP and ISFJ, which share no letters in common, can share functions. This will allow for more nuance than previous research has allowed.

Types	Rational	Irrational
ENTP & INTP & ESFJ & ISFJ	FeTi	NeSi
ENFJ & INFJ & ESTP & ISTP	FeTi	SeNi
ENFP & INFP & ESTJ & ISTJ	TeFi	NeSi
ENTJ & INTJ & ESFP & ISFP	TeFi	SeNi

Table 5: Rational and Irrational Function Combinations

5 Machine Learning Algorithms

5.1 Logistic Regression

Logistic regression was one of the most successful algorithms across multiple studies with MBTI prediction modeling [8, 12, 13]. This is not surprising, as logistic regression is a very

effective method to classify binary data, and most of the previous research makes each type into four binary dimensions.

Logistic regression is a statistical approach to machine learning. The goal of the model is to predict the probability of an event occurring based on a logistic curve. Logistic regression has two types of models: binary logistic regression and multinomial logistic regression [8]. This research uses both types of models for different data breakdowns of each type.

Binary logistic regression uses a logistic curve to split the data points and defines a decision boundary, as seen in the example in Figure 2. This allows for the classification of the data to either class 0 or class 1 [4, 5]. In this research, the binary classes will be the focus of the irrational functions and the focus of the rational functions, however, for the other models, this will not work.

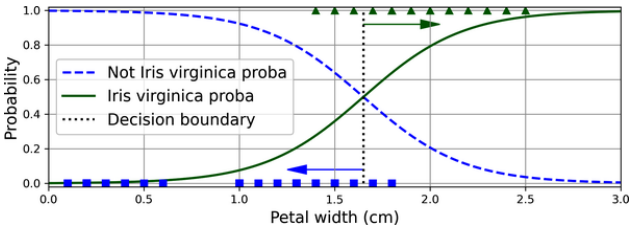

Figure 2: Estimated Probabilities and Decision Boundary for Petal Classification [4]

Multinomial logistic regression is used when a classification has more than 2 classes, which is utilized for the majority of models in this research. For each instance, multinomial logistic regression first calculates a score for each class. The softmax function is then applied, which estimates the probability of each class [4, 5]. This allows for the prediction of dominant and auxiliary functions as well as irrational and rational functions. An example of how classes are classified using multinomial logistic regression is in Figure 3.

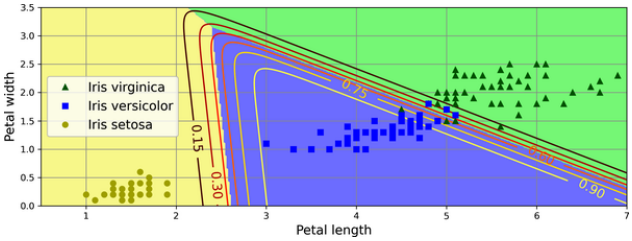

Figure 3: Multinomial Regression Decision Boundaries for Petal Classification [4]

5.2 Support Vector Machine

Support vector machine (SVM) is an algorithm that also utilizes linear separation [4, 9]. Figure 4 shows data being classified using a linear average classifier. Some points are misclassified due to the data being below the average line. The SVM finds the support

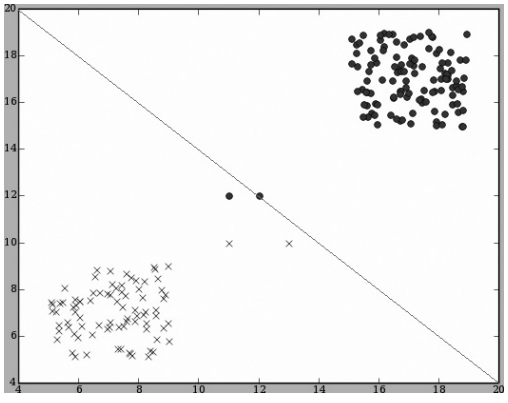

Figure 4: Linear Average Classifier [4]

vectors which are chosen based on their proximity to the other class. These vectors will be used to find lines parallel to the support vectors and thus will place the dividing line in between them, as seen in Figure 5 [4, 9].

These models can be altered to include non-linear lines using kernels. Kernels are mapping functions that return the data transformed into a higher dimensional space [9]. This allows for more complex data, which is nonlinearly separable, to be classified [4, 9]. Common kernel functions include the linear kernel, radial basis function (RBF) kernel, and polynomial kernel. The choice of kernel and parameters for the kernel can impact the model's performance immensely [4, 9].

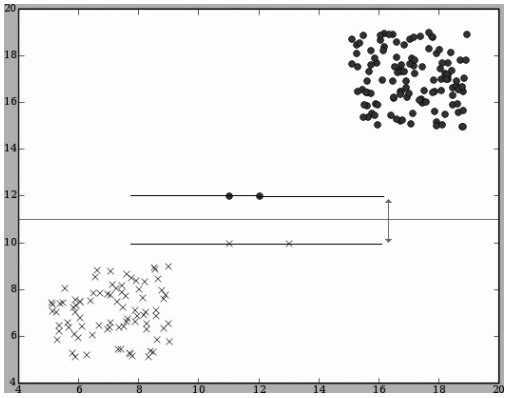

Figure 5: Support Vectors [4]

SVMs can be used in multi-class situations, similar to logistic regression [3]. There are two types of multi-class classification for SVMs: One-Versus-One (OVO) and One-Versus-ALL (OVA). For each class, OVO-SVMs construct binary SVM classifiers for each distinct pair of classes. Each classifier assigns one class positive and the other negative. To combine each classifier, Max-Wins Voting (MWV) is commonly used. To classify new instances, MWV determines which class has the largest votes from each of the binary classifiers and assigns the instance as such. For OVA-SVMs, one class is chosen to be our comparative class, and binary models are created for each of the remaining classes against the chosen comparative class. In this case, Winner Takes All (WTA) and whichever class has the highest decision function value among all classifiers [3].

6 Methodology

As stated previously in Section 4, the data comes partially pre-processed, however, some posts include MBTI-related verbiage. Thus, all terms referencing the types or cognitive functions were removed. As an extra measure, we additionally process the data using the Natural Language Toolkit (NLTK) Python library and regular expressions, excluding stopwords, special characters, and any extra spaces. Lastly, we will lemmatize all words, which

reduces each word to its base or dictionary form (lemma), to standardize all words. This is done using the NLTK library, as well.

In order to transform the words into information that the machine learning algorithms can understand, we use Scikit-Learn's TfidfVectorizer. This method calculates Term Frequency-Inverse Document Frequency (TF-IDF) scores to represent each word in the dataset. No parameter alterations were made to the TfidfVectorizer object.

6.1 Models

To be able to compare our models with previously done studies, we are investigating both the accuracy score and F1 score with the average set to micro, which prevents the favoring of any class.

We have twelve models in total, six logistic regression models and six SVM models. This means we have one model for each algorithm to classify the six new classifications. Each model's data was split into training and testing sets using an 80/20 ratio.

While modeling, we test different variations to our models to see if the performance measures increase. We explore different options for standardizing the sample size, such as undersampling and the Synthetic Minority Over-sampling Technique (SMOTE). These do not yield performance improvements and thus are not used. Truncated SVD, a technique that reduces linear dimensionality, is also tested, however, it decreased model performance across all classification tasks.

7 Results

Both performance metrics had almost identical results, thus, Table 6 shows the F1 scores for each classification with logistic regression (LR) and SVM models. The SVM models performed slightly higher than the logistic regression models consistently. As one can see, the highest-performing models were the rational and irrational function combinations, our bi-

Classification	LR	SVM
Dominant	0.73	0.75
Auxiliary	0.73	0.74
Rational Dominant	0.78	0.79
Irrational Dominant	0.77	0.78
Rationals	0.78	0.79
Irrationals	0.79	0.80

Table 6: Model F1 Scores

nary models. To complement the performance metrics in Table 6, Figures 6-9 show the confusion matrices for the SVM models.

Figure 6: SVM - Rational Dominant Functions

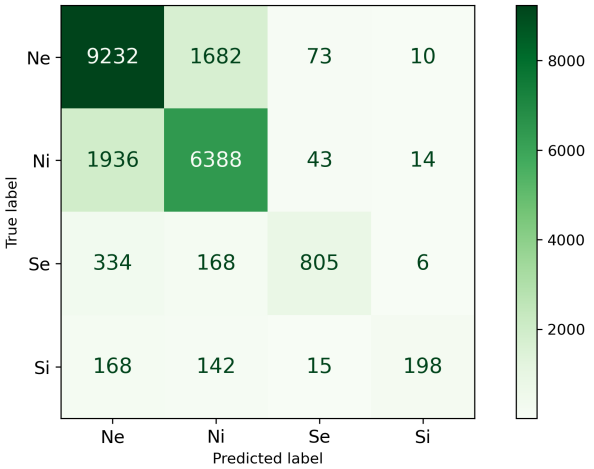

Figure 7: SVM - Irrational Dominant Functions

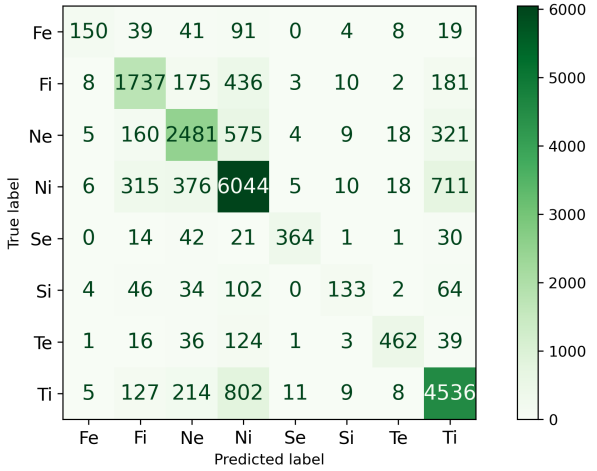

Figure 8: SVM - Dominant Functions

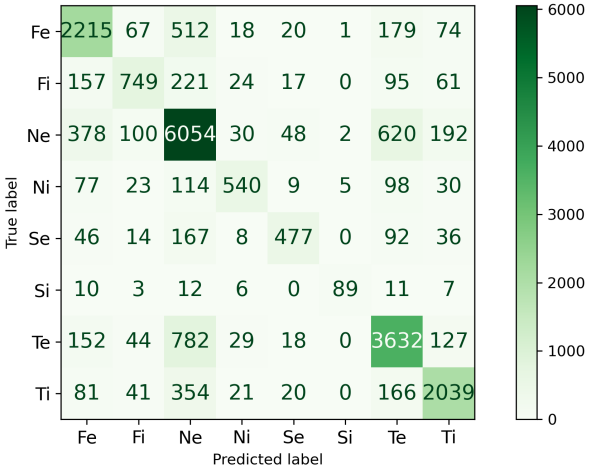

Figure 9: SVM - Auxiliary Functions

As one can see from Figures 6 and 7, the models are not performing well for some functions more so than others. For some more insight, misclassification rates can be explored. In Figure 6, Fe has the highest misclassification rate with 29% mistyped. On the other hand, Ti has the lowest misclassification rate of 14%. In Figure 7, one can see that the model mistyped Ne users as Ni users and vice versa, however, Ne has the lowest misclassification rate of 16%, Ni with the second lowest of 0.24. Se and Si were not well identified, having misclassification

rates of 0.39 and 0.62 respectively.

For the dominant and auxiliary classifications, it is clear that only certain types were correctly typed. In Figure 8, the lowest misclassification rate is 19% with Ni, however, in Figure 9, the lowest misclassification rate is 18% with Ne. In comparison, the most misclassified dominant function is Si with 65% mistyped, and the most misclassified auxiliary function is Fi with 43% mistyped.

8 Discussion

As the most notable pattern, the dominant and auxiliary models did not perform as well as the other models with a 0.04-0.06 performance difference. These specific models have the most number of classes compared to the other models, and additionally is the most specific to each type, as it depends on the stack hierarchy. As seen in Table 3 previously, only two MBTI types share the same dominant or auxiliary function. The weak performance may be due to the complexity of the function hierarchy specifically. According to Jung's typological model, the auxiliary function provides balance to the dominant, particularly when the dominant is introverted and the auxiliary is extroverted [6]. This additionally suggests that types sharing dominant rational or irrational functions may exhibit more similar linguistic patterns, regardless of their full stack structure.

There is a fascinating pattern between the confusion matrices for these models, as seen in Figure 8 and 9. In the dominant functions, the lowest mistyped function are Ni followed by Ne, however, in the auxiliary functions, the lowest mistyped functions are Ne followed by Ni. First, this suggests that the types with the most unique verbiage are intuitive types, regardless of their expression. This aligns with the location of data collection, as more individuals who are prone to speculation or brainstorming different ideas, such as typology, tend to be intuitive types [1, 6, 7]. This aligns with previous research as the N/S dimension is the highest performing model across all studies [8, 13].

There are many additional interesting patterns that are highlighted specifically by the confusion matrices. Consistently across all models, the intuitive and thinking cognitive functions were confused with the different expressions, as well as with each other. For

example, in Figure 8, the most identified function is Ni followed by Ti. If one looks at the intersection where these functions were mistyped as each other, one can see that these numbers are higher than all other mistypes and even larger than some of the correctly predicted functions. This may be due to the similarities in how these functions express themselves through text. Overall, Ni and Ti are very similar, as they both are introspective on ideas, however have different goals. This could be why the model confused the two.

This pattern follows into Figure 9 with Te and Ne. While Te and Ne are not as similar as Ti and Ni, it could be that intuition and thinking both appear similarly through the chosen words of the individuals with those functions at the top of their stack.

It is to be noted that the dominant function model more often correctly predicts introverted functions while the auxiliary function model more often correctly predicts extroverted functions. This could be due to the relation of the dominant function and the auxiliary function, meaning that it was identifying types with NiTe (INTJ) and TiNe (INTP) the strongest. This could also be the result of introverted types being the most introspective individuals, which would lead them to have similar verbiage while chatting online about typology.

There are also some concerns about how the individuals typed themselves. Some may have typed themselves using Myers and Briggs' theory, while others may have used John Beebe's theory. Some individuals may have gotten formally typed by a professional, and some may have taken online tests. The reliability of the classifications is not very high, which could be why models using this dataset have not been able to perform as high as it could.

Future research could investigate some of these discrepancies through different testing data, as word choice may not be clear enough to define an individual's exact function stack. The data is also scaled to exclusively individuals on this site, who are interested in talking with others about typology. By collecting data independently, one could ensure that the types are collected by the same source. Additionally, future research may look into unsupervised learning and mapping or analyzing how the results could be related to different personality theories, such as MBTI.

9 Conclusion

This study aimed to analyze the results of logistic regression and support vector machine models trained to predict the cognitive functions based on John Beebe's 8-Function model built from Carl Jung's personality theories. The models performed relatively well, with F1 scores and accuracy scores ranging from 0.73 to 0.80 and the SVM models slightly outperforming the logistic regression models. Future research will investigate unsupervised models trained on survey data, in which the collections will be compared to different personality theories and potentially mapped to the cognitive functions.

References

[1] J. Beebe. *Energies and Patterns in Psychological Type*, chapter Understanding Consciousness Through the Theory of Psychological Types. Routledge, 2016.

[2] L. V. Berens and D. Nardi. *Understanding Yourself and Others: An Introduction to the Personality Type Code*. Telos Publishing, 1998.

[3] K.-B. Duan, J. C. Rajapakse, and M. N. Nguyen. Evolutionary computation,machine learning and data mining in bioinformatics. In *One-Versus-One and One-Versus-All Multiclass SVM-RFE for Gene Selection in Cancer Classification*, 2007.

[4] A. Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition*. O'Reilly Media, Inc., 2022.

[5] D. W. H. Jr., S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression, 3rd Edition*. Wiley, 2013.

[6] C. Jung. *Collected Works of C.G. Jung, Volume 6: Psychological Types*. Princeton University Press, 1971.

[7] D. Pittenger. The Utility of the Myers-Briggs Type Indicator. *Review of Educational Research*, 63(4):467–488, 1993.

[8] G. Ryan, P. Katarina, and D. Suhartono. MBTI Personality Prediction Using Machine Learning and SMOTE for Balancing Data Based on Statement Sentences. *Information*, 2023.

[9] T. Segaran. *Programming Collective Intelligence*. O'Reilly Media, Inc., 2007.

[10] R. Stein and A. B. Swan. Evaluating the validity of Myers-Briggs Type Indicator theory: A teaching tool and window into intuitive psychology. *Social and Personality Psychology Compass*, 13, 2019.

[11] N. Sumaiya, E. Promitee, and M. Khan. Understanding Human Learning and Decision Making Process based on Jungian Analytical Psychology by incorporating Python and Data Science. Master's thesis, Islamic University of Technology, 2021.

[12] P. Sánchez-Fernández, L. G. Baca Ruiz, and M. del Carmen Pegalajar Jiménez. Application of classical and advanced machine learning models to predict personality on social media. *Expert Systems with Applications*, 216:119498, 2023.

[13] N. Vaddem and P. Agarwal. Myers briggs personality prediction using machine learning techniques. *International Journal of Computer Applications*, 2020.

[14] D. J. Wilde. *Jung's personality theory quantified*. Springer London, 2011.