Myers-Briggs Personality Classification: Using Post Content to Determine Personality Type

Chris O'Brien
Bredesen Center
University of Tennessee
Knoxville, TN USA
cobrien8@vols.utk.edu

Carrie Sanford
Bredesen Center
University of Tennessee
Knoxville, TN USA
csanfor5@vols.utk.edu

Jay Pike
Bredesen Center
University of Tennessee
Knoxville, TN USA
jpike1@vols.utk.edu

***Abstract*— The Myers-Briggs Type Indicator (MBTI) is a test that assigns one of sixteen personality types based on four dimensions: Introversion (I) vs. Extroversion (E), Intuition (N) vs. Sensing (S), Feeling (F) vs. Thinking (T), and Judging (J) vs. Perceiving (P). MBTI tests are used for making compatible groups in professional and educational settings, but these evaluations can be time consuming. Machine learning can be used to remove the need for the full MBTI evaluation. In this work, machine learning methods were evaluated for their ability to predict MBTI types using forum posts. Term-frequency inverse document frequency was used to extract text data. The prediction capabilities of models using logistic regression, Linear Support Vector Classification (LinearSVM), and XGBoost algorithms were used to classify personality types (N=16) and dimension (N=2, 4 times). The most accurate model for predicting personality type was LinearSVM with a 42% accuracy for 16-class classification and 87% and 78% accuracy for N vs. S and F vs. T, respectively. Logistic regression models were most accurate at predicting I vs. E (78%) and J vs. P (70%). The dataset used was neither uniformly distributed nor representative of the general population which had significant impacts on classification accuracy. Classification accuracy was observed to be positively correlated with class distribution in the data. Datasets with better distributed data may result in models with improved MBTI predictions.**

Keywords — Myers-Briggs, machine learning, logistic regression, XGBoost, Linear SVC, Random Forest

I. Introduction

Psychoanalyst Carl Jung hypothesized that behavior differences amongst the population are not random, but the result of how people use their perception and judgement [1]. Katherine Cook Briggs became interested in the diverse ways people respond to the world around them when she first met her daughter Isabel Briggs Myers future husband. Myers and Briggs came upon Jung's theory and wanted to make it more accessible to the public as World War II was currently happening. They believed that if people understood different personality types, they could learn how to interact with one another better and have less conflict. It took Myers and Briggs 20 years to develop the first version of the Myers-Briggs Type Indicator (MBTI), published in 1962 [8].

The MBTI assigns personality types based on responses to situational or relational questions. There are 16 different personality types with four dimensions. The first dimension is where one puts their attention, are they focused on their own thoughts and ideas (Introversion; I) or on the people and things around them (Extroversion; E). The next dimension is how one receives and uses information. Are they hands on, picking up information through their senses (Sensing; S)? Or do they use patterns and theories to deeper understand meaning behind information (Intuition; N)? The third dimension is how one makes decisions. Do they use objective information and logic (Thinking; T)? Or do they use personal views and impact of those involved (Feeling; F)? The last dimension is how one goes about their daily life. Do they prefer structure, planning, and organization (Judging; J)? Or do they prefer flexibility and adaptability (Perceiving; P) [1]? For example, if someone is "tolerant and flexible, quiet observers until a problem appears, then act quickly to find workable solutions," they would be described as having a ISTP personality.

The Myers-Briggs parent company states the purpose of the MBTI is to "help people understand personality differences in the general population" to improve their interactions with others [2]. The MBTI system is commonly used in schools or workplaces to create effective combinations of personality types when forming collaborative groups. The purpose of MBTI in these settings is to improve workplace efficiency. Teams made of individuals with compatible personality types are typically more productive. There are currently four different formats of the MBTI system: Form M (computer-based or self-scoring) with 92 questions, Step II Form Q with 144 questions, and Step III with 222 questions. Steps II and III provide more in-depth analyses than the 16 personality types. These tests include five facets for each dimension and are administered by professionals [1]. The testing procedure for all forms is time consuming. A successful machine learning model to predict personality types would significantly reduce the time needed to form collaborative groups.

II. Related Work

The use of machine learning models to predict MBTI personality types using text is reported in literature. Text data used for MBTI classification has included long format writing such as essays, variable length social media posts, and tweets [4 – 7]. Multiple lexical resources and classification methods have been tested. Linguistic Inquiry and Word Count (LIWC) uses 68 dictionaries that correspond to concepts or psychological constructs like various emotions or money. The method uses vector counts of words corresponding to each concept. EmoSenticNet assigns one of six affect emotion labels, while ConceptNet uses a graph of concepts and related phrases. Other lexicon methods like Mairesse, SenticNet, NRC Emotion Lexicon, and VAD Lexicon have also been tested. Support Vector Machines (SVM) use the distance between hyperplanes and the closest data points for classification. Naïve Bayes classification is based on probability, using class probabilities (class instances out of total instances) and conditional probabilities (instances of attribute values within a class value out of total instances of that class). Deep learning methods like BERT, Albert, and Robert have also been used with text for personality type predictions. The model with the reported greatest accuracy (88%) used LIWC, EmoSenticNet, and ConceptNet with tweets [5]. Deep learning methods with longer form social media posts and essays had lower accuracies than the short form tweets, ranging from 38% - 77% [6,7]. In this work, additional machine learning methods were tested to improve the prediction accuracy of MBTI personality types using social media posts.

III. Data/Methods

A. Data

The data for this work was obtained from Kaggle collected through PersonalityCafe [3]. The data contains MBTI types of 8,765 people and their last 50 posts on PersonalityCafe, which is "a forum community dedicated to all ranges of personality types and people." Posts are separated by '|||' and range from strictly text to hyperlinks to GIFs. The dataset does not have an equal distribution of personality types (Fig. 1).

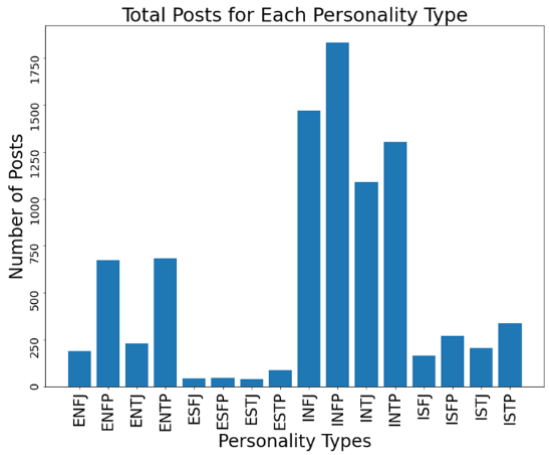

Fig. 1. Distribution of personality types in dataset by number of posts.

Introverted and intuition types (INTJ, INTP, INFP, INFJ) have the most occurrences with over 1000 datapoints each. Extroverted and sensing types (ESTJ, ESTP, ESFJ, ESFP) each contain less than 200 datapoints. Extroverted, intuition, and perceiving types (ENFP, ENTP) have around 700 occurrences each while extroverted, intuition, and judging types (ENFJ, ENTJ) and introverted and sensing types (ISTJ, ISTP, ISFJ, ISFP) occur between two and four hundred times. Within the dimensions, there are three times more posts from introverts than extroverts, eight times more posts from intuitive than sensing types, and 1.5 times more posts from perceiving than judging types. The distribution of personality types within the dataset is not representative of the distribution in the general population (Fig. 2).

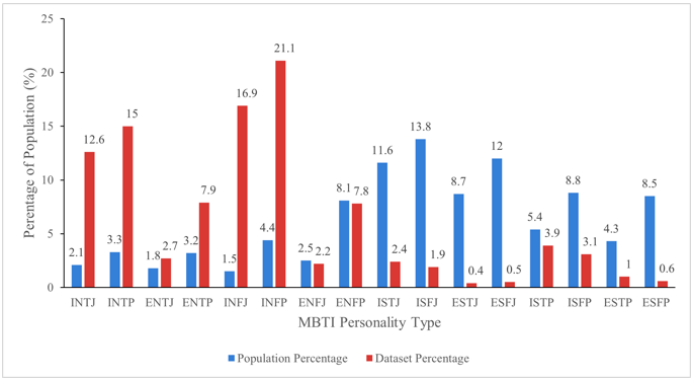

Fig. 2. Proportion of personality types in general population and in the Kaggle dataset.

For example, INFP personality types make up 21.1% of the dataset, but only 4.4% of the population. Meanwhile, the most common personality types, ISFJ and ESFJ, representing 13.8% and 12% of the general population only make up 1.9% and 0.9% of the dataset, respectively. There is not a significant difference in post length between the personality types.

B. Methods

Text data was converted to all lowercase letters and links, MBTI personality labels, nonwords, punctuation, and extra spaces were removed. These cleaning techniques reduced the average words per post from 25 to 12. The personality type labels were encoded to be an integer from zero to fifteen. Term Frequency-Inverse Document Frequency (TF-IDF) was used for feature extraction to convert the text data into machine learning compatible features. TF-IDF is based on the Bag of Words method and weights features or phrases directly to its frequency within each document and inversely to its frequency within the complete set of documents. Words occurring less than three times and common English stop words were removed. Unigrams and bigrams were extracted. The data set was then split into training (70%), validating (15%), and testing (15%) sets.

Two classification problems were tested: prediction of personality type (N=16) and prediction of each of the four dimensions (N = 2). Classifying across dimension required four separate 2-class classification problems to be solved. The N=16 class classification problem was the initial focus of this project and the purpose of the Kaggle competition where the

dataset was obtained. Logistic regression was used as the baseline model. The other algorithms evaluated were Linear SVM, Random Forest, and XGBoost. Linear SVM works well with large multi-class datasets and uses a N-1 dimension hyperplane to classify the data [5]. Random Forest Classification is useful when there are numerous input variables and is one of the most accurate machine learning algorithms, using the average predication of a group of decision trees (the forest). Random Forest corrects for overfitting on the training set as well. XGBoost uses gradient boosting which trains models successively so that each model learns from the error of the previous model until the final model can no longer be improved.

Hyperparameter optimization was done for each model to optimize its classification accuracy. A total of 20 models were trained and underwent hyperparameter tuning. Table 1 summarizes the hyperparameter tuning efforts.

TABLE 1: HYPERPARAMETER TUNING

Algorithm	Parameter	Variables
Logistic Regression	0.4155	Liblinear
	Penalty	L2, L1
	C values	0.0001, 0.001, 0.01, 0.1, 1, 10, 100
Linear SVM	Penalty	L2, L1
	C values	0.0001, 0.001, 0.01, 0.1, 1, 10, 100
Random Forest	# of Estimators	100, 1000
	Max depth	5, 10
	Min samples split	2, 5
	Min samples split leaf	1, 2
	Bootstrap	True, False
XGBoost	Subsample	0.5
	Subsample	5, 10
	Eta	0.05, 0.1

With the tuned parameters, the models were fit using the training dataset and the prediction capabilities determined with the testing set. Training and testing times were measured. Classification reports were generated for each model evaluating precision, recall, F1-score, support, and accuracy. Additionally, our models were evaluated through examining confusion matrices, Receiver Operating Characteristics (ROC) curves and Area Under the Curve (AUC) values.

IV. RESULTS AND DISCUSSION

Two classification problems were evaluated to test the ability to predict Myers-Briggs personality types using forum. Model parameters were tuned separately for N=16 classification and N=2 classifications. The resulting parameters (Table 1) were used when training each model.

TABLE 2: CLASSIFICATION REPORT FOR N=16 MODELS

Algorithm	Accuracy	Train Time (s)	Test Time (s)
Logistic Regression	0.4155	**13.86**	**0.0221**
Linear SVM	**0.4237**	19.97	0.0237
Random Forest	0.2919	25.39	0.4674
XGBoost	0.3648	186.8	0.1320

As shown in Table 2, the model that performed best in classifying the 16 Myers-Briggs types was the Linear SVM model. SVM was also observed to have the best performance by Bharadwaj et al. It was initially assumed that Linear SVM would have low performance due to an expected high number of margin violations; however, the methods which performed better (Logistic Regression and Linear SVM) were more susceptible to overfitting than other algorithms with worse performance (Random Forest and XGBoost). Overfitting is a negative trait for a classifier, so the higher classification accuracies observed for Logistic Regression and SVM are artifacts of the skewed class distribution.

The results of the two best performing algorithms for 16-class classification were explored using confusion matrices. In Figures 3 and 4, the confusion matrices from Logistic Regression and Linear SVM are shown. It is evident that the four most common predictions coincided with the four majority types within the dataset. The dataset consisted of mainly INFJ, INFP, INTJ, or INTP personality types, and the four most predicted were of the same type. Additionally, almost half of the personality types, ENFJ, ENTJ, ESFJ, ESFP, ESTJ, ESTP, and ISFJ, were not predicted during testing which corresponded to the seven least frequent personality types in the dataset. Therefore, as expected, the class imbalance had a large effect on the 16-class classification results.

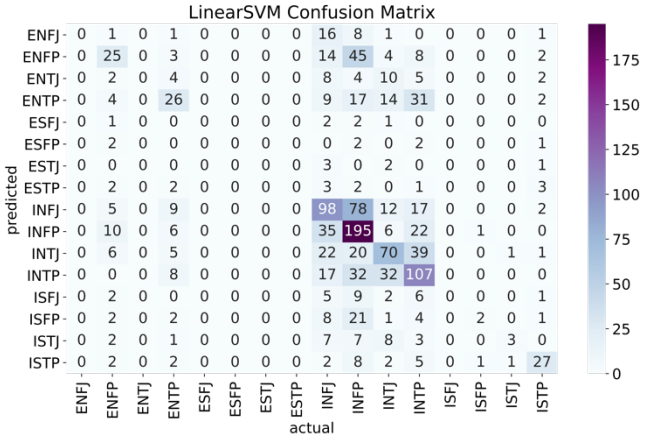

Fig. 3. Confusion matrix for LinearSVM model for N=16 classification. Actual Myers-Briggs types included on x-axis and model predictions on y-axis.

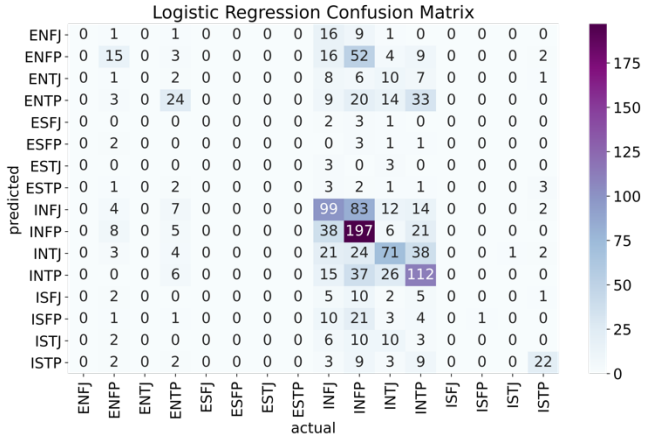

Fig. 4. Confusion matrix for LinearSVM model for N=16 classification. Actual Myers-Briggs types included on x-axis and model predictions on y-axis.

The ROC curves for 16-class classification were plotted to show the tradeoff between sensitivity and specificity. The ROC curve for Logistic Regression was included to show that the predictions have both poor sensitivity and specificity. Due to the high number of classes the ROC curve appeared crowded and was deemed to be less effective in comparison to a confusion matrix.

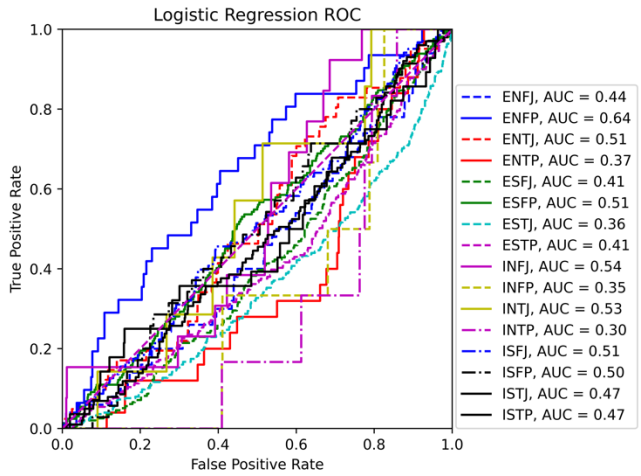

Fig. 5. Receiver Operating Charateristics curve for Logistic Regression model for N=16 classification.

Across the Extroversion/Introversion dimension (I/E) the best performing algorithm was Logistic Regression with a classification accuracy of 78.42% (Table 3). The other models all had the same classification accuracy of 76.96%, and thus it was determined that they predicted that all datapoints belonged to Introversion class. This result was due to the severely skewed class distribution.

TABLE 3: INTROVERSION/EXTROVERSION CLASSIFICATION REPORT

Algorithm	Accuracy	Train Time (s)	Test Time (s)
Logistic Regression	**0.7842**	1.032	0.0022
Linear SVM	0.7696	**0.1627**	**0.0021**
Random Forest	0.7696	1.551	0.3694
XGBoost	0.7696	38.37	0.1381
Fraction of I in Dataset	0.7696	N/A	N/A

As shown in Table 4, comparable results were found with the Intuition/Sensing (N/S) class where all the models had a close classification accuracy. The same phenomenon occurred where most of the predictions were Intuition which was the majority of the datapoints in the dataset. The Linear SVM had the best performance at 87.02%, however, this is only 0.82% better than classifying all datapoints as Intuition.

TABLE 4: INTUITION/SENSING CLASSIFICATION REPORT

Algorithm	Accuracy	Train Time (s)	Test Time (s)
Logistic Regression	0.8648	2.040	**0.0026**
Linear SVM	**0.8702**	7.256	0.0027
Random Forest	0.8618	**1.806**	0.3587
XGBoost	0.8679	32.20	0.1341
Fraction of N in Dataset	0.8620	N/A	N/A

Logistic Regression produced a 79.26% classification accuracy across the Thinking/Feeling (T/F) dimension (Table 5). Whereas the I/E and N/S dimensions' classification algorithms produced results which were almost equivalent to the baseline of the ratio of the majority dimension in the dataset, the classification algorithm for T/F is 25.15% better than its baseline. This result may be because the distribution of T/F dimensions within the dataset was much more uniform.

TABLE 5: THINKING/FEELING CLASSIFICATION REPORT

Algorithm	Accuracy	Train Time (s)	Test Time (s)
Logistic Regression	0.7926	**0.9804**	**0.0022**
Linear SVM	**0.7811**	2.0433	0.0023
Random Forest	0.7358	190.4	3.750
XGBoost	0.7404	26.34	0.1297
Fraction of F in Dataset	0.5411	N/A	N/A

For a third time, the Logistic Regression model had the highest classification accuracy at 69.66% across the

Judging/Perceiving (J/P) dimension. The classification across this dimension split the difference between the behavior of the I/E and N/S dimensions and the T/F dimension. The distribution of the J/P dimension within the dataset was between the imbalanced I/E and N/S dimensions and the well-balanced T/F dimension. However, regardless of dimension distribution, this dimension had the lowest classification accuracy of all four.

TABLE 6: JUDGING/PERCEIVING CLASSIFICATION REPORT

Algorithm	Accuracy	Train Time (s)	Test Time (s)
Logistic Regression	**0.6966**	**1.185**	0.0027
Linear SVM	0.6935	2.709	**0.0024**
Random Forest	0.6175	15.93	0.4118
XGBoost	0.6536	44.10	0.1456
Fraction of P in Dataset	0.6041	N/A	N/A

The E/I ROC curves (Fig. 6) were included to describe the sensitivity and specificity of the classifiers. The ROC curves for N/S, T/F, and J/P all exhibit the same behavior, so the E/I plot has been included to summarize all of them. The ROC curve highlights the difference in performance depending on the class majority. For example, the Introversion dimension (majority class) curves for three of the classification algorithms all have better specificity and sensitivity in comparison to their Extroversion dimension (minority class) counterparts. The gap between the class-wise curves is greater for the Logistic Regression than Random Forest and XGBoost. The more consistent class-wise classification accuracy for the ensemble learners is attributed to their superior generalization.

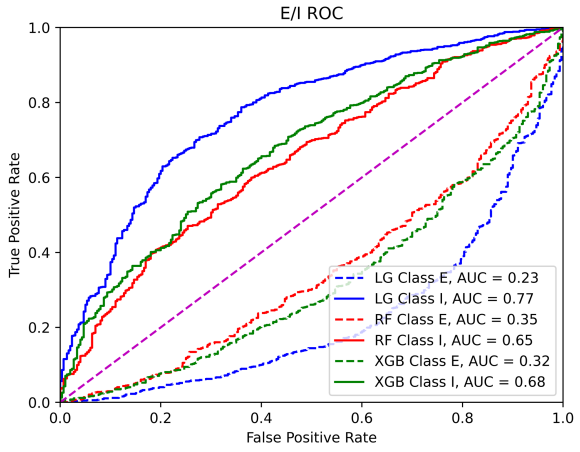

Fig. 6. Reciever Operating Characteristics Curve for N=2 Classification of Introversion versus Extroversion dimension. Inclusive of results from Logistic Regression, Random Forest, and XGBoost models.

The most commonly occurring misclassified words by logistic regression for N=16 classification were: “like”, “just”, “don’t”, “think”, “people”, “know” with 747, 743, 739, 732, 727, and 723 misclassifications, respectively (Appendix A). Other relevant words that were misclassified commonly were those related to the dimensions of the MBTI such as related to oneself (I: I’m, I’ve, I’ll, I’d), external relationships (E: world, friends, friend, person, talk, talking, tell), thinking (T: thought, believe, don’t think, thinking), and feeling (F: feel, feeling). There were a few related to the other dimensions such as “sounds” (S) and “know” (N). Other common words were those with positive sentiments (good, best, better, great) and negative sentiments (mean, hard, bad). The number of misclassifications for each word varied slightly between models, but there was minimal variation in the most common words.

V. CONCLUSION

Four different Machine Learning algorithms were used to classify the MB personality types of the forum users across the MB types (16 classes) and across the MB dimensions (4 classes). Linear SVM and Logistic Regression produced higher classification accuracies than Random Forest and XGBoost. For 2-class classification, the class accuracies of the algorithms were affected by the unequal dimension distribution within the dataset. The skewed class distribution that impacted the 2-class classification had a similar impact on the 16-class classification. This was proven with confusion matrices which showed that classes were predicted in relation to their occurrences in the dataset.

One of the main ways to improve the classification results is to use techniques which account for the skewed class distribution. One technique to distribute the classes more evenly before training is oversampling. This would involve copying or creating new samples of the datapoints from less frequent classes. Oversampling would help to decrease the class imbalance of the dataset which may lead to better results. Additionally, methods to improve the algorithms through weighting the classes with fewer samples could be implemented.

Differences in distribution of personality types amongst the general population and the dataset could be due to a poor sampling method or certain personality types being likely to use chat forums. In this regard, it may be difficult to obtain a representative sample of the general population using social media. Other text sources that may have less bias towards certain personality types that could be used in the future include text messages, emails, or essays.

Many different data preprocessing techniques and algorithm adjustments could be attempted in future work. Different feature extraction techniques could be used to improve the significance of personality relevant vocabulary. These methods include using LIWC, which identifies psychologically significant words, EmoSenticNet with assigns affect emotion labels, or NRC Emotion Lexicon which is a list of words associated with the eight basic emotions and positive and negative sentiments. Text normalization could also be used to transform words into standard form to account for misspellings and abbreviations that are very commonly used on social media.

References

[1] Anon. MBTI® Basics. Retrieved November 5, 2021 from https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/

[2] Anon. MBTI Facts. Retrieved November 5, 2021 from https://www.themyersbriggs.com/en-US/Support/MBTI-Facts

[3] Mitchell J. 2017. (MBTI) Myers-Briggs Personality Type Dataset. Retrieved November 5, 2021 from https://www.kaggle.com/datasnaek/mbti-type

[4] Gjurković, M. and Šnajder, J., 2018, June. Reddit: A gold mine for personality prediction. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (pp. 87-97).

[5] Bharadwaj, S., Sridhar, S., Choudhary, R. and Srinath, R., 2018, September. Persona Traits identification based on myers-briggs type indicator (MBTI)-a text classification approach. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1076-1082). IEEE.

[6] Cui, B. and Qi, C., 2017. Persona Traits Identification based on Myers-Briggs Type Indicator (MBTI)-a text classification approach.

[7] Khan, A.S., Ahmad, H., Asghar, M.Z., Saddozai, F.K., Arif, A. and Khalid, H.A., 2020. Personality Classification from Online Text using Machine Learning Approach. *International Journal of Advanced Computer Science and Applications*, *11*(3).

[8] Anon. Myers-Briggs History. Retrieved December 6, 2021 from https://eu.themyersbriggs.com/en/tools/MBTI/Myers-Briggs-history

VI. Appendix

A. Top 100 Misclassified Words

Word	Misclassifications	Word	Misclassifications
like	747	youre	501
just	743	thought	489
dont	739	said	485
think	732	pretty	476
people	727	friends	472
know	723	yes	471
Im	718	got	463
really	686	did	462
time	682	try	461
good	651	read	460
way	624	didnt	459
say	622	best	455
Ive	613	better	454
want	612	doesnt	452
things	612	probably	448
feel	597	work	444
make	591	maybe	442
love	572	look	441
life	566	Ill	434
thats	562	does	434
going	562	mean	429
right	550	Id	428
lot	549	long	424
type	544	little	421
thing	544	thread	411
actually	518	kind	409
need	505	makes	408
sure	504	years	405
person	503	different	404
thinking	402	getting	361
friend	401	yeah	359
tell	400	bit	357
trying	393	agree	356
come	390	use	355
post	389	guess	353
hard	387	types	351
having	385	feeling	351
dont know	385	times	350
understand	385	thanks	346
day	384	help	346
world	382	believe	343
mind	376	quite	342
usually	375	talk	342
bad	375	isnt	340
oh	373	used	334
doing	371	talking	330
point	371	dont think	327
interesting	366	true	325
new	365	sounds	325
great	363	theres	322