Myers-Briggs Personality Prediction

Kishan Sathish Babu, Pranav Kanth Anbarasan, Sreeja Vepa

sathishbabu.ki@northeastern.edu, anbarasan.p@northeastern.edu, vepa.s@northeastern.edu

Khoury College of Computer and Information Science

Northeastern University

1. Objectives and Significance

The primary objective of this project is to explore potential connections between an individual's Myers-Brigg Type Indicator (MBTI) and their personal writing style in social media data. The Myers-Briggs Type Indicator is a personality assessment tool used to categorize individuals into one of 16 personality types. The goal of this project is to analyze the textual data using learning models such as logistic regression, support vector machine (SVM), and random forests to predict an individual's MBTI type.

The possibility of a presence of patterns between MBTI types and personal writing styles gives insight into the general validity of the personality test's ability in analyzing, predicting, and categorizing behavior. Having a better understanding of one's personality can be useful in a variety of ways such as improving personal counseling and education effectiveness, focused ad-marketing, and focused digital content. This large range of practical applications is one of the key motivators of this project. Additionally, although the Myers-Briggs Type Indicator is not a recent addition to the landscape of personality assessments, it has shown its ability to remain consistently relevant despite the countless newer tests created after. Thus, this project is also motivated by a desire to understand the rationale behind this steady relevance.

The data has been sourced from two distinct social media related datasets: the Personality Café Forum and Twitter. Following data collection, a range of features have been extracted from the comments utilizing diverse methods such as sentiment analysis and the application of natural language processing tools like nltk for grammatical tagging. Due to an imbalance in class distribution within the dataset, a Synthetic

Minority Over-sampling Technique (SMOTE) has been employed to generate artificial samples for the minority classes. Upon completion of feature extraction, various modeling techniques, including logistic regression, decision trees/xgboost, naive bayes, and support vector machines, have been employed to facilitate predictions. Notably, the classification problem at hand is treated as a multilabel one, where each letter in the final label is considered an independent label. The outcomes of the models indicate comparable accuracies across all three approaches. Specifically, predictions for labels pertaining to Extroverted/Introverted and Sensors/Intuitive exhibit an accuracy of approximately 73%. In contrast, predictions for labels associated with Feelers/Thinkers and Judgers/Perceivers demonstrate a noticeable decrease to around 60%, signifying a substantial deviation compared to the accuracy achieved for the other two labels.

2. Background

2.1 Myers-Brigg Type Indicator

The Myers-Brigg Type Indicator (MBTI) is a widely known psychological assessment designed to evaluate personality preferences and classify individuals' personalities into one of 16 distinct personality categories. There are four dichotomies in the MBTI test: Extroversion (E) versus Introversion (I), Sensing (S) versus Intuition (N), Thinking (T) versus Feeling (F), and Judging (J) versus Perceiving (P). Each individual is assigned one value in each of these categories.

E/I: Extroverts may derive energy from social interactions, while introverts may derive more energy from activities done alone.

S/N: Sensors may focus on concrete and factual information leading them to be considered more practical. Those in the intuition category may focus on patterns and impressions leaning towards innovative decisions.

T/F: This dichotomy measures how an individual may make a majority of their decisions. A thinker may rely on logic and objective reasoning in comparison to feelers who tend to

think of personal values the most and the impact their decisions have on their surroundings.

J/P: Judgers often prefer structure and order often liking to plan-ahead. On the other hand, perceivers are more inclined to prefer flexibility and spontaneity.

2.2 Logistic Regression

Logistic regression models the probability of a discrete outcome given a certain input variable. The statistical model shows the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. The decision boundary of a logistic model is determined by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other.

This model has been used on each of the four dichotomies in the MBTI personality type. The probability from the logistic function is based on the sigmoid function given by:
$p(x) = 11+e-wTx$

2.3 Support Vector Machines

Support Vector Machines (SVM) is a supervised learning algorithm used for classification and regression tasks. SVM finds optimal hyperplanes by maximizing the margins between the classes in the data. The method is also built to handle high-dimensional data well.

This project relies on SVM to create the optimal subspaces between each of the four dichotomies in the MBTI types using hyperplanes by separating the two classes by the largest possible margins. Using SVM will also help assess the effectiveness of this project's chosen engineered features.

2.4 Random Forests

Decision trees are standalone algorithms that recursively split data based on feature values. While the model does offer simplicity, it is prone to overfitting. Decision trees may lack the predictive power of more complex models. To overcome these boundaries, XGBoost was also employed.

2.5 Extreme Gradient Boosting (XGBoost)

XGBoost is an ensemble learning algorithm that employs boosting to combine the predictions of multiple weak models, usually decision trees using regularization techniques, such as L1 and L2 regularization, to control the complexity of individual trees and the overall ensemble. The algorithm employs gradient boosting for optimization and has built-in mechanisms for handling missing values and early stopping.

2.6 Naïve Bayes Model

This probabilistic machine learning algorithm stems from Bayes' Theorem. The Naïve Bayes model is valuable in personality prediction project due to its efficient handling of textual data such as the social media posts analyzed in this study. The model's adeptness at handling multiple features and high-dimensional data also contributes to its usefulness in this project.

2.5 Previous Work

"Personality Traits on Twitter":

This study utilizes Twitter users' self-reported MBTI personality types to create a sizable dataset of tweets. It builds models for each Myers-Briggs test letter, using logistic regression. Successful modeling is achieved for Introvert/Extrovert and Feeling/Thinking traits, but challenges arise in predicting other characteristics. Despite differences in Twitter user data compared to the real world, the language usage in the data accurately represents real language patterns with significant sample sizes. The study suggests that analyzing user attributes on a large scale with an open vocabulary can improve model accuracy and offer valuable insights into personality types.

"Reddit: A Gold Mine for Personality Prediction":

This paper introduces the MBTI9K dataset, designed specifically for MBTI personality type detection using data from Reddit. The dataset looks to address limitations found in previous datasets related to user non-anonymity and limited topic diversity. Truth labels are established using Reddit Flairs, commonly used in MBTI discussion subreddits to

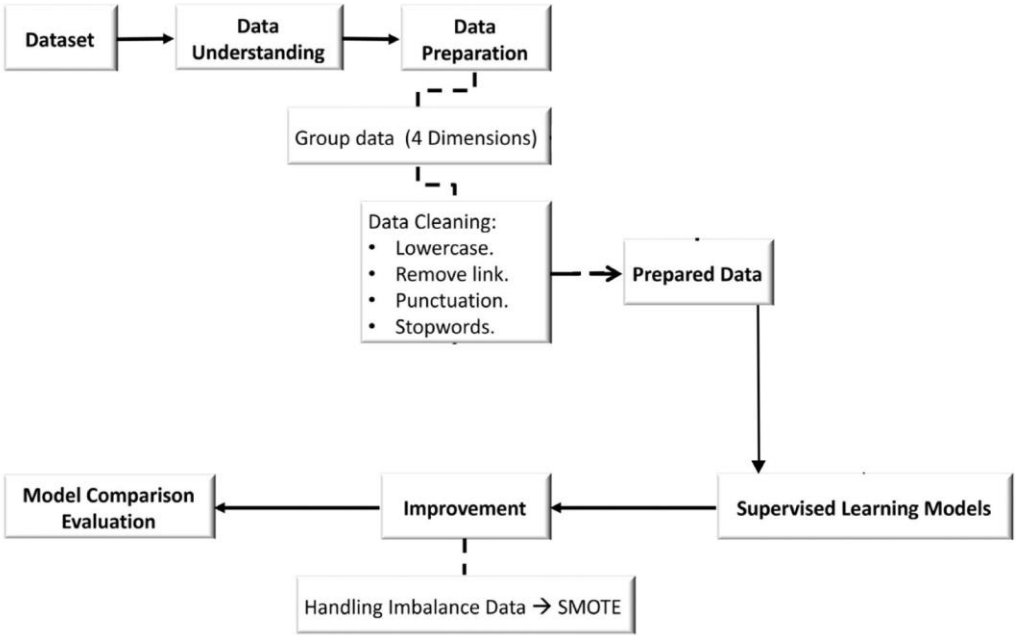
indicate personality types. The dataset includes comments and posts annotated with corresponding MBTI types, enabling the training of personality prediction models that outperform baseline performance.

2.6 Distinction of this Project

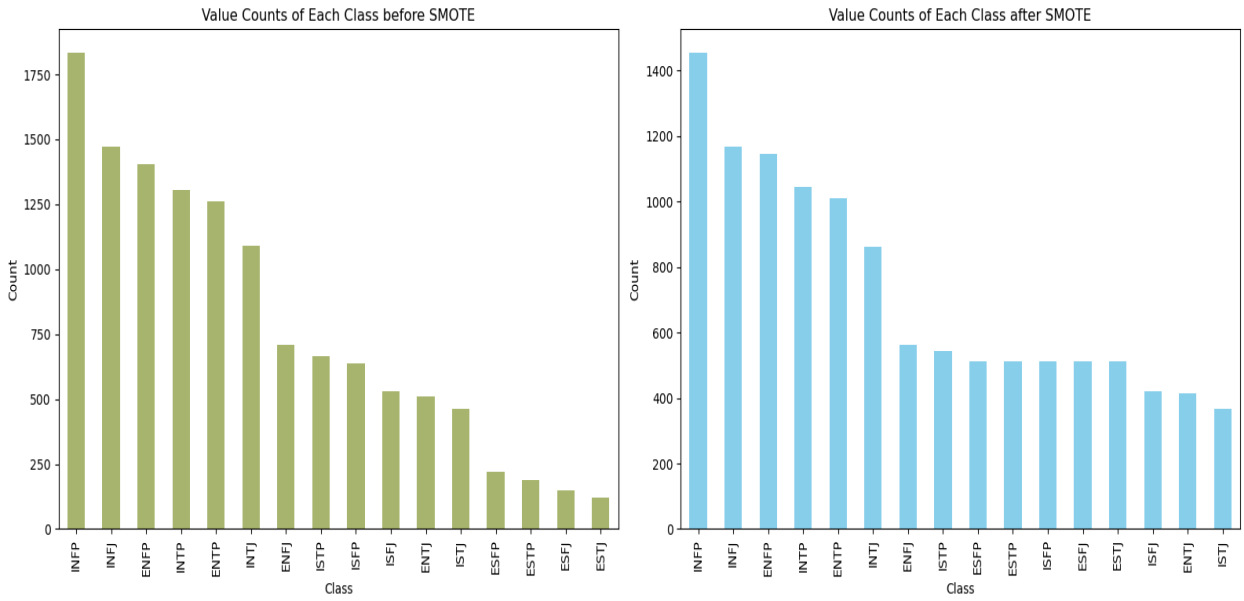
While previous works have explored personality prediction using diverse personality tests, this project distinguishes itself through the integration of two distinct datasets. Unlike the literature surveyed, none have combined these particular datasets. The text data utilized for analysis are from PersonalityCafe.com and Twitter, both in social media post form. The primary source of the dataset was PersonalityCafe.com with specific samples of data from the Twitter dataset added to address imbalance issues. Also, VADER, a nltk module, was utilized to perform sentiment analysis due to its efficiency in social media data.

Before initiating the project, a preliminary examination of the data revealed a significant skew despite the added data. Some MBTI types had considerably more data than others. To address this imbalance, the project also employed SMOTE analysis to equalize the weighting of the data, mitigating disparities in class distribution.

3. Methods



This project's dataset was a combination of two datasets from Personality Café Forum and Twitter, respectively. Both of these datasets are now accessible on Kaggle as well. The Personality Café data set had around 8600 rows and the Twitter data set had over 7800 rows of data. The Personality Café data was used fully, but the imbalance was large. Thus, data relating to specific minority MBTI types were drawn from the Twitter dataset to achiever higher prediction accuracies. The final combined dataset consisted of approximately 12,555 rows, each row with posts from a unique user. Both datasets had "|||" pipelines within each row separating different comments made from users. Upon removal of the pipelines symbol, the entire dataset revealed a total of 975,145 posts. The initial column of the final dataset indicates the user's MBTI type and the second column contains the individuals' posts.



The bar graph illustrates the distribution of the dataset among different MBTI types, revealing imbalances in several types. To address this issue, reorganization of the classes into four dichotomous categories instead of the original sixteen personality categories was decided. This involved a meticulous data cleaning process and the application of synthetic minority oversampling techniques (SMOTE) to mitigate the imbalances in class representation. The post-SMOTE class counts are presented in the

bar graph below. The four minority classes up sampled were the minority classes: ESTP, ESFJ, ESFP, and ESTJ.

Up sampling minority classes using SMOTE, can be beneficial for a variety of reasons such as improved model generalization and increased sensitivity to minority classes. If generating a more balanced distribution of classes, the model may become better able to generalize across all classes and any possible bias towards the majority classes may be minimized. Also, generation of synthetic samples in the minority classes allows the model to have more diverse examples to learn from.

The initial set of features encompasses the extraction of syntactical values, including metrics such as average word count per comment, variance of word count per comment, punctuation usage, and references to external websites.

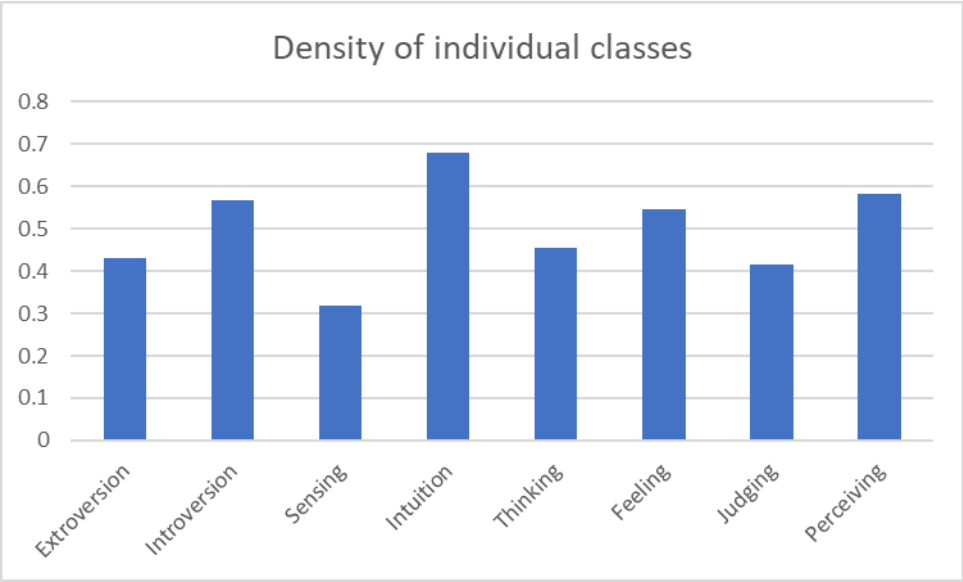
The subsequent set of features comprises semantic attributes, specifically focusing on parts of speech such as nouns, verbs, adverbs, and other linguistic elements. To achieve this, we utilized the 'averaged_perceptron_tagger' module from the NLTK library for word tagging and analysis.

An important portion of this project was sentiment analysis. To properly represent the social media textual data, we employed the NLTK library's VADER module. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a rule-based sentiment analysis tool designed to handle English text. This tool was originally created due to its specific efficiency in handling textual data with informal language such as social media data. The decision to use VADER was driven by the nature of the combined dataset, which originated from two social media-oriented sources. By utilizing VADER, it was possible to better acknowledge the nuances and expressions common in online communication. Another important aspect of the sentiment analysis was data preprocessing using the NLTK library's stopwords list. Stop words are common words such as "and", "the", "is", etc., that are often removed so the focus can remain on content-carrying words such as "bad" or "good". Categorization of terms based on their semantic orientation through VADER's polarity score functionality allowed for a text sentiment to be represented numerically on a scale from –1 to 1. This scale was then pieced into five categorical

values: “Strong Positive”, “Positive”, “Negative”, “Strong Negative”, and “Neutral” for better analysis.

The task of MBTI personality classification involves a classification paradigm. Consequently, we have identified five fundamental machine learning classifiers for this purpose, namely: Naïve Bayes, Support Vector Machine, Logistic Regression, Random Forests, and XGBoost. The latter two being relatively similar in nature.

Given the individualistic nature of MBTI classification parameters, we partitioned the output type into its four constituent classes.



In this project, a diverse set of methodologies were employed to assess the performance of the trained models. The primary metric for evaluation was accuracy, given the substantial balance achieved in the dataset after the up-sampling process. Additionally, the F1 score was utilized due to its suitability for evaluating models on imbalanced data. Furthermore, ROC AUC scores and k-fold cross validation were implemented to better provide a comprehensive evaluation of the model effectiveness.

4. Results

The sentiment analysis module used, VADER, was chosen due to its adeptness in analyzing social media data. Among various implementations of text sentiment analysis

like TextBlob and SpaCy, the Vader module from the NLTK library is the most suitable for the current data. This preference arises from its capability to understand emoticons, which are commonly used in comments, and its ability to discern neutral scores. The following snippets of code depict the polarity value contributions that VADER allows compared to other sentiment analysis measures.

Code 1:

```
individual_scores = sia.polarity_scores("dont hurt them :(")
individual_scores
```

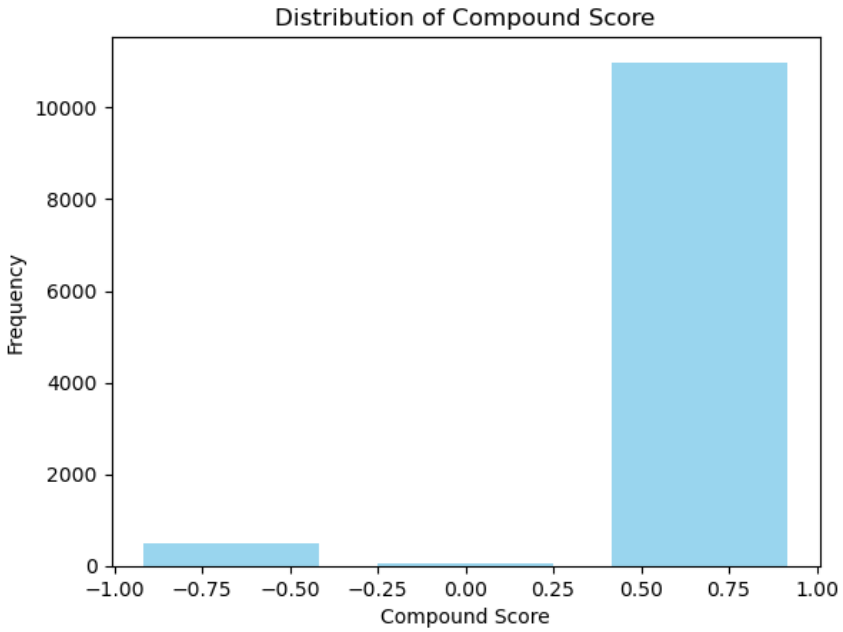
```
{'neg': 0.0, 'neu': 0.278, 'pos': 0.722, 'compound': 0.6348}
```

Code 2:

```
individual_scores = sia.polarity_scores("dont hurt them :)")
individual_scores
```

```
{'neg': 0.342, 'neu': 0.276, 'pos': 0.383, 'compound': 0.0762}
```

As seen, the different emojis greatly impact the final compound score with a difference between 0.0762 and 0.6348.

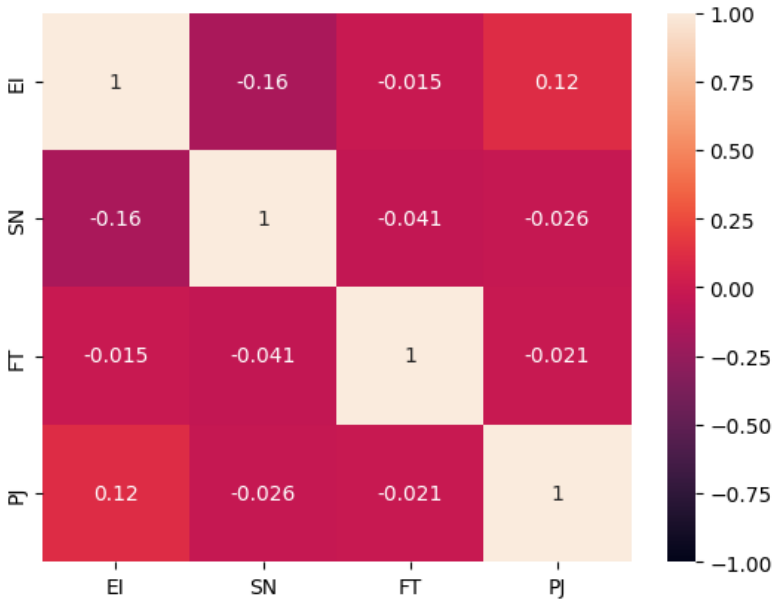


The above histogram shows that most text polarity scores had positive polarity values between approximately 0.4 and 0.9. There were very few values near the neutral value of 0.

When modeling the algorithm with a single output encompassing 16 possibilities, a substantial decline in accuracy is observed, and in certain instances, the model fails to

converge. Given the nature of the MBTI dataset, where each class is independent of the others, it is more prudent to treat this as four separate classifiers. This phenomenon is also evident in the correlation matrix.

Correlation of output parameters



Base Model

Base Model Metrics					
	E/I	S/N	F/T	P/J	Overall
Trivial	0.56	0.68	0.54	0.58	0.12
Random	0.5	0.5	0.5	0.5	0.0625

The probability of the trivial classifier is derived from the priors obtained from the dataset.

Support Vector Machines:

Support Vector Machines Metrics				
	Separate F1 Scores	Average F1 Score	AUC Score	Cross-Fold Validation Mean Accuracy
Extrovert/ Introvert	0.80, 0.59	0.73	0.72	0.73
Sensing/ Intuitive	0.85, 0.50	0.76	0.71	0.75
Feeling/ Thinking	0.53, 0.70	0.63	0.54	0.64
Perceiving/ Judging	0.11, 0.73	0.59	0.50	0.60

The above table indicates the performance of the SVM classifier on the test dataset. Note the listed F1 Scores, E/I and S/N seem to have accuracies around and slightly above 0.70 while F/T and P/J are around 0.60. However, the precision and recall distributions among the eight personality categories are vastly different.

The E/I and S/N categories both have high precision and low recall scores. As precision accounts for positive predictive value, it represents the ratio of correctly predicted positive observations to the total predicted positives. High precision shows the model has a low false-positive rate. Thus, when it predicts a positive outcome, it is likely correct in its classification. Recall, also known as sensitivity, is the true positive rate. A high recall indicates that the model can identify a large proportion of the positive instances. Despite E/I and S/N being correct in their positive classifications about 0.70 of the time, it seems that the model may be missing many positive instances.

The F/T and P/J categories have the opposite issue with the recall score being high and the precision score low, meaning the model is generating a large number of false positives. However, the model is capturing a sizable portion of the actual positive instances as the recall values are around 0.70.

Achieving a balanced precision-recall ratio is ideal, and in the above data it can be seen that although the F1 score seems relatively stable between the four rows, the precision-recall ratio is unbalanced.

Additionally, the k-fold cross validation scores are very similar to the F1 scores. However, the AUC scores deviate in the F/T and P/J categories. A lower AUC score may be attributed to the model's weaker ability in discriminating between positive and negative instances. This could be due to class imbalance or insufficient feature representation in the data set.

Additionally, running a classification system directly on the "types" column, which had all 16 personality types, was attempted. However, the model did not work. This is most likely due to the data not being linearly separable.

Random Forest and XGBoost:

In both the Random Forest and XGBoost models, significant improvements in predictions are evident compared to the base models. Notably, the E/I and S/N classes exhibit substantial enhancements, while the F/T class demonstrates a moderate increase in accuracy. However, the P/J class shows minimal improvement, a trend observed in other studies addressing the same problem. The critical challenge arises from the binary nature of the MBTI test, where an incorrect prediction may lead to an output in the opposite extreme, profiling the user in the opposite class.

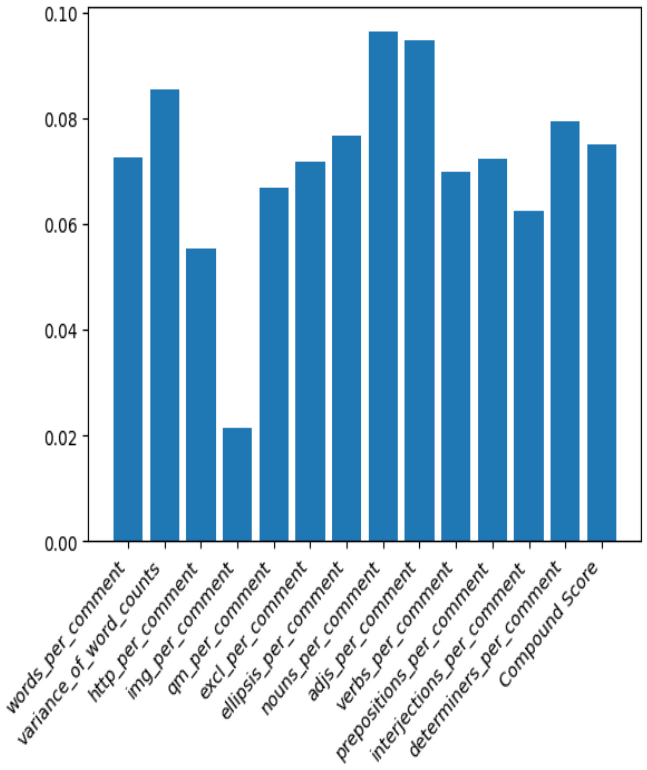
We are discussing Random Forest and XGBoost in tandem due to the similarities in their algorithms.

Random Forest and XG Boost Metrics						
		E/I	S/N	F/T	P/J	Overall
Random Forest	F1 score	0.665	0.583	0.584	0.513	0.586
	Accuracy	0.712	0.764	0.587	0.532	0.170
	AUC ROC score	0.655	0.575	0.587	0.509	0.581

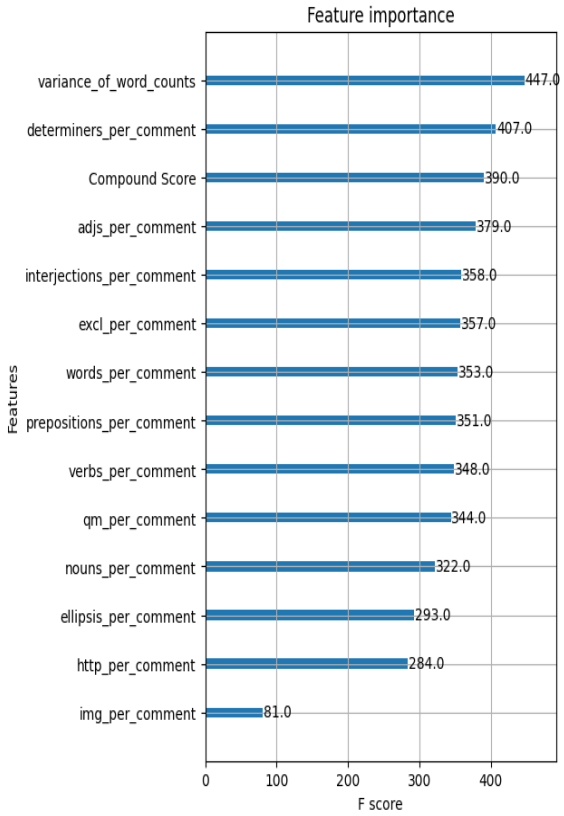
XGBoost	F1 Score	0.561	0.291	0.650	0.693	0.548
	Accuracy	0.726	0.772	0.622	0.567	0.197
	ROC score	0.674	0.574	0.622	0.513	0.598

A key observation emphasizes the critical role of sentiment analysis scores (Compound Score) as vital features in both models, underscoring the importance of sentiment analysis in their predictive performance.

Feature importance in Random Forest:



Feature Importance in XGBoost:



Logistic Regression:

In the initial approach, logistic regression was attempted as a binary classifier for the multi-class problem, where the target variable represented four distinct personality traits

(E/I, S/N, F/T, P/J). However, logistic regression, being inherently binary, faced convergence issues when applied directly to the multi-class scenario. To overcome this challenge, the target variable was transformed into four independent binary classes, each corresponding to one of the personality traits.

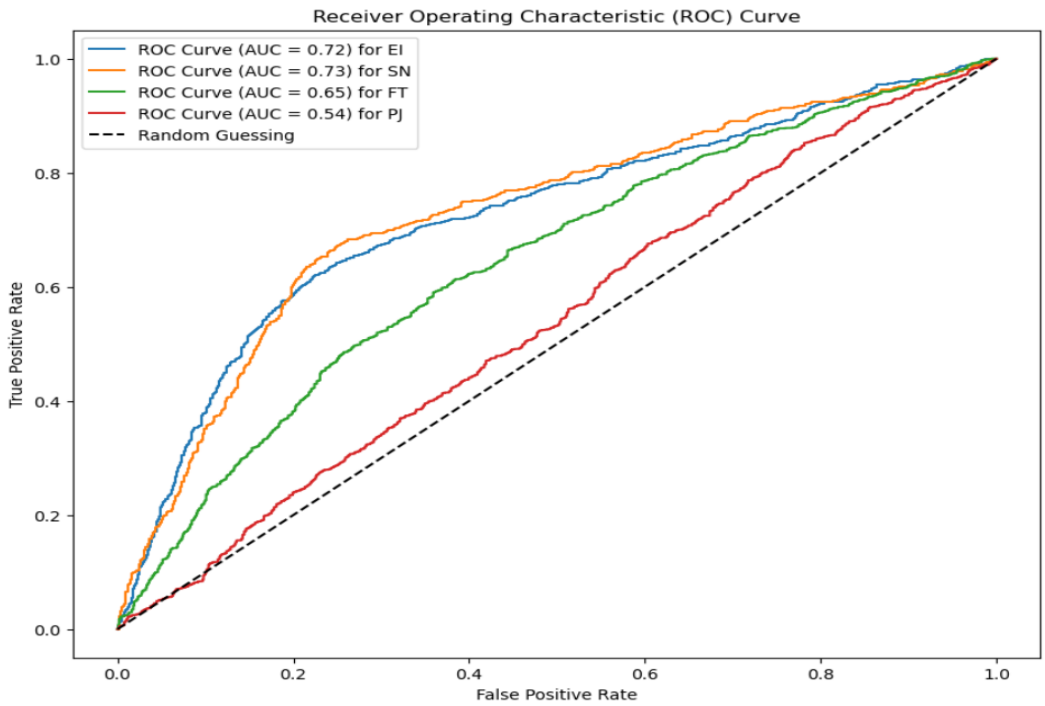
By treating each personality trait as a binary classification problem, logistic regression was able to converge successfully. This approach allowed the model to independently predict the presence or absence of each trait, facilitating better convergence and more reliable predictions. While this transformation introduces some simplification by addressing each trait in isolation, it provides a practical workaround for logistic regression, enabling effective modeling in the context of personality trait prediction.

Logistic Regression Metrics			
	Accuracy	F1 Score	Cross Validation Accuracy
Extraverted/ Introverted	0.72	0.59	0.72
Sensors/ Intuitive	0.75	0.53	0.72
Thinkers/ Feelers	0.61	0.67	0.58
Judgers/ Perceivers	0.58	0.73	0.58

The binary classification models for personality traits exhibit diverse performance metrics. Extrovert/Introvert and Sensors/Intuitive achieved accuracies of 72.84% and 75.35%, respectively, with corresponding F1 scores of 59.60% and 53.56%. Thinkers/Feelers demonstrated 61.37% accuracy and a notable F1 score of 67.32%, while Judgers/Perceivers achieved a lower accuracy of 58.74% but with a noteworthy

F1 score of 73.46%. Cross-validation results further validate the models, showcasing consistent performance across different personality dimensions.

Area Under the Curve (AUC):



The AUC (Area Under the Curve) is a metric used to assess the performance of a classification model. Specifically, it represents the area under the Receiver Operating Characteristic (ROC) curve, which visualizes the trade-off between true positive rate (sensitivity) and false positive rate. The AUC scores and accuracies for personality trait classification show some correlation. E/I (0.72 AUC, 0.73 accuracy) and S/N (0.73 AUC, achieving 0.75 accuracy) exhibit alignment between higher AUC values and predictive performance. However, for F/T (0.65 AUC, 0.61 accuracy) and P/J (0.54 AUC, 0.59 accuracy), differences suggest potential complexities in precisely distinguishing between classes.

5. Conclusions

The main objective of this research was to consider the efficacy of MBTI personality type prediction using text and sentiment analysis. This was done using various libraries and modules in Python such as pandas, sklearn and nltk. To reach the objective, sentiment analysis was combined with data balancing measures such as data set combination and SMOTE, general pre-processing measures, and detailed implementations of five machine learning classifiers

Notably, the strategic implementation of SMOTE yielded promising outcomes across all models. By generating synthetic instances of the minority class, SMOTE effectively addressed the skewed distribution, allowing the models to better capture the nuances of the minority classes. The observed stabilizing of the F1 score in the XGBoost algorithm as we increased the depth highlighted a crucial relationship between parameters, especially in the context of an imbalanced dataset. The flattening of the graph shows the relation between the depth of the model and the imbalanced dataset. This highlights the requirement of a large balanced dataset to increase the F1 score and build a better model.

The F1 score proved to be a pivotal metric in the context of data classification, it demonstrated reliable performance, but it could have improved. There were some interesting nuances in the data that indicated the need for improvements in the SVM models' F1 accuracies. From the results section's SVM Metrics table, the F1 scores' basis, which leverages on the precision and recall scores, was unbalanced. For all four dichotomies, either the precision was high with a low recall or the opposite. These imbalances have a high chance of being due to imbalanced data. Additionally, in SVM, the AUC scores were low for two of the dichotomies: F/T and P/J. This could be due to class imbalance or insufficient feature representation in the data set. By creating a larger dataset with more information related to those in categories such as Extrovert, we would be able to better represent all MBTI types and achieve better scores in AUC, F1, and more.

Considering the target variables as independent made it feasible to use logistic regression. The binary models for personality traits exhibit varying performances. E/I and S/N show strong alignment between AUC and accuracy while F/T and J/P indicate

complexities in classification. This variability in performance may be attributed to the potential inadequacy of the model in capturing sufficient information.

Looking forward, if we want to make improvements to the project, we could add more data sources to better correct data imbalances, especially in the minority classes. Advanced text analysis tools like tokenization and word2vec could also be implemented. It is also possible to try machine learning algorithms such as LSTMs to improve the predictions about MBTI personalities. Additionally, to improve the sentiment analysis we could implement new word-embedding techniques, such as BERT, which helps capture subtle meanings between words.

6. Individual Task

Pranav was responsible for implementing the SMOTE (Synthetic Minority Over-sampling) to address class imbalance and was also responsible for implementing logistic regression. Sreeja was responsible for implementing Support Vector Machine (SVM) and sentiment analysis. Kishan was responsible for feature extraction and implementing Xgboost/Random forests and Naïve Bayes.

7. References

Colab file:
https://colab.research.google.com/drive/1LbgUmB09XSW4E5rvL4lEKkBP7Ws3UKQy?usp=sharing

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Matej Gjurković and Jan Šnajder. 2018. 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, pages 87–97, New Orleans, Louisiana, USA. Association for Computational Linguistics.

(MBTI) Myers-Briggs Personality Type Dataset. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/datasnaek/mbti-type

MBTI Personality Type Twitter Dataset. (n.d.). Www.kaggle.com. Retrieved November 1, 2023, from https://www.kaggle.com/datasets/mazlumi/mbti-personality-type-twitter-dataset/code

Plank, B., & Hovy, D. (2015). Personality Traits on Twitter—or—How to Get 1,500 Personality Tests in a Week. WASSA@EMNLP.

Ryan, G.; Katarina, P.; Suhartono, D. MBTI Personality Prediction Using Machine Learning and SMOTE for Balancing Data Based on Statement Sentences. Information 2023, 14, 217. https://doi.org/10.3390/info14040217