International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Myers Based Personality Prediction System

Diksha Singh[*1], *Divya Ramani*[*2], *Prof Neelam Sharma*[*3]

[*1,2] Student, Computer Science and Engineering, Shri Shankaracharya Technical Campus, India
[*3] Professor, Computer Science and Engineering, Shri Shankaracharya Technical Campus, India

ABSTRACT*:*

Nowadays, personality stands out as a highly researched and captivating subject in psychology. The recognition of user personalities finds extensive utility in various research fields such as recommendation systems and human-robot interaction. Conventional recommendation systems encounter issues like insufficient user preference data, free-rider problems, and data sparsity. Understanding the personality traits of identified users aids in comprehending their preferences. The MBTI test demonstrates a test-retest reliability with an error rate of approximately 0.5. Upon retesting, individuals tend to exhibit 3–4 type preferences around 75–90% of the time. Our methodology offers greater accuracy compared to existing tests, empowering users to trust their results. Personality classification through digital data emerges as a more effective substitute for traditional psychological assessments.

Keywords: Personality recognition, Personality traits, Traditional psychology, Data sparsity.

INTRODUCTION

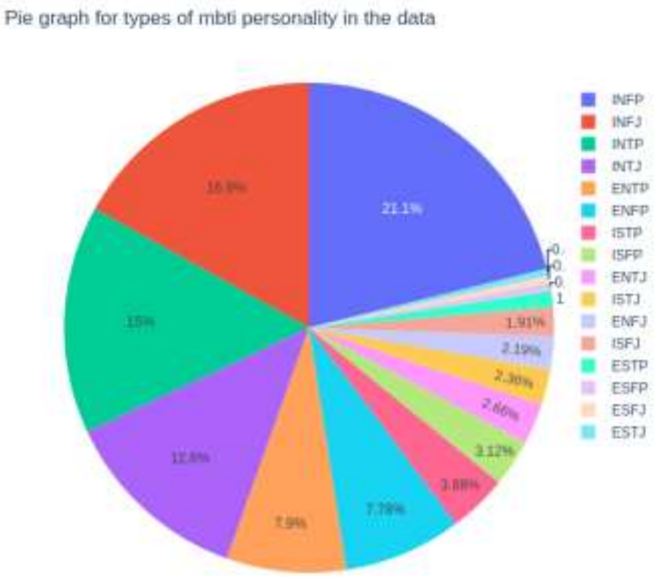
The Myers Briggs Type Indicator (MBTI) holds its status as a widely utilized personality assessment tool globally. It categorizes individuals into 16 distinct personality types by assessing four dimensions: Introversion (I) versus Extroversion (E), Intuition (N) versus Sensing (S), Thinking (T) versus Feeling (F), and Judging (J) versus Perceiving (P). The MBTI's predictions of personality traits maintain fundamental similarities with traditional personality constructs. Researchers employ machine learning and deep learning algorithms extensively to forecast personality and psychological attributes from digital data.

We are in the process of developing an MBTI personality classifier utilizing machine learning models. This classifier aims to predict an individual's personality based on their most recent 50 social media posts. We have observed correlations between a person's MBTI personality type and their writing style. Additionally, the classifier serves to validate the accuracy of the MBTI test. Our approach involves the utilization of a substantial volume of annotated personality data extracted from social media platforms. Moreover, our model operates on a larger dataset compared to conventional personality tests, enhancing the reliability of the results for users.

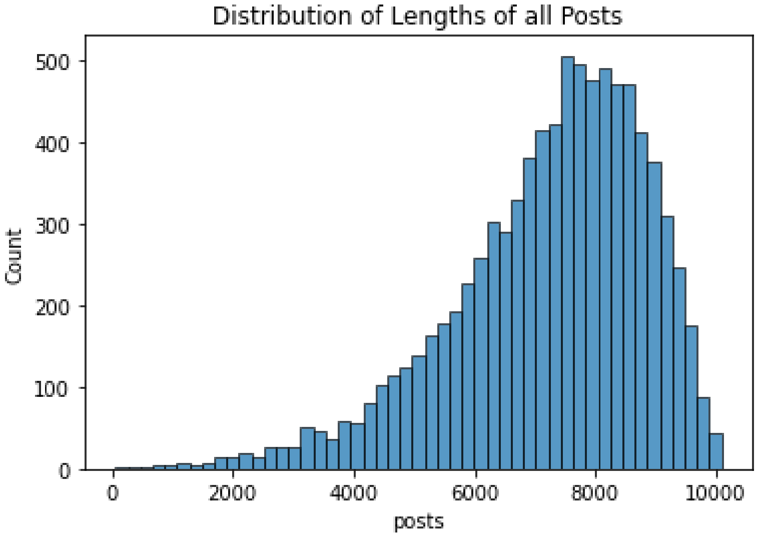
MATERIAL AND METHODS

Dataset Description and Analysis

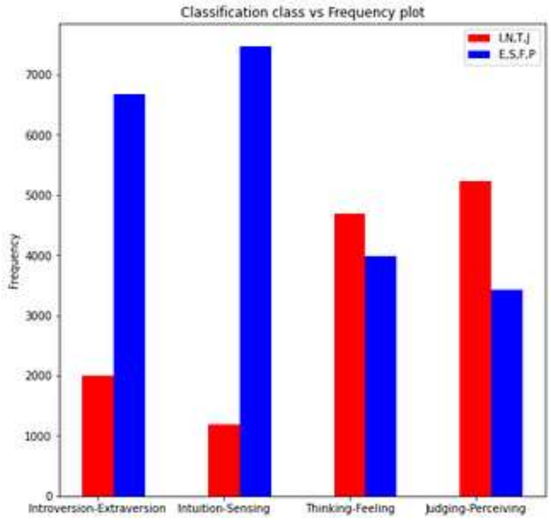
The dataset comprises 8675 entries and is organized into 2 columns: 'type' and 'posts'. Within the 'posts' column, each entry consists of the 50 most recent social media posts per user. The 'type' column encompasses 16 distinct labels, each corresponding to one of the MBTI personality types, and there are no NULL values present in this column.

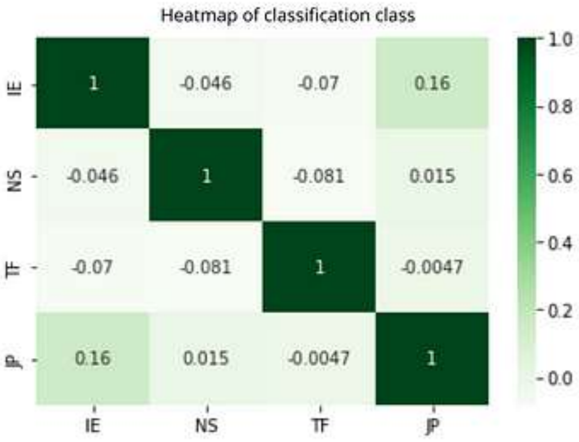


The pie chart depicting the distribution of posts across different personality types indicates an imbalance within the dataset. Additionally, the plot illustrating the length distribution of all posts reveals that certain posts contain fewer than 2000 words, while others range from 4500 to 9000 words.



The MBTI classifier encompasses four primary dimensions: 'Introversion-Extraversion' (IE), 'Intuition-Sensing' (NS), 'Thinking-Feeling' (TF), and 'Judging-Perceiving' (JP). To facilitate analysis, four additional columns have been appended to the dataset. Within each column, '1' denotes the presence of the first part of the respective dimension (I, N, T, J), while '0' signifies the presence of the second part of the dimension (E, S, F, P), respectively.





TF and JP data exhibit a nearly balanced distribution, whereas IE and IS data display imbalance. The heatmap illustrates a positive correlation between IE-JP and NS-JP, while other pairs demonstrate negative correlations (excluding correlations with themselves).

Preprocessing steps are implemented to enhance feature extraction from the textual data within the 'posts' column:

- Conversion to lowercase: Textual data is converted to lowercase to ensure uniformity.

- Removal of URLs/links: URLs are eliminated as they do not directly contribute to personality classification and can be considered irrelevant.

- Elimination of special characters and numbers: Special characters and numbers are removed to reduce noise and outliers in the data.

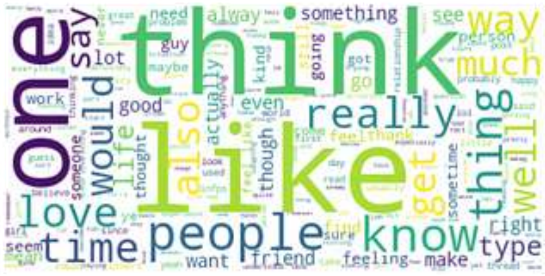
- Removal of extra spaces: Extraneous spaces are eliminated to maintain data cleanliness.

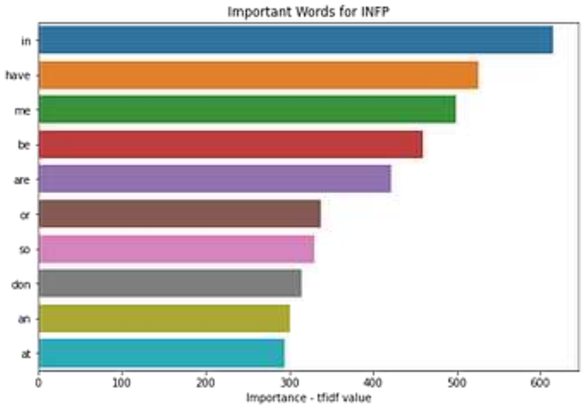
- Extraction of stopwords: Common English words, such as 'for,' 'them,' and 'you,' are removed as they do not provide meaningful information for feature extraction.

- Exclusion of MBTI personality names: MBTI personality names mentioned in posts are omitted to prevent biasing the results.

- Lemmatization: Words with similar meanings are grouped together using lemmatization, thereby reducing redundancy (e.g., 'gone,' 'going,' 'went' are treated as one).

Feature extraction involves addressing the imbalance in the dataset by employing TF-IDF and count vectorization techniques. Initially, countVectorizer is used to convert posts into token count matrices for the model. Subsequently, TF-IDF normalization scales the features from the count vectorizer into floating-point values, assessing the relevance of words to the corpus and their importance in the dataset. Following vectorization, each user post is represented by 1500 features.





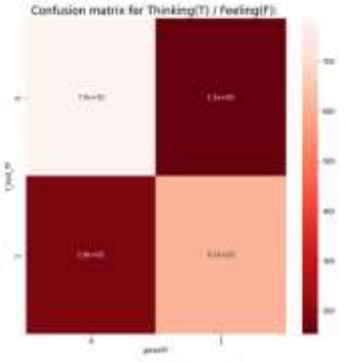
A word cloud illustrates the frequency of words for the INFP type, with each word's size corresponding to its frequency in the top posts. Additionally, a bar graph depicts the significance of various words for the INFP type using TF-IDF values. Similarly, we generated TF-IDF value graphs and word clouds for each personality type.

Methodology

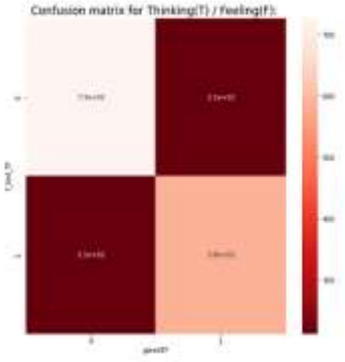
The methodology used in our project involved several steps, starting with preprocessing and feature extraction of the dataset. The resulting dataset was then divided into training and testing sets in an 80:20 ratio. Due to slight imbalance observed in the IE and NS dimensions, Stratified K-Fold cross-validation using GridSearchCV was applied to improve accuracy.

Various machine learning models were employed for MBTI personality classification, including Logistic Regression, Naive Bayes, Random Forest Classifier, K-Nearest Neighbor (KNN), SGD Classifier, and Support Vector Machines (SVM). These models were built using sklearn, NumPy, and pandas libraries.

- Logistic Regression: Achieved over 80% accuracy in IE, NS, and TF classification, and 71% in JP classification.

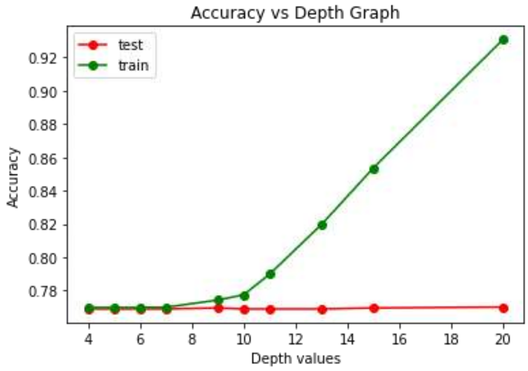


Logistic Regression

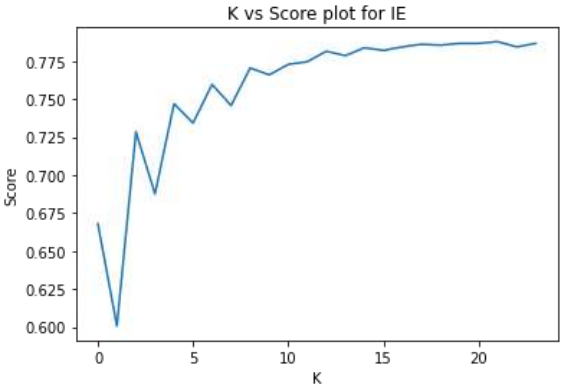


Naive Bayes

- Naive Bayes: Gaussian naive Bayes yielded 75% accuracy in NS and TF classification, and over 64% accuracy in IE and JP classification.



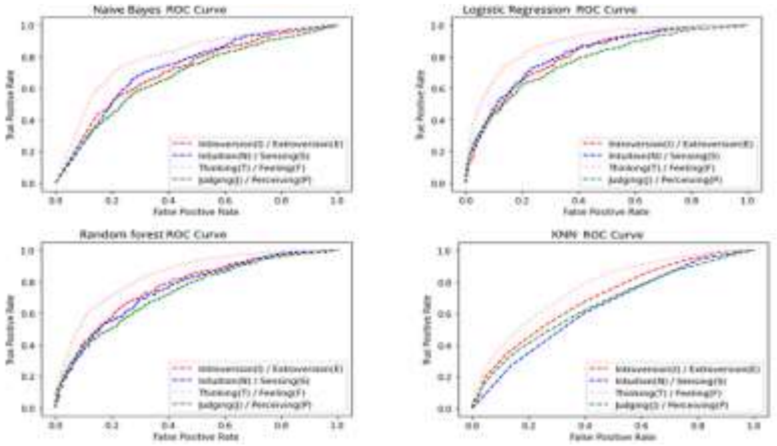
- Random Forest: After hyperparameter tuning, models for IE, NS, and TF classification attained accuracy greater than 70%, while JP classification achieved 60% accuracy.



- KNN Classifier: Achieved above 60% accuracy for FT and JP, and over 78% for IE and NS.

- SGD Classifier: Provided accuracy above 80% for IE, NS, FT, and 70% for JP.

- Support Vector Classification: Predicted accuracy above 80% for IE, NS, FT, and 72% for JP.



RESULT AND DISCUSSION

• Gaussian Naive Bayes showed low accuracy (about 60%) on the testing dataset for all features, indicating poor performance in terms of precision and ROC curve.

• SVC and SGD performed similarly to Logistic Regression, with good precision and recall values.

• Random forest and KNN exhibited good but relatively lower performance, with accuracy around 75%. The ROC curve for KNN indicated poor performance.

• Logistic Regression emerged as the best-performing model for personality classification based on The Myers Briggs Personality Model, supported by the ROC curve analysis.

Conclusion

Our model effectively forecasted MBTI personalities through analysis of social media posts employing six supervised machine learning algorithms. Among these, logistic regression yielded the highest accuracy. Enhanced precision can be achieved through training models on expansive and refined datasets. This system holds potential for refining recommendation systems and aiding governments in outlier identification and comprehension of targeted individuals' personalities. Similarly, companies can leverage MBTI personality test outcomes to gain insight into their employees' behaviors, strengths, weaknesses, and cognitive approaches to data perception, processing, and interpretation.

References

[1] "6839354.pdf." Accessed on January 08, 2022. [Online]. Available at: https://web.stanford.edu/class/archive/cs/c s224n/cs224n.1184/reports/6839354.pdf

[2] Antonio, B. "Data Science Final Project: Myers-Briggs Prediction," Medium, May 05, 2018. https://medium.com/@bian0628/data-science-final-project-myers-briggs-prediction-ecfa203cef8 (accessed January 08, 2022).

[3] Abidin, N. H. Z. et al., "Improving Intelligent Personality Prediction using Myers-Briggs Type Indicator and Random Forest Classifier," IJACSA, vol. 11, no. 11, 2020, doi: 10.14569/IJACSA.2020.0111125.

[4] Novikov, P., Mararitsa, L., and Nozdrachev, V., "Inferred vs traditional personality assessment: are we predicting the same thing?," arXiv:2103.09632 [cs], March 2021. Accessed on January 08, 2022. [Online]. Available at: http://arxiv.org/abs/2103.09632

[5] Patel, S., Nimje, M., Shetty, A., and Kulkarni, S., "Personality Analysis using Social Media," International Journal of Engineering Research, vol. 9, no. 3, p. 4, 2021