International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Myers-Briggs Personality Prediction Using Machine Learning Techniques

Ankur Vipin Sheth[1], Dev Rohit Pandhare[2]

[1]Student, Computer Science Engineering, Manipal University Jaipur
[2]Student, Computer and Communication Engineering, Manipal University Jaipur

Abstract

The Myers-Briggs Type Indicator (MBTI) is a popular and widely used personality assessment tool that categorizes individuals into one of sixteen different personality types based on their preferences for certain psychological traits. This makes MBTI a fascinating subject for research in the fields of psychology and artificial intelligence. Machine learning has advanced significantly in recent years and has become a powerful tool for predicting outcomes based on data. By using machine learning algorithms to analyse data from MBTI tests, researchers can potentially identify patterns and develop models that accurately predict an individual's personality type based on their responses to the test. The ability to predict an individual's personality type could have a range of practical applications, such as in the workplace or in personalized marketing. For example, employers could use personality predictions to build more effective teams, while marketers could use them to tailor advertisements and products to specific personality types.

Keywords: MBTI

1. INTRODUCTION

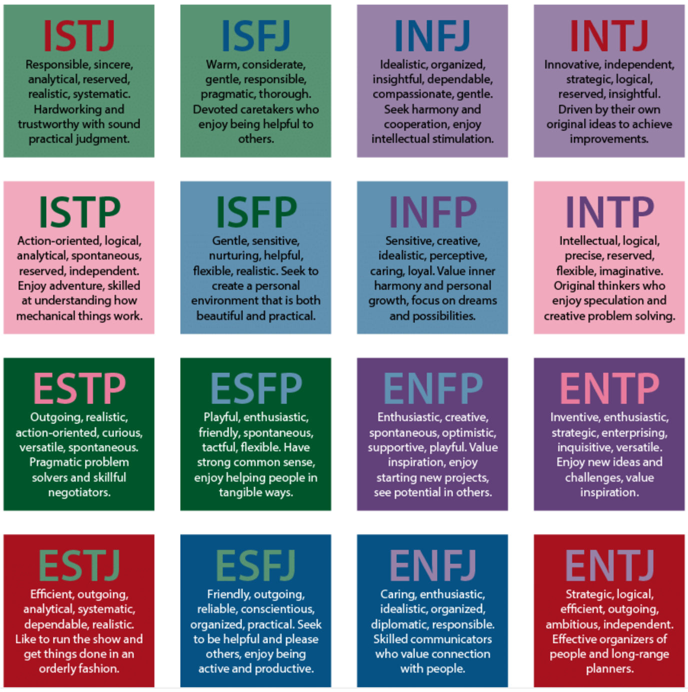
A person's personality affects all facets of life. It defines the thought, feeling, and conduct patterns that impact daily activities such as emotions, preferences, motivations, and health and forecast and characterise an individual's actions. Various personality models, such as the Myers-Briggs Type Indicator (MBTI), serve as the foundation for these applications.

Katharine Briggs and Isabel Briggs Myers, a mother-daughter team, created the Myer-Briggs type indicator, or MBTI. Making Carl G. Jung's theory of psychological types approachable and applicable to everyday life is the aim of the MBTI personality inventory. Four functional kinds were used by C. G. Jung to categorise personality types:

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

As per Jung, a person can be either extrovert or introvert plus either be sensing or intuitive plus either be thinker or feeling plus either judging or perceiving. This way there are in total sixteen different personality types in MBTI (e.g. – ESTJ, INTP, ESFP etc).

Figure 1 : MBTI Personality Types.



ISTJ
Responsible, sincere, analytical, reserved, realistic, systematic. Hardworking and trustworthy with sound practical judgment.

ISFJ
Warm, considerate, gentle, responsible, pragmatic, thorough. Devoted caretakers who enjoy being helpful to others.

INFJ
Idealistic, organized, insightful, dependable, compassionate, gentle. Seek harmony and cooperation, enjoy intellectual stimulation.

INTJ
Innovative, independent, strategic, logical, reserved, insightful. Driven by their own original ideas to achieve improvements.

ISTP
Action-oriented, logical, analytical, spontaneous, reserved, independent. Enjoy adventure, skilled at understanding how mechanical things work.

ISFP
Gentle, sensitive, nurturing, helpful, flexible, realistic. Seek to create a personal environment that is both beautiful and practical.

INFP
Sensitive, creative, idealistic, perceptive, caring, loyal. Value inner harmony and personal growth, focus on dreams and possibilities.

INTP
Intellectual, logical, precise, reserved, flexible, imaginative. Original thinkers who enjoy speculation and creative problem solving.

ESTP
Outgoing, realistic, action-oriented, curious, versatile, spontaneous. Pragmatic problem solvers and skillful negotiators.

ESFP
Playful, enthusiastic, friendly, spontaneous, tactful, flexible. Have strong common sense, enjoy helping people in tangible ways.

ENFP
Enthusiastic, creative, spontaneous, optimistic, supportive, playful. Value inspiration, enjoy starting new projects, see potential in others.

ENTP
Inventive, enthusiastic, strategic, enterprising, inquisitive, versatile. Enjoy new ideas and challenges, value inspiration.

ESTJ
Efficient, outgoing, analytical, systematic, dependable, realistic. Like to run the show and get things done in an orderly fashion.

ESFJ
Friendly, outgoing, reliable, conscientious, organized, practical. Seek to be helpful and please others, enjoy being active and productive.

ENFJ
Caring, enthusiastic, idealistic, organized, diplomatic, responsible. Skilled communicators who value connection with people.

ENTJ
Strategic, logical, efficient, outgoing, ambitious, independent. Effective organizers of people and long-range planners.

The Myers Briggs Type Indicator is a personality type system that divides a person into various distinct personalities based on introversion, intuition, thinking and perceiving capabilities. We can identify the personality of a person from the type of posts they put on social media. The objectives of the project are as follows:

- Understanding individual personality traits: The project can be aimed at understanding the personality traits of individuals such as introversion, extroversion, neuroticism, etc. This can be done by analysing text data such as social media posts, emails, or interviews.
- Improve the accuracy of the existing models: The existing models which people have developed earlier have an accuracy which we believe that is not suitable for any practical application. They all had an accuracy below 75 percent and we aim to take it above.
- Bring Real world Applicability: We want to create a model which, because of its high accuracy, will be applicable on wide variety of domains like psychology, employee recruitment, chatbot development, etc.

2. LITERATURE REVIEW

Numerous different machine learning algorithms have used by experimenters in the study of personality prediction. All exploration in this field involved several stages, including data gathering, pre-processing, rooting features and perform bracket to determine the delicacy of the model. This section highlights the Myers – Briggs Type Indicator ®(MBTI) and affiliated workshop from former experimenters using machine literacy algorithms.

DI XUE, ZHENG HONG in 2017 worked on the Personality prediction for the data on social network platform for the data with the marker distributing the data for literacy. Experimenters used the new way of prediction ways in the machine literacy model named marker distribution literacy in the field and it may be admired with the purpose. Personality prediction will be a cerebral construct which will aim to explain the new wide variety with which a human's geste and the persons pungency way of carrying social media platforms in terms of a stable and standard way of personality prediction must be important and a cerebral construct which may regard for collectively available differences in the way of prediction perspective.

Current exploration on prognosticating MBTI personality types from textual data is meagre. Nonetheless, important strides have been made in both machine literacy and neuroscience. once work has discovered the neural supplements of the Big Five personality disciplines. Specifically, the activation patterns of independent functional services in the brain, responsible for cognitive and affective processing, were shown to be statistically different among individualities differing on the colourful Big Five personality confines. Likewise, there was a functional imbrication between these linked regions responsible for differences in personality type and written communication. This justifies our attempt at prognosticating patient personality traits from textual data.

Over the once many times, numerous studies use colourful machine learning algorithm for prognosticating personality types. Studies developed a new machine literacy system for automating the process of meta programme discovery and personality type prediction grounded on the MBTI personality type index. The natural language processing toolkit (NLTK) and XGBoost, which is grounded on grade Boosting library in Python is used for enforcing machine literacy algorithms.

Exploration into prediction of personality styles from textual data is limited. still, huge strides have been taken through machine literacy in this bid. Classic neural networks and machine literacy ways have been used for textbook bracket, translation discovery, and prognosticating MBTI personality types.

3. METHODOLOGY

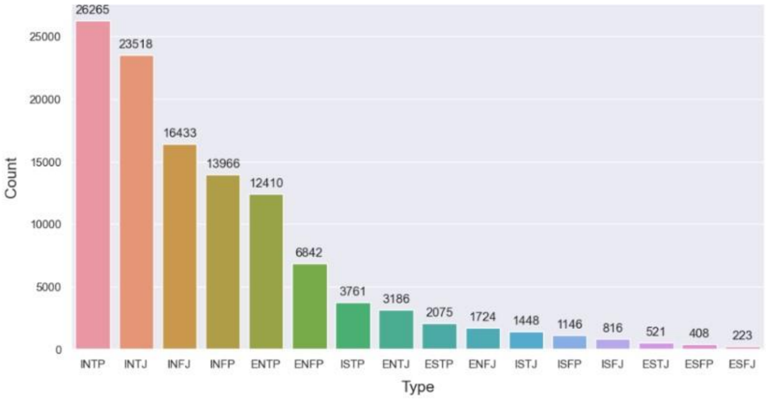
- **Dataset Description**

In this proposed methodology, the popularly used and open-source Myer-Briggs Type Indicator dataset is tested for personality analysis. The MBTI dataset is premised on the psychologist Carl G. Jung's personality traits hypothesis and distinguish a person into sixteen personality labels among four dimensions. The dimensions are listed below:

- Introversion-Extroversion (I-E)
- Sensing-Intuition (S-I)
- Thinking-Feeling (T-F)
- Judging-Perceiving (J-P)

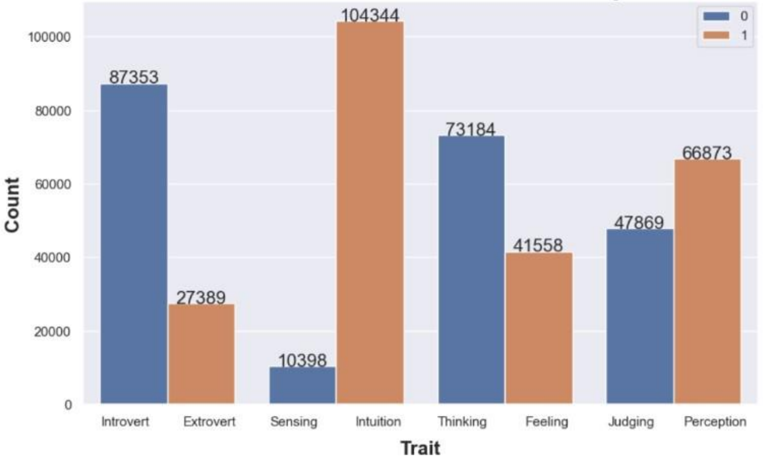
Regarding the topic of the project, two major datasets were available, one dataset of 8675 rows and the other dataset of 106067 rows. These two datasets were similar; both contained two columns, one column of independent variable called 'Posts' and the other column, the dependent variable called 'Type' which is to be predicted.

Figure 2: Number of Instances of Each MBTI Type



In the 'posts' column of both the datasets, there are approximately 30-50 tweets or 500 English words along with certain numeric characters, special symbols, links to certain websites etc. and all of the tweets were separated by 3 pipe '|||' symbols. The corresponding dependent variable is a categorical non-ordered non-ranked variable which depict one of 16 MBTI personality types in combination of four letters (so it is one word). Since this dataset has already the actual or expected outputs, so it is an example of supervised machine learning.

Figure 3: Number of Instances of Each Personality Trait



Now, most of the previous work which we have encountered were using either the former dataset (most of them) or they were using the dataset now not available on the internet or not freely accessible or from their own survey.

- **Data Pre-processing**

After analysing the dataset structure, natural language processing methods were employed to recognize the important patterns and features and removing the irrelevant and meaningless entities and keywords from the raw text document.

- **Word Embedding**

Machine learning algorithms cannot work with raw text directly; the text must be converted into numbers. Specifically, vectors of numbers. For learning from a text data, we need to apply something called as a feature extraction. Feature extraction can be done through word embedding techniques.

- **Model/Object Creation/Initialization:**

Create multiple objects of multiple models for personality type as well as individual trait-pair and initialize them with appropriate parameters. All the models were fitted in grid search on certain chosen hyper parameters suitable for the model and the problem statement.

- **Model Training:**

After we have initialized our models, we will pass our dataset to these models and train them on those datasets. Models will be trained on the dataset with a combination of values in defined space of hyper parameters.

- **Testing and Comparison:**

Models are trained to evaluate their efficiency using different performance matrices.

4. WORK DONE

Most existing work the team encountered utilized the former dataset or datasets that were no longer accessible. Opting to use the larger dataset, the team combined both datasets, resulting in a dataset of 114742 rows. Due to time constraints, the team decided to limit the dataset to 10704 rows.

The project was executed on Jupyter Notebook, an IDE/code editor available in the Anaconda distribution of Python. Jupyter Notebook facilitated the execution of Python scripts and the display of visual data, including graphs. The Anaconda distribution came pre-equipped with certain Python libraries, but additional libraries needed to be imported or installed for specific project requirements.

The initial steps involved importing the required libraries and bringing the dataset, in CSV format, into the notebook using the Pandas library. Given that the project centred around a text classification problem, pre-processing was crucial. This phase aimed to clean the 'Posts' column by removing irrelevant content, ensuring that the model would learn the necessary information in the desired manner.

The pre-processing steps included:

1. Remove Extra Whitespaces from posts: If there are more than one Whitespaces between any two words, truncate them.
2. Remove Hyper-Links from posts: Hyperlinks do not convey any meaning to either to any model, nor to the personality, so remove them.
3. Expand Contractions from posts: Slangs such as becoz for because and others must be replaced with their proper word because they are not unique, so the model might treat two different slangs (belonging to the same word) differently.
4. Remove Stop Words from posts: Stop words such as pronouns, conjunctions etc. do not convey any meaning towards personality so remove them.
5. Lower Casing each word in posts: Lower case all the words just in case some letters in a word are small or capital in the wrong places, basically not following the English syntax.
6. Remove Punctuations from posts: Punctuations again do not convey any meaning to either to any model, nor to the personality, so remove them.
7. Remove Special Characters except Exclamation symbol '!' from posts: Again, special characters do not convey any meaning to either to any model, nor to the personality, so remove them except the exclamation symbol because they represent emotion so they might have an impact on personality of a person. So, keep it.
8. Lemmatization on posts: Lemmatization means bringing a word to its root.
9. For example – lemmatization of walking, walked, walk will give us walk.
10. This is done to reduce unnecessary treating similar words differently.
11. Remove Numbers from posts: Numbers again do not convey any meaning to either to any model, nor to the personality, so remove them.
12. Remove Non-English Vocabulary words: It might be possible that lemmatization might have created some words that have no meaning, so it is better to remove them for reducing unnecessary complexity of the models.
13. Remove MBTI Personality words: It was found that certain tweets contained MBTI Personality words and therefore it was needed to remove them so that our model does not ‘emphasize’ on those words only; and focuses on the whole corpus.

After the pre-processing phase is done, the 16 MBTI personalities are made of eight traits (4 pairs of mutually exclusive traits). So, they can be split into the 4-letter category of personality into four columns of binary data (0 or 1) as per what one of the two personalities they include of the pair; one column for each pair.

Next, the analysis focused on evaluating the dataset's distribution of available data to determine whether it exhibited balance or imbalance. Leveraging Python's Matplotlib and Seaborn libraries, the visual representation included graphs for the sixteen categories, detailing the total record count for each category. Additionally, four pairs of traits (representing the newly added columns) were graphically presented to highlight the total records containing either of the two mutually exclusive traits. Following this comprehensive analysis, it was observed that the dataset displayed a notable imbalance.

To achieve a balanced dataset, the first step involved extracting all rows from the imbalanced dataset and creating separate Excel files for each personality type. Each Excel file contained only the corresponding records of that personality type. Subsequently, these records from the sixteen Excel files were gradually incorporated into the original dataset. This process aimed to create a highly balanced dataset, specifically targeting the first quarter of the total number of records.

Moving on to the training phase, the dataset was split into an 80:20 ratio, allocating 80 percent of the data for training the model and reserving the remaining 20 percent for evaluating model accuracy.

Given that machine learning algorithms cannot directly handle raw text, the next step involved converting the text into numerical representations, specifically matrices of numbers. Feature extraction was carried out through word embedding techniques, with a focus on experimenting with Bag-of-Words and TF-IDF. These techniques returned vectors of numerical values reflecting various linguistic properties of the text.

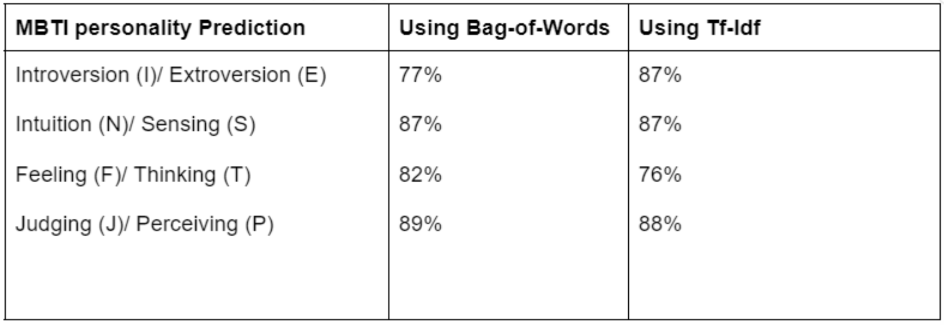
Following feature extraction, different models were applied to the dataset, specifically to the generated vectors, allowing them to learn underlying patterns. An essential step in the process was the grid search, where hyper parameters in each model were systematically evaluated. This involved exploring a limited set of values for these hyper parameters to identify the combination that resulted in the best model accuracy.

Multiple models were employed in the study which includes Multinomial Naive-Bayes Model, Random Forest Classifier, Support Vector Machine Classifier, K Nearest Neighbour Classifier, and Decision Tree Classifier. Once the models were trained, they were evaluated for accuracy.

5. RESULTS

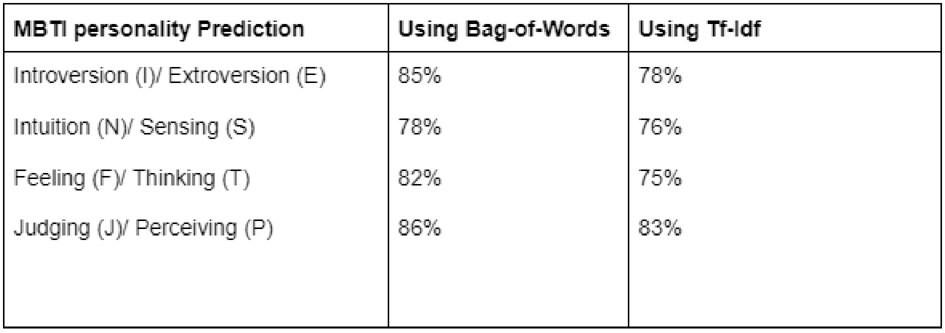
The accuracy of every model using both text features i.e. Bag-of-Words and Tf-idf is listed in the table below:

Table 1: Accuracy using Multinomial Naïve Bayes Model



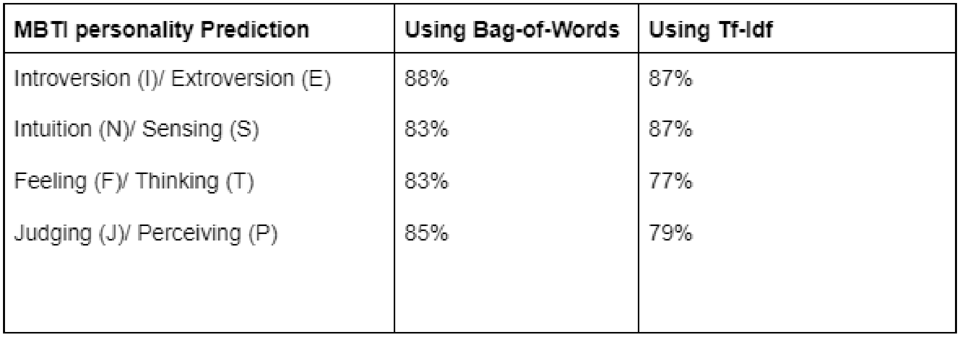
MBTI personality Prediction	Using Bag-of-Words	Using Tf-Idf
Introversion (I)/ Extroversion (E)	77%	87%
Intuition (N)/ Sensing (S)	87%	87%
Feeling (F)/ Thinking (T)	82%	76%
Judging (J)/ Perceiving (P)	89%	88%

Table 2: Accuracy using K Nearest Neighbours Model



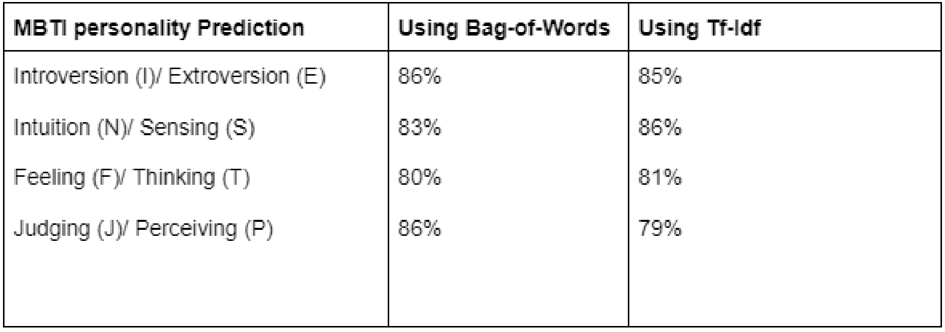
MBTI personality Prediction	Using Bag-of-Words	Using Tf-Idf
Introversion (I)/ Extroversion (E)	85%	78%
Intuition (N)/ Sensing (S)	78%	76%
Feeling (F)/ Thinking (T)	82%	75%
Judging (J)/ Perceiving (P)	86%	83%

Table 3: Accuracy using Support Vector Machine Model



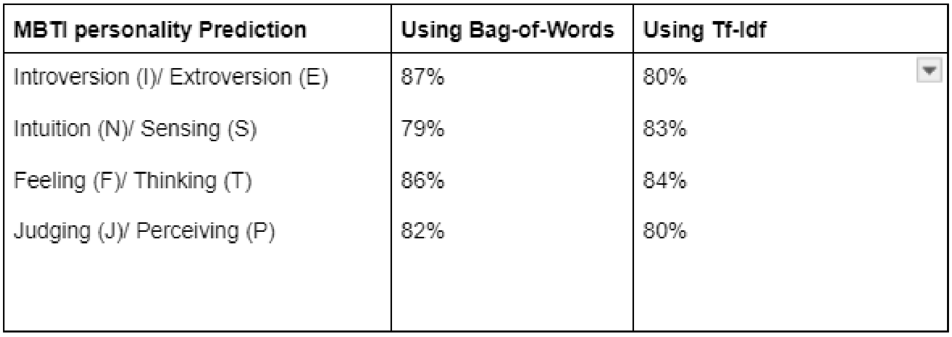
MBTI personality Prediction	Using Bag-of-Words	Using Tf-Idf
Introversion (I)/ Extroversion (E)	88%	87%
Intuition (N)/ Sensing (S)	83%	87%
Feeling (F)/ Thinking (T)	83%	77%
Judging (J)/ Perceiving (P)	85%	79%

Table 4: Accuracy using Decision Trees Model



MBTI personality Prediction	Using Bag-of-Words	Using Tf-Idf
Introversion (I)/ Extroversion (E)	86%	85%
Intuition (N)/ Sensing (S)	83%	86%
Feeling (F)/ Thinking (T)	80%	81%
Judging (J)/ Perceiving (P)	86%	79%

Table 5: Accuracy using Random Forest Model



MBTI personality Prediction	Using Bag-of-Words	Using Tf-Idf
Introversion (I)/ Extroversion (E)	87%	80%
Intuition (N)/ Sensing (S)	79%	83%
Feeling (F)/ Thinking (T)	86%	84%
Judging (J)/ Perceiving (P)	82%	80%

6. CONCLUSION AND FUTURE SCOPE

In summary, this study aims to explore the fascinating field of personality prediction using the Myers-Briggs Type Indicator (MBTI) through the lens of machine learning. Our research leveraged a large dataset spanning over 1 million social media tweets, examining subtle patterns embedded in language to decipher personality traits. This study was conducted using careful preprocessing techniques to transform raw data into a structured format suitable for training and testing machine learning models. The use of various preprocessing techniques played an important role in refining the dataset. Tokenization, stemming, and stop word processing were essential steps to convert raw text data into a format suitable for model ingestion. This preprocessing step significantly improved the model's ability to detect subtle linguistic nuances and extract meaningful features related to personality types.

This study used a comprehensive suite of machine learning models, including Naive Bayes, Support Vector Machine (SVM), Random Forest, Decision Tree, and k-Nearest Neighbours (KNN). Each model brought to the fore its unique strengths and offered diverse approaches to address the multifaceted challenges of personality prediction. The experimental results reflect different levels of effectiveness of the models and reveal their respective capabilities and limitations in this particular situation. In particular, Naive Bayes models have demonstrated commendable performance, leveraging their simplicity and efficiency to provide robust predictions. SVM, with its ability to delineate complex decisions, also showed remarkable results, highlighting the importance of considering nonlinear relationships in the data. Random forests and decision trees were distinguished by their ensemble nature and differed in computational complexity, but were better at capturing complex patterns. On the other hand, ANNs that rely on the proximity of data points have demonstrated the ability to recognize local patterns within a data set.

However, the model's predictive performance also highlighted the inherent challenges in disentangling the complex relationship between language and personality.

The nature of social media discourse, characterized by abbreviations, slang, and contextual expressions, poses a major obstacle to accurately predicting personality types based solely on textual content. This work provides a stepping stone into the fascinating field of personality prediction using machine learning, and its findings and methods pave the way for further exploration and refinement.

As technology advances and datasets become more pervasive, the possibility of uncovering the complex relationships between language and personality becomes an increasing challenge, prompting researchers to embark on a journey of discovery and innovation.

7. REFERENCES

1. https://www.kaggle.com/datasets/datasnaek/mbti-type
2. https://www.kaggle.com/datasets/zeyadkhalid/mbti-personality-types-500-dataset
3. Nishita Vaddem, Pooja Agarwal (2020) - “Myers Briggs Personality Prediction using Machine Learning Techniques”, International Journal of Computer Applications (0975 – 8887) Volume 175
4. Prajwal Kaushal, Nithin Bharadwaj B. P., Pranav M. S., Koushik S., and Anjan K. Koundinya (2021) – “Myers-briggs Personality Prediction and Sentiment Analysis of Twitter using Machine Learning Classifiers and BERT”, I.J. Information Technology and Computer Science DOI: 10.5815/ijitcs.2021.06.04
5. Gayathri Kadam C., Dr. D. Preethi (2022) – “Comparative Study of Personality Prediction Using Machine Learning Algorithms”, International Journal of Science and Research (IJSR) ISSN: 2319-7064 SJIF (2022): 7.942
6. Hernandez, Rayne, Knight, Ian Scott – “Predicting Myers-Briggs Type Indicator with Text Classification”, https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6839354.pdf

7. Sakdipat Ontoum, Jonathan H. Chan (2022) – "Personality Type Based on Myers-Briggs Type Indicator with Text Posting Style by using Traditional and Deep Learning", https://arxiv.org/pdf/2201.08717.pdf
8. Mohammad Hossein Amirhosseini, Hassan Kazemian (2020) – "Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator", https://www.mdpi.com/2414-4088/4/1/9
9. Dr Swetha P., Sunil B., Viresh, Vivek S. Shyavi (2022) – "Personality Prediction with social media using Machine Learning", International Research Journal of Engineering and Technology (IRJET) Volume: 09 Issue: 06 e-ISSN: 2395-0056 p-ISSN: 2395-0072
10. Alam Sher Khan, Hussain Ahmad, Muhammad Zubair Asghar (2020) – "Personality Classification from Online Text using Machine Learning Approach", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol.11
11. Shruti Garg, Ashwani Garg (2021) – "Comparison of machine learning algorithms for content-based personality resolution of tweets", https://www.sciencedirect.com/science/article/pii/S2590291121000747