International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1460 - 1464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0208

Personality Prediction Using Machine Learning

Swathy Sree C S [1], Akshayavarthini B [2], Sruthi K [3], Supraja G [4], Dr. Tamilvizhi T [5]
[1,2,3,4]UG Scholar, Department of Computer science and Engineering, Panimalar Engineering college, chennai, India.
[5]Faculty, Department of Computer science and Engineering, Panimalar Engineering college, chennai, India.
Emails: *swathysre@gmail.com[1], akshayavarthiniav@gmail.com[2], sruthi28705@gmail.com[3], suprajasupraja220@gmail.com[4], tamilvizhi.phd.it@gmail.com[5]*

Abstract
We discover the utility of diverse machine mastering algorithms to are expecting persona sorts primarily based on textual facts using the Myers-Briggs Type Indicator (MBTI) dataset. The dataset incorporates user-generated posts, which are pre-processed through a sequence of herbal language processing (NLP) strategies, including text normalization, stopword elimination, and lemmatization. We appoint the TF-IDF (Term Frequency-Inverse Document Frequency) method to convert the textual information into numerical features. Several classifiers—Gaussian Naive Bayes, Multinomial Naive Bayes, Random Forest, XGBoost, LightGBM, Support Vector Machine (SVM), and Logistic Regression—are skilled and evaluated to predict the MBTI persona kinds. The models are in comparison primarily based on accuracy and certain type reviews. Among the models tested, the XGBoost classifier outperforms others with an accuracy of 67.55%, demonstrating its effectiveness for this multi-elegance text class hassle. This venture highlights the ability of device getting to know in personality prediction from textual records and offers a comparative analysis of diverse type algorithms for this purpose.
Keywords: *Personality Prediction MBTI Classification, Text Preprocessing, Natural Language Processing (NLP), Machine Learning Models.*

1. Introduction
Attempts to predict male or female genders from text data often use standard NLP methods and simple machine gaining knowledge of models which includes TF-IDF and Naive Bayes. These capture phrase frequencies however often miss deeper which means styles wanted for correct gender prediction. These fashions tend to simplify the trouble by specializing in constrained features. This results in much less accuracy and potential to work properly with specific textual content sorts. Current structures additionally take a reactive method. They are expecting gender primarily based on collected information in preference to changing in real-time. Also many models don't use advanced NLP strategies like deep gaining knowledge of. These can enhance general performance and capability to handle massive amounts of information. [1]

2. Literature Survey
"Smart-Hire Personality Prediction Using ML" (May 2023) via Isha Gupta and Manasvi Jain: This have a look at underscores the practical implications of character type sorts through ML predictions, can actively interact in self-development efforts. The paper emphasizes the ability effect of such insights on personal and professional improvement. It shows that individuals, upon coming across their character [2].

"A Study on Personality Prediction & Classification Using Data Mining Algorithms" (August 2022) by Pavitha N., Somesh Kamnapure, and Ayush Gundawa:
Highlighting the importance of character in private and expert contexts, this work explores information mining algorithms to unexpectedly predict and categorize an person's persona. The researchers suggest for integrating ML strategies, especially via intuitive input strategies like questionnaires, to decorate prediction performance.

"Language Style Matters: Personality Prediction from Textual Styles Learning" (November 2023)

e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1460 - 1464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0208

by Meiling Li and Hezi Liu:
This research delves into psycholinguistic literature, emphasizing the role of language styles in unveiling personality factors. The paper contends that language patterns provide insights into users' personalities, including social networks and intellectual health. Textual styles learning is supplied as a treasured approach for character prediction.

"Personality Prediction using Machine Learning" (June 2022) via Hima Vijay and Neenu Sebastian:
Acknowledging the importance of sorting people primarily based on persona types, this work emphasizes the packages of ML algorithms in accomplishing this aim. The paper contributes to the literature by using exploring the capacity advantages and implications of persona prediction the use of ML.

"Personality Prediction from Textual Data" by Plank and Hovy (2015)
Early studies has centered on using textual information from social media, blogs, and forums to expect personality types. Techniques like Bag-of-Words (BoW) and TF-IDF were typically implemented to symbolize textual content numerically. For instance, he used TF-IDF and word embeddings to are expecting Big Five persona trends from Twitter posts, displaying that personality prediction from quick textual content is possible. [3]

"Machine Learning Algorithms for Personality Classification" by using Verhoeven et al. (2016) and Gjurković and Šnajder (2018):
Classical system learning models together with Naive Bayes, SVM, and Random Forest had been extensively carried out for personality classification. Studies by means of them proven slight fulfillment in using these models for character kind prediction, with Random Forest and SVM often outperforming simpler fashions like Naive Bayes.

"Deep Learning Approaches" by way of Kim et al. (2020):
Recent research has become too deep mastering, specifically RNNs and transformer models like BERT. LSTM networks were used to seize textual content series styles, improving persona class, whilst transformers inclusive of BERT offer brand new results by shooting complex contextual relationships. These models, however, require big datasets and massive computational electricity.

"Ensemble Learning and Boosting Methods" via Li et al. (2019):
Boosting algorithms like XGBoost and LightGBM have grow to be popular in persona prediction because of their capability to handle high-dimensional records and decrease both bias and variance. He have shown that these models outperform traditional classifiers, mainly when mixed with sturdy characteristic extraction techniques like TF-IDF.

3. Existing System

Attempts to predict male or female genders from text data often use standard NLP methods and simple machine gaining knowledge of models which includes TF-IDF and Naive Bayes. These capture phrase frequencies however often miss deeper which means styles wanted for correct gender prediction. These fashions tend to simplify the trouble by specializing in constrained features. This results in much less accuracy and potential to work properly with specific textual content sorts. Current structures additionally take a reactive method. They are expecting gender primarily based on collected information in preference to changing in real-time. Also many models don't use advanced NLP strategies like deep gaining knowledge of. These can enhance general performance and capability to handle massive amounts of information. [4]

4. Proposed System

The new device to expect personality types from textual content information improves on older techniques through using modern NLP and modern system learning. It applies smart textual content representation strategies like word embeddings (Word2Vec GloVe) and context-aware embeddings from transformer models (BERT) to comprehend deeper meanings. Advanced mastering algorithms along with LSTM and BERT, assist understand complex language styles, which makes predictions greater correct. This machine can expect personalities in real-time and adjusts to new inputs. It additionally attempts to paintings nicely with unique text resources thru switch mastering and provides more functions like psycholinguistic markers, sentiment analysis, and tone detection for a more complete persona classification. This method leads to extra precise bendy, and expandable personality predictions than older methods. (Figure 1)

5. Architecture Diagram

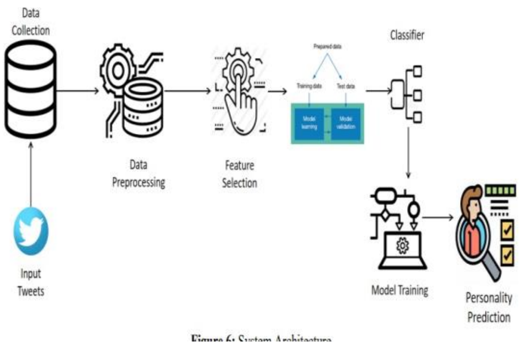

Figure 1 Architecture Diagram of the Prediction Process

- **Data Collection:** The device starts evolved by way of gathering input statistics inside the form of tweets from Twitter. These tweets function the uncooked textual content records for personality prediction**.**
- **Data Preprocessing:** The amassed records undergoes preprocessing, which includes steps inclusive of cleansing, normalization, stopword elimination, and lemmatization. This level prepares the data for characteristic extraction by getting rid of beside the point or redundant facts.
- **Feature Selection:** After preprocessing, vital capabilities are extracted from the textual content the usage of techniques inclusive of TF-IDF, word embeddings, or other techniques to transform the textual records into numerical formats. This step ensures that simplest the most applicable features are fed into the model for education**.**
- **Model Training:** The prepared facts is then cut up into training and testing datasets. The schooling information is used to build the system studying fashions, at the same time as the take a look at records is reserved for validation. Various classifiers are implemented, which include Naive Bayes, Random Forest, XGBoost, and others to research the patterns within the information.
- **Classifier:**The decide on classifier version processes the statistics and is trained to understand styles that correspond to distinctive character kinds based on the tweets. The skilled model is confirmed to make sure accuracy.
- **Personality Prediction: based** Finally, the skilled model is used to are expecting the character kind of a person primarily based on them enter tweets. The output is a character prediction, indicating the most probably persona type primarily based on the textual content information, permitting insights into the user's psychological tendencies. [5]

6. Algorithm

Step 1: Data Loading and Preprocessing

Load dataset from CSV
DATA <- load_csv("mbti_dataset.csv")
Rename columns (if needed) and format Date column (if present)
DATA <- rename_columns(DATA, {"col1": "Post", "col2": "Type"}) Data preprocessing steps (if relevant, depending on dataset)
For personality prediction, preprocess text data
DATA["Post"] <- preprocess_text(DATA["Post"])
Define feature matrix X (e.g., textual posts converted using TF-IDF) and target y (e.g., MBTI personality types)
X <- convert_to_TFIDF(DATA["Post"])
y <- DATA["Type"]

Step 2: Label Encoding

Initialize Label Encoder
LABEL_ENCODER <- LabelEncoder()
Fit the label encoder on target variable y (Personality Types)
y_encoded <- LABEL_ENCODER.fit_transform(y)

Step 3: Data Splitting

Split dataset into training and testing sets using 80/20 ratio
X_train, X_test, y_train, y_test <- train_test_split(X, y_encoded, test_size=0.2)

Step 4: Model Training

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1460 - 1464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0208

Initialize XGBoost Classifier with proper hyperparameters

```
model <- XGBClassifier(use_label_encoder=False, eval_metric='mlogloss')
```

Train XGBoost model on the training data

```
model.fit(X_train, y_train)
```

Step 5: Make Predictions

Predict on the test data

```
y_pred <- model.predict(X_test)
```

Step 6: Evaluate Model Performance

Calculate accuracy, precision, recall, and F1 score

```
accuracy <- calculate_accuracy(y_test, y_pred)
precision <- calculate_precision(y_test, y_pred)
recall <- calculate_recall(y_test, y_pred)
f1_score <- calculate_f1(y_test, y_pred)
```

Optionally, plot the confusion matrix to evaluate classification

```
plot_confusion_matrix(y_test, y_pred)
```

Step 7: Future Predictions

Load or define new data for future predictions

```
NEW_DATA <- load_new_data("new_tweets.csv")
```

Preprocess new data (same steps as training data)

```
NEW_DATA["Post"] <- preprocess_text(NEW_DATA["Post"])
new_X <- convert_to_TFIDF(NEW_DATA["Post"])
```

Predict personality types for new data

```
future_predictions <- model.predict(new_X)
```

Decode predicted numeric labels back to original form

```
future_predictions_decoded <- LABEL_ENCODER.inverse_transform(future_predictions)
```

Step 8: Save or Display Results

Append predicted labels to new data

```
NEW_DATA["Predicted Personality"] <- future_predictions_decoded
```

Optionally save the results to a CSV file for future use or reporting

```
save_csv(NEW_DATA, "predicted_personalities.csv")
```

7. Result and Discussion

The experiments compare multiple machine learning classifiers based on accuracy, precision, recall, and F1-score. The dataset is split into an 80-20 ratio for training and testing. Table 1 presents the comparative performance of the models:

Table 1 Accuracy of Different ML Models

Model	Accuracy (%)	Precision	Recall	F1-Score
Naïve Bayes	38.0%	40.2%	37.5%	38.8%
SVM	65.4%	66.0%	64.8%	65.4%
Random Forest	66.0%	67.2%	65.9%	66.5%
LightGBM	67.4%	68.1%	66.9%	67.5%
XGBoost	**67.55%**	**68.2%**	**67.1%**	**67.6%**

Among the tested models, XGBoost demonstrated the best performance, achieving an accuracy of 67.55%. The confusion matrix analysis revealed that certain personality types, such as Introverted-Sensing-Feeling-Judging (ISFJ) and Extroverted-Intuitive-Thinking-Perceiving (ENTP), were often misclassified due to linguistic similarities.This superior performance can be attributed to its ability to handle feature interactions effectively, reduce overfitting using regularization, and optimize computational efficiency through parallel tree boosting. Additionally, XGBoost's ability to capture non-linear relationships in textual data makes it well-suited for personality classification tasks. (Figure 2) These findings indicate that while ensemble learning techniques outperform traditional classifiers, deep learning approaches such as transformer models (e.g., BERT) may provide additional improvements. Future research will explore hybrid architectures combining traditional ML models with contextual embeddings to capture deeper semantic meanings from textual data. Additionally, real-time personality prediction applications will be considered to enhance user interaction in fields such as personalized marketing, recruitment, and mental health analysis.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1460 - 1464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0208

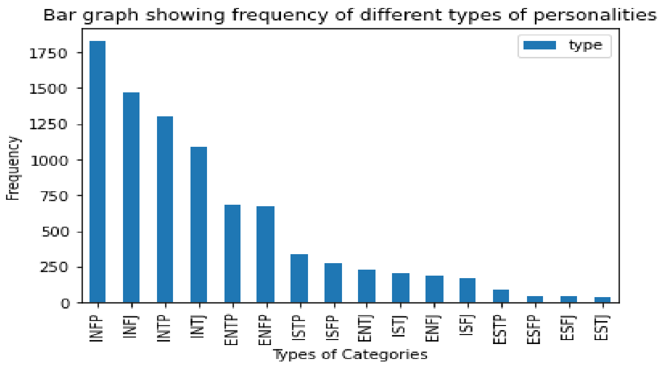

Figure 2 Bar Graph Showing Different Types of Personalities

Conclusion

This undertaking aimed to categorise Myers-Briggs Type Indicator (MBTI) persona sorts the usage of textual records from on line posts through the software of numerous device getting to know models. After preprocessing the information, inclusive of text cleansing, stop word removal, and function extraction the use of TF-IDF, more than one fashions had been skilled and evaluated for overall performance.

The fashions explored encompass Gaussian Naive Bayes, Multinomial Naive Bayes, Random Forest, Support Vector Machine (SVM), Logistic Regression, LightGBM, and XGBoost. Among these models, XGBoost executed the best test accuracy of sixtyseven.55%, followed carefully by LightGBM with 67.38%. On the alternative hand, less complicated models like Gaussian Naive Bayes and Multinomial Naive Bayes carried out poorly, highlighting that extra complicated algorithms are wished for this multi-elegance, text-based totally classification trouble.

Future Work

Incorporating deep getting to know fashions, which include recurrent neural networks (RNNs) or transformer-based totally models (eg: BERT), could further enhance prediction accuracy with the aid of better shooting the sequential and contextual relationships in the text.

- Addressing the imbalanced nature of the dataset ought to improve overall performance for underrepresented personality kinds, likely through strategies like records augmentation or artificial facts technology.
- Further exploration of advanced ensemble strategies and hyperparameter tuning should push the models towards higher accuracy and generalization.

References

[1]. Faisal, L. (2024) “Everyone is an Alien Somewhere: Investigating the Skills Needed for Effective Learning Advising” 21(9),3074.

[2]. Beckley, J. (2024) “Personality prediction using medias” Do emotion motives constrain the selections of visual, auditory, and taste stimuli to up- and down- regulate emotions. [IEEE].

[3]. Meiling Li and Hezi Liu (2023) “Language Style Matters: Personality Prediction from Textual Styles Learning” pp.49-67.

[4]. H. N. Desai and R. Patel (2020) "A Study of Data Mining Methods for Prediction of Personality Traits," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 58- 64.

[5]. S. Wang, L. Cui, L. Liu, X. Lu and Q. Li (2020), “Personality Traits Prediction Based on Users: Digital Footprints in Social Networks via Attention RNN," pp. 78-92.