**PERSONALITY PREDICION ANALYSIS USING MACHINE LEARNING**

**SWATHY SREE C S AKSHAYAVARTHINI A SRUTHI K**

**Computer Science Engineering Computer Science Engineering Computer Science Engineering**

**Panimalar Engineering College Panimalar Engineering College Panimalar Engineering College**

**Chennai, India Chennai, India Chennai, India**

[**swathysre@gmail.com**](mailto:swathysre@gmail.com)[**akshayavarthiniav@gmail.com**](mailto:akshayavarthiniav@gmail.com) **sruthi28705@gmail.com**

**SUPRAJA G Dr. TAMILVIZHI T**

**Computer Science and Engineering Computer Engineering College**

**Panimalar Engineering College Panimalar Engineering College**

**Chennai, India Chennai, India**

**Suprajasupra220@gmail.com tamilvizhi.phd.it@gmail.com**

**-------------------------------------------------------------------------------------------------------------**

***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 preprocessed thru 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**

**I.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 frequency 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 useadvanced NLP strategies like deep gaining knowledge of. These can enhance general performance and capability to handle massive amounts of information

**II.LITERATURE SURVEY**

**1**. "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.

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

4. "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.

5. “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.

6. “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.

7. “Deep Learning Approaches” by way of **Kim et al.** (2020):

Recent research has became to 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.

8. “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.

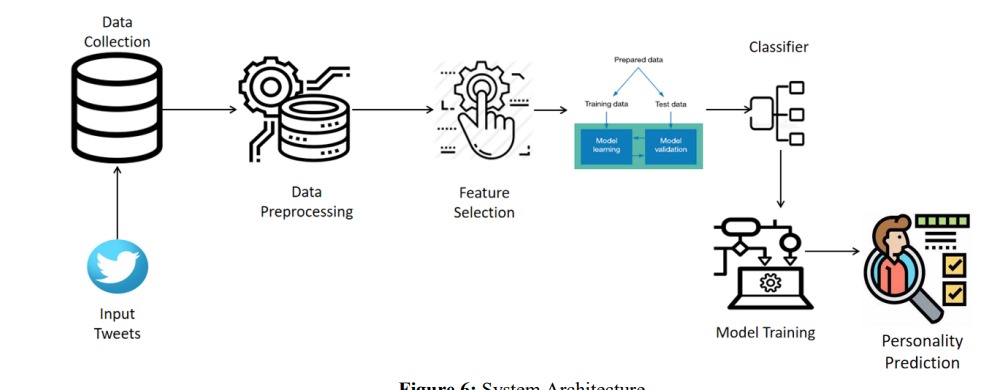
**III.EXISTING SYSTEM**

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 frequency 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.

**IV.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, andexpandable personality predictions than older methods.

**V.ARCHITECTURE DIAGRAM**

****

**Data Collection:**

The device starts offevolved 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 decided 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 their enter tweets. The output is a character prediction, indicating the most probably persona type primarilybased on the textual content information, permitting insights into the user's psychological tendencies**.**

**VI. 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

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

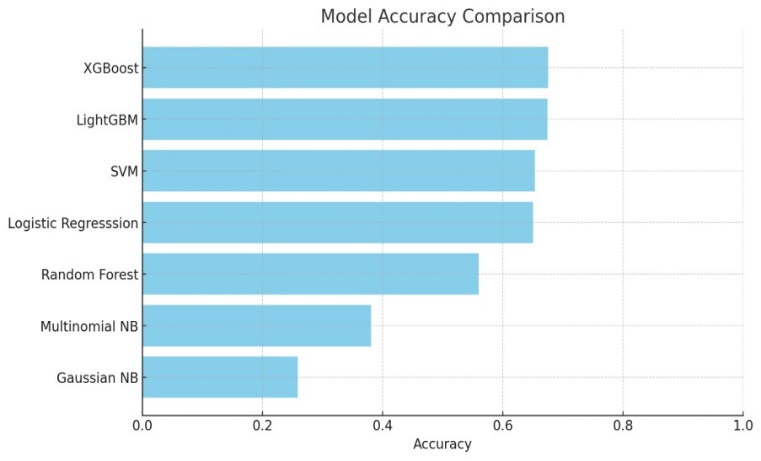
NEW\_DATA["Predicted Personality"] <- future\_predictions\_decoded

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

save\_csv(NEW\_DATA, "predicted\_personalities.csv")

**VII. RESULT AND DISCUSSION**

In this task, we evaluated numerous machine mastering classifiers to expect persona sorts based totally on the Myers-Briggs Type Indicator (MBTI) from tweets. The performance of classifiers discovered sizeable variations in accuracy, with XGBoost attaining the highest accuracy of 67. 5%, accompanied closely by using LightGBM at 67.4%. Other models like Support Vector Machine and Logistic Regression achieved reasonably properly, with accuracies of sixty five.4% and sixty five.Zero%, respectively, even as Gaussian and Multinomial Naive Bayes confirmed confined effectiveness with accuracies of 25.9% and 38.Zero%. The analysis of function importance highlighted key linguistic trends that correlate with character kinds, demonstrating the capacity of the usage of advanced herbal language processing strategies for personality prediction. Despite challenges which include records nice and class imbalance, the findings underscore the feasibility and effectiveness of system studying in expertise and predicting personality from textual records.

****

**VIII. 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, stopword 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.

**IX. FUTURE WORK**

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

1.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.

2.Further exploration of advanced ensemble strategies and hyperparameter tuning should push the models towards higher accuracy and generalization.

**X. 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.