Predicting MBTI Personality Types from Text Using Logistic Regression

Group No. 07

Shashank Singh(202401100300227)
Sneha Yadav(202401100300249)
Suryank Batham(202401100300257)
Tanishk Gupta(202401100300259)
Uday Singh Kushwaha(202401100300269)

Project Overview: The Challenge of Personality Classification

Problem Statement

The core challenge of this project is to accurately classify MBTI personality types from unstructured written text. This involves transforming raw text into a format suitable for machine learning algorithms, capturing the subtle linguistic nuances that differentiate personality types.

Proposed Solution

- Implement robust text preprocessing to clean and normalize raw user data.
- Apply Natural Language Understanding (NLU) techniques to extract meaningful features from text.
- Train and evaluate a supervised machine learning model for personality prediction.

Key Objectives: Guiding Our Predictive Journey

Classify MBTI Personality Types

Predict one of the 16 MBTI personality types using NLP and machine learning techniques, aiming for high accuracy in discerning personality from written expression.

Apply NLP Techniques for Text Preprocessing

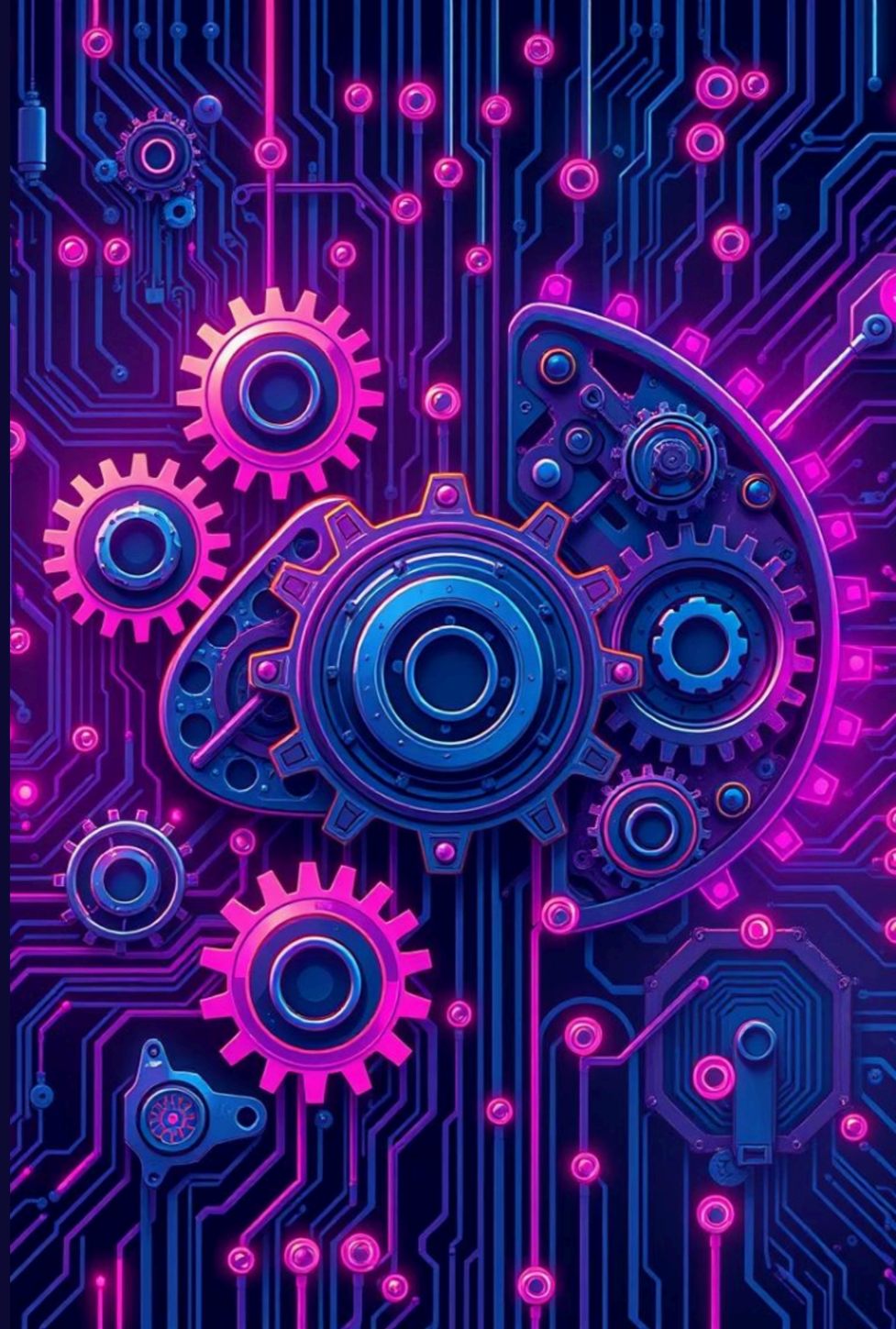
Clean and normalize raw user text data by removing punctuation, stopwords, and URLs, and applying lemmatization to standardize word forms.

Convert Textual Data to Numerical Features

Transform user posts into structured numerical data using TF-IDF vectorization, essential for feeding into machine learning models.

Build & Train a Supervised Classification Model

Develop and train a Logistic Regression model to learn patterns between text features and their corresponding personality types.



Methodology: A Structured Approach to Prediction

1

Data Loading

Upload and load the MBTI dataset containing text posts and associated personality types.

2

Text Preprocessing

Convert text to lowercase, remove URLs, punctuation, extra spaces, and stopwords.

3

Feature Extraction

Transform cleaned text into numerical vectors using TF-IDF to quantify word importance.

4

Model Training

Train a Logistic Regression classifier on the prepared training data.

5

Evaluation & Prediction

Test the model on unseen data, measure performance, and use it to predict new inputs.

Our methodology ensures a systematic progression from raw data to a trained predictive model, emphasizing rigorous preprocessing and robust evaluation.

Data Preprocessing: Refining Raw Text for Analysis

Loading Data

The MBTI dataset, in CSV format, was efficiently loaded using the Pandas library, preparing it for subsequent processing steps.

Text Cleaning

Through regular expressions, text was converted to lowercase, and extraneous elements like URLs, punctuation, digits, and extra spaces were meticulously removed.

Stopwords Removal & Lemmatization

Common English stopwords were filtered out using NLTK to reduce noise, and words were reduced to their base forms via WordNetLemmatizer for consistency.

TF-IDF Vectorization

Cleaned text was transformed into numerical features using TF-IDF, selecting the top 5000 most relevant words to represent the textual data.

This meticulous preprocessing ensures that the data fed into the model is clean, consistent, and optimized for accurate feature extraction, minimizing noise and maximizing signal.

Made with GAMMA

Model Implementation: Building the Predictive Engine

Train-Test Split

The dataset was partitioned into an 80% training set and a 20% testing set using **train_test_split()**. This division is crucial for evaluating the model's generalization capability on unseen data.

Model Training

A **LogisticRegression()** model was instantiated and trained on the TF-IDF vectorized features. During this phase, the model learns the complex relationships between the textual patterns and the corresponding MBTI personality types.

Prediction & Evaluation

The trained model then generated predictions for the test data. Its performance was rigorously assessed using the **Accuracy Score** for overall correctness and a **Classification Report** to detail precision, recall, and F1-score across all personality classes.

The implementation phase focuses on standard machine learning practices to ensure the model is robustly trained and reliably evaluated, providing confidence in its predictive power.

Results & Analysis: Performance of the Logistic Regression Model

70%

Accuracy Achieved

The Logistic Regression model, utilizing TF-IDF features, demonstrated approximately 70% accuracy in predicting MBTI types from text posts.

16

MBTI Types Predicted

The model successfully attempted to classify all 16 distinct MBTI personality types present in the dataset.

0.7

F1-Score for Major Classes

The classification report revealed F1-scores around 0.7 for well-represented classes, indicating reasonable balance between precision and recall.

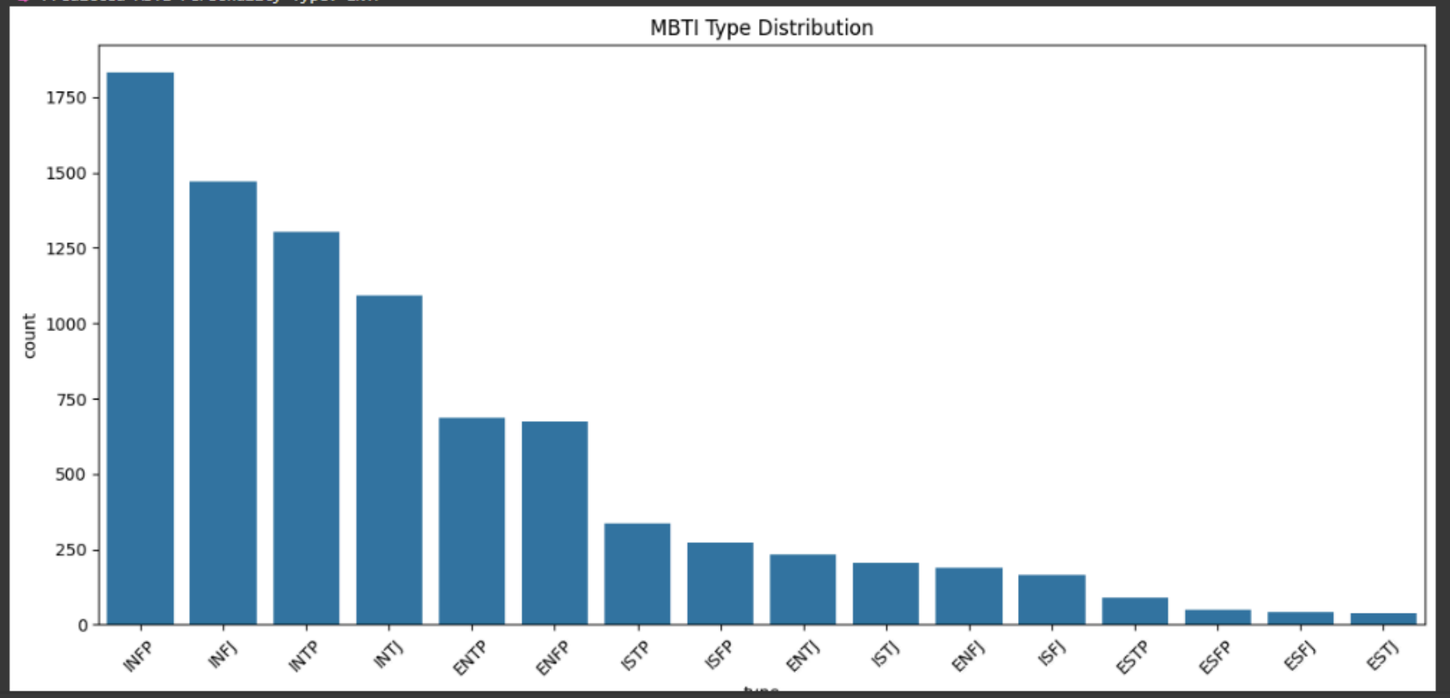
Despite challenges such as class imbalance, the model performed reasonably well, establishing a solid baseline for text-based personality prediction. This indicates the effectiveness of our preprocessing steps in cleaning the data for optimal model performance.

Output

```
✅ Accuracy: 0.6352

🖍️ Enter your text (a paragraph that reflects your thoughts/behavior):
> I enjoy deep conversations about abstract topics and often reflect on my thoughts

🧠 Predicted MBTI Personality Type: INTP
```



Conclusion & Future Enhancements

Summary of Achievements

- Successfully built a baseline MBTI personality prediction model.
- Demonstrated effective text preprocessing and TF-IDF feature extraction.
- Achieved good accuracy and practical prediction capability with Logistic Regression.

Future Enhancements

- Integrate more advanced NLP models (e.g., Transformers) for complex language patterns.
- Implement techniques to address class imbalance and improve performance on minority classes.
- Expand the dataset size and diversity for more robust generalization.
- Explore transfer learning for enhanced feature representation.

While the current model serves as a strong foundation, there is ample room for improvement. Incorporating state-of-the-art NLP techniques and further refining data handling strategies will undoubtedly lead to a more sophisticated and accurate personality prediction system. The project provides a clear path for continued research and development in this fascinating interdisciplinary field.