International Journal of Development Mathematics Vol 1 Issue 3, Page No. 174 - 185

INTERNATIONAL JOURNAL OF DEVELOPMENT MATHEMATICS

ISSN: 3026-8656 (Print) | 3026-8699 (Online)

journal homepage: https://ijdm.org.ng/index.php/Journals

Smart Hostel Allocation System Using Naïve Bayes Algorithm

Mikailu Habila[a], Jeremiah Amos[a], Chukuemeka N. Florence[b], Mujtaba K. Tahir[a], Auwal Aminu[c]

[a]*Department of Computer Science, Nigerian Army University Biu, Nigeria*

[b]*ICT Unit, Bingham University Karu, Nasarawa State, Nigeria*

[c]*Department of Software Engineering, Nigerian Army University Biu, Nigeria*

ARTICLE INFO

Article history:
Received 17 April, 2024
Received in revised form 15 July, 2024
Accepted 19 July, 2024

Keywords:
Hostel room, Machine learning, Personality traits, Prediction, Smart systems.

MSC 2020 Subject classification:
90B50, 68T30

ABSTRACT

This research explores the potential of automated personality classification using supervised machine learning techniques, with a specific focus on the Myers-Briggs Type Indicator (MBTI) for hostel allocation. The problem is the manual and resource-intensive allocation of hostel accommodations in educational institutions, leading to personality mismatches among roommates, which negatively impacts students' academic engagement and well-being. The aim of this research is to improve hostel room allocation by predicting student personality traits, thus minimizing personality mismatches among roommates. A Naïve Bayes classifier was applied to the MBTI personality dataset acquired from Kaggle. Various preprocessing techniques, such as tokenization, stop-word removal, and word stemming, were used to prepare the dataset for feature selection. CountVectorizer and TF-IDF methods were employed to convert text into numerical vectors for machine learning analysis. Additionally, SMOTE was implemented to address class imbalance within the dataset, ensuring better model performance. The Naïve Bayes classifier showed promising results, with scores achieved across all personality traits as 73% precision, 70% accuracy, F1 Score 70% and Recall 70% in predicting MBTI personality traits from input text. Visualizations and performance metrics such as accuracy, precision, recall, and F1-score were used to evaluate the model's efficiency. The research findings suggest that supervised machine learning techniques can effectively predict personality traits from social media data, aiding in improved hostel room allocation and potentially benefiting broader applications in education and computational psychology. This system is recommended for educational institutions and residential complexes seeking efficient roommate pairing.

1. Introduction

Modern computer applications increasingly rely on internal or back-end features that incorporate human psychology for enhanced user experiences. Personality prediction is a significant aspect of personalization in various applications such as e-commerce sites and mobile apps, leveraging user-provided personal information such as demographic attributes (gender, age, and past selections). Computationally analyzing personality and integrating it into systems can improve user-friendliness and help forecast future conduct, especially in educational data mining and computational psychology. Traditional methods of personality analysis, such as written tests, interviews, and surveys, are time-consuming, costly, and often yield inefficient outcomes. The rise of social networking sites such as Facebook and Twitter has led to an increase in online interactions, where individuals share thoughts, feelings, and opinions, reflecting their attitudes and behaviors. This correlation between a person's temperament and their online conduct through comments or tweets has attracted significant attention from various studies. As a result, it is simpler to evaluate a person's actions and draw conclusions about their personality using social media platforms (Khan *et al*., 2020; Katiyar *et al*., 2020).

Different sorts of personality classifications can be determined using a variety of psychological tests. Tests like the MBTI, Big Five, and DISC are popular. One of the most prominent and extensively used personality tests or descriptions is the Myers-Briggs Type Indicator (MBTI) (Agarwal *et al*., 2021). Briggs Myers is one of the key figures in the development of the idea of personality prediction. Carl Jung's thoughts were basically cited and put into practice

* Corresponding author. Tel.: +2348069588787
E-mail address: habilamikailu.hm@gmail.com (Mikailu H.)
https://doi.org/10.62054/ijdm/0103.13
© 2024 Department of Mathematics, Modibbo Adama University.

while adding more insights. Carl's theories served as inspiration for Briggs Myers, who used Sensing and Intuition, Introversion and Extroversion, and Thinking and Feeling as the three original pairs of references in his topology. Briggs Myers introduced a fourth pair after researching the three existing pairs (Katiyar *et al*., 2020). The MTBI then lists the following personality traits: Introversion vs. Extroversion, Sensing vs. Intuition, Thinking vs. Feeling, and Judging vs. Perceiving. This subject of personality prediction did not advance as rapidly as expected during its time, progressing slowly instead. By utilizing a questionnaire that considers various personality traits exhibited by an individual, a person's personality type can be determined. This quiz helps individuals find their suitable match, enabling them to identify the personality type that is most conducive to their success (Katiyar *et al*., 2020).

The problem involves the manual and resource-intensive allocation of hostel accommodations in educational institutions at the beginning of each academic session, often resulting in personality mismatches among roommates, which can lead to conflicts detrimental to students' academic engagement and overall well-being. However, a persistent problem in educational institutions lies in hostel accommodation allocation, which involves manual, resource-intensive processes at the start of each academic session. These processes often lead to personality mismatches among roommates, causing conflicts. Unlike family dynamics, roommates tend to be less tolerant of incompatible behavior, as noted by (Omonijo *et al*., 2015). The quality of roommate relationships, as argued by Erb *et al*. (2014) affects students' academic engagement and overall well-being. Adverse effects, such as reduced sleep disturbances, depression, discomfort, absenteeism can also result from incompatibility in housing (Acquah, 2016).

The aim of this work is to classify the personality traits of students from input text by applying supervised machine learning techniques, namely Naïve Bayes classifier, on the benchmark dataset of MBTI personality. This research work intends to improve compatibility in pairing students in a room during hostel allocation with the understanding of the predicted personality characteristic.

2. Literature Review

Supervised learning algorithms are comprised of unlabeled data/variables which is to be determined from labeled data, also called independent variables. Many researches have been conducted on personality recognition using supervised learning. Most researchers have employed and developed different approaches using supervised learning methodologies to improve technology utilization. Khan *et al*., (2020) proposed a system for analyzing online text and classifying personality accordingly and resolving challenges due to skewness of dataset discovered in prior works. The work mainly emphasized data acquisition, re-sampling pre-processing, feature selection and machine learning algorithm for prediction.

The feature vectors were constructed using different feature selection techniques such as TF/IDF, etc. The obtained feature vectors were used during training and testing of machine learning algorithms, like KNN, Decision Tree, Random Forest, MLP, Logistic Regression (LR), SVM, XGBoost, MNB. However, XGBoost with all feature vectors achieved best accuracy across all dimensions of Myers-Briggs Type Indicator (MBTI) types (Khan *et al*., 2020). According to Fahrudy *et al.* (2022), the changing personality behavior of students gives rise to various problems. Thus, there is a need for an intelligent system to determine students' personality types. Subsequently, the researcher applies such an intelligent system with a classification method using the Naïve Bayes to determine the personality of the student based on Sanguine, Choleric, Melancholic, and Phlegmatic classes against Hippocrates Galenus typology. The attributes used include gender, age, year of class, answers to test A, answers to test B, answers to test C, and answers to test D. System testing was carried out with a scheme for sharing training data and test data. Data used was collected via questionnaire based on the Hippocrates-Galenus typology which was filled out by 130 students. The data collected was divided into training data set and a test data set. The proportion is 90:10 using 10-fold cross validation. The data held are then calculated using the naive Bayes algorithm. Based on the results, 12 students were correctly predicted and 1 student was not predicted correctly. An accuracy of 92.31% was obtained.

Chishti and Sarrafzadeh (2015) used unsupervised machine learning methodology to identify the traits and personalities of network visitors. The website's quantitative materials served as the foundation for this suggested endeavour. The results showed that the technique could also be used to forecast the personality attributes of website and network visitors. In order to study how different personality types interact on Twitter, (Celli & Rossi 2012) also developed an unsupervised personality detection algorithm. The method developed by Celli and Rossi made use of statistical and linguistic variables. It was tested on a dataset that has been annotated with personality models based on human judgment. The study discovered that neurotic users tended to build longer chains of users who participated with

© 2024 Department of Mathematics, Modibbo Adama University.

the posts and posted less frequently than secure users.

Sun *et al*., (2019) aimed to examine group-level personality detection via unsupervised feature learning. The adawalk algorithm was employed for this. According to the results, adawalk performed better in the case of Micro-F1 by at least 7% for Wiki, 3% for Cora, and 8% for BlogCatlog. Furthermore, the suggested method achieved 97.74% Macro-F1 (means macro-average of the F1 scores for each class in a classification task) in the case of the SoCE personality dataset. Alsadhan and Skillicorn (2017) proposed a method for predicting personality from social mediabased writing using word count. It supported 8 distinct languages and the MBTI and Big5 personality models. For both Big5 and BMTI, four different types of labeled corpus were employed for the experiments. 1000 of the most commonly used terms from each corpus are chosen. The Big5 "openness" attribute had the highest prediction accuracy across all corpora, although the MBTI S/N dimension had the highest prediction accuracy compared to other dichotomies.

The goal of Lukito *et al*. (2016) study was to extract MBTI personality qualities from Bahasa Indonesian tweets. Out of 142 responders, 97 users were chosen for the study; the former sent an average of 2500 tweets. WEKA was used to generate the training and classification set, and three different prediction methods—lexicon-based, machine learning based, and rule-driven—were employed in the training set. Of all the approaches, Naive Bayes performed the best, with an accuracy rate of 80% for I/E and 60% for the remaining attributes (S/N, J/P, and T/F). The Myers-Briggs Type Indicator (MBTI) is a survey tool whose results provide insight into a person's personality and reveal various psychological tendencies. People use the results to notice the world around them, analyze it, and then make judgments about it. By considering all statistics, it was determined that it was designed for average individuals, which distinguishes it from a natural occurrence(Varvel *et al*.,2003)."The premise on which MBTI was built in that it could all have some similar categories in behaviour that makes up personalities and some technique to forecast traits as well. These common categories may be the means of providing personal information, strengths, weaknesses, happiness, and devotion.

For any particular instance commonly 4 types of behaviors are described by four letters: These are ESTJ: Extraversion (E), Sensing (S), Thinking (T), Judgment (J), INFP: Introversion (I), Intuition (N), Feeling (F), Perception (P). Different behavioral criteria formerly utilized by the MBTI have a few correlation matrices with the aforementioned range of personality behaviors; these personality traits are typically shared by users on social media (Katiyar *et al.,* 2020). Varvel *et al*. (2003) built McCrae's personality prediction theory on Costa's five theory models. To determine whether a person is suitable for a particular designation, McCrae defines correlation matrix and measures these personality qualities. One may forecast a person's skill and experience using a correlation matrix, according to (Menon & Rahulnath 2016). Research on personality prediction has been conducted in two directions as a result of the increased usage of mobile devices and the sharing of information on social media in recent years: Personality Recognition via smart devices and Personality Recognition on web (Xue *et al*., 2017).

Muhammed *et al*. (2012) proposed a system that could easily manage the hostel details, room details, student records, mess expenditure, mess bill calculation, easy way of room allocation and hostel attendance. Thus, there were a lot of repetition that were easily evaded which has reduced the data redundancy (Ocho & Okeke, 1997).
According to Daniel (1994), several researchers have explored the relationship between residence hall living and satisfaction with the college experience. Researcher has demonstrated that physical environment and social factors can have a substantial impact on student's satisfaction with their residence hall experiences. Residence hall experiences, in turn, have shown to have positive impact on student perceptions of their undergraduate experiences, mend ships and faculty student relations.

3. Methodology

The working procedure of this system involves the following steps: (i) Data acquisition and initial resampling, (ii) Preprocessing and feature selection, (iii) Applying SMOTE (Synthetic Minority Oversampling Technique) for addressing class imbalance, and (iv) Applying the Naïve Bayes algorithm for personality classification.

3.1 Method of data collection

The public available benchmark dataset was acquired from Kaggle. This data set is comprised of various numbers of rows, where every row represents a unique user. Each user's social media posts are included along with that user's MBTI personality type (e.g. ENTP, ISJF). As a result, a labeled data set comprising of records, was obtained

© 2024 Department of Mathematics, Modibbo Adama University.

in the form of excerpt of text along with user's MBTI type.

3.2 Model setup

3.2.1 Data level re-sampling approach:

Data manipulation sampling approaches focus on rescaling the training datasets for balancing all class instances. It takes further, as it hopes to identify any large deviation from assumption. The re-sampling of an imbalanced dataset occurs before the training of the prediction model. In this research, we employed the Synthetic Minority Oversampling Technique (SMOTE), a numerical method utilized for re-sampling imbalanced datasets. SMOTE addresses the imbalance by generating synthetic samples for the minority class, thus mitigating the overfitting issue associated with random sampling methods. Specifically, SMOTE operates within the feature space, facilitating the creation of new instances through interpolation between closely situated positive instances. This approach contributes to a more balanced and representative dataset, enhancing the performance and reliability of our predictive models.

3.2.2 Preprocessing and feature selection

In this research, we employed various preprocessing techniques and feature selection methods to delve deeper into the personality traits exhibited in text data. Our preprocessing procedures encompassed word stemming, stop-word removal, URL elimination, exclusion of user mentions and hash tags, tokenization, and feature selection utilizing TF-IDF (Term Frequency-Inverse Document Frequency). These methods collectively served to refine and enhance the quality of the textual data, enabling us to extract meaningful features essential for personality analysis.

1. **Preprocessing**: The following preprocessing steps on mbti_1 dataset were applied before classification.
   - i. **Tokenization**: Tokenization enable the division of words into small fractions of text. For this reason, Python-based NLTK tokenizer is utilized.
   - ii. **Dropping Stop Word**: Stop words don't reveal any idea or information. A python code is executed to take out these words utilizing a pre-defined words inventory. For instance, "the", "is", "an" and so on are called stop words.
   - iii. **Word stemming**: It is a text normalization technique. Word stemming is used to reduce the inflection in words to their root form. We then produced Stem words by eliminating the pre-fix or suffix used with root words.

2. **Feature Selection**: The following feature selection steps were accomplished using different machine learning classifiers.
   - i. **CountVectorizer**: Using machine learning algorithms, it cannot execute text or document directly, rather it may first be converted into matrix of numbers. This conversion of text document into number vector is called tokens. The count vector is a well-known encoding technique to make word vector for a given document. CountVectorizer takes what's known as the Bag of Words approach. Each message or document is divided into tokens and the number of times every token happens in a message is counted. CountVectorizer perform the following tasks:
     - a. It tokenizes the whole text document.
     - b. It constitutes a dictionary of defined words
     - c. It encodes the new document using known word vocabulary.
   - ii. **Term Frequency Inverse Document Frequency**: In our approach, we leveraged TF-IDF (Term Frequency-Inverse Document Frequency) scoring to assess the significance of words within our dataset. TF-IDF enables us to strike a balance between the importance of common words and less frequently used ones. It achieves this by calculating the frequency of each token within individual tweets (term frequency) and weighing it against the frequency of the token across the entire dataset (inverse document frequency). This process assigns a value to each token, reflecting its importance within the dataset as a whole. Unlike traditional frequency counts, TF-IDF emphasizes the relevance of terms, offering insights into their importance rather than simply their occurrence frequency.

© 2024 Department of Mathematics, Modibbo Adama University.

Naive Bayes Multinomial Algorithm

i. **Data Representation**:
Each instance in the dataset is represented by a feature vector $x=(x_1, x_2, \ldots, x_n)$, where x_i denotes the frequency or count of the i-th feature.

ii. **Model Assumptions**:
Independence Assumption: The model assumes that the presence of a particular feature in a class is independent of the presence of other features, given the class label.
Multinomial Distribution: It assumes that features follow a multinomial distribution, hence the name "multinomial Naive Bayes".

iii. **Parameter Estimation**:
Prior Probability P(Ck): This is the probability of class C_k occurring in the dataset.
Conditional Probability $P(x_i \mid C_k)$: This is the probability of feature x_i occurring given that the class is C_k.

iv. **Training**:
Compute Class Priors: Estimate $P(C_k)$ as the frequency of class C_k in the training set.
Compute Conditional Probabilities: Estimate $P(x_i \mid C_k)$ for each feature x_i and class C_k using Laplace smoothing to handle zero frequencies.

v. **Classification**:
Input: Receive a new instance x_{test}.
Prediction: Calculate the posterior probability $P(C_k \mid x_{test})$ for each class C_k using:
$P(C_k \mid x_{test}) \propto P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)^{xi}$
where x_i is the count of the i-th feature in x_{test}
Assign Class: Assign x_{test} to the class C_k with the highest posterior probability.

Students' classification into matching groups for hostel accommodation:

Features: Each student is represented by a feature vector where each feature could be:
Preferences: Ratings or scores (frequencies) given by the student for various aspects of accommodation (e.g., cleanliness, noise level, proximity to campus, room size).
Personal Details: Gender, age, dietary preferences, etc.
Data Collection: Gather historical data on past student preferences and their assigned accommodation groups.
Feature Extraction: Convert preferences and personal details into numerical features.
Count Vectorization: Use count-based vectorization (multinomial representation) to convert text or categorical data into numerical format suitable for Naive Bayes.
Fit a multinomial Naive Bayes model using the training data. This involves calculating:
Prior Probabilities: Probability of each accommodation group (class).
Conditional Probabilities: Probability of each feature value given the accommodation group.
Classification Phase:
Input: Receive new student data (preferences and details).
Feature Extraction: Convert new student data into the same feature vector format as used during training.
Prediction: Use the trained Naive Bayes model to predict the most probable accommodation group/class for the new student.
Output: Assign each student to a suitable accommodation group based on their profile.

Table 1. Sample Tweets of MBTI Dataset Obtained

Personality Type	Post
ESTP.	Most things hands on. For me, music. I'm very tactile. I like to write too
ENTP	You're fired
INTP	I'm in this position where I have to let go of the person, due to some reasons

© 2024 Department of Mathematics, Modibbo Adama University.

ENTJ	Lol. Its not like our views were unsolicited. What a victim.
INFP	That more or less finds myself in agreement, honey cookie.
ESTP	Most things hands on. For me, music. I'm very tactile. I like to write too.
INFJ	I'm not sure that is a good question
ENFP	He doesn't want to go on the trip without me , so me staying behind wouldn't be an option for him
ISFJ	We are always willing to help those in need

3.2.3 Working procedure of the system for personality traits prediction

As depicted in Fig. 1, first, the proposed model is trained by giving both labeled data (MBTI type) and text (in the form of tweets). After training the model, it is evaluated for efficiency. For better prediction, the dataset will be split into three phases (training phase, validating phase and testing phase). The validating step will reduce over-fitting of data.

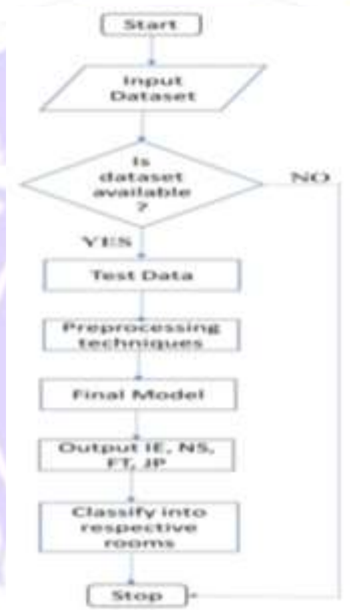

Figure 1: Predicting Phase

Figure 1 depicts a flowchart of the smart hostel allocation system. The system will take in the dataset and from the available dataset test the data. The data pre-processing is executed before the final model and the finally the output with respective room allocation.

3.2.4 Proposed pseudo code for the entire system.

Input: A set of tweets from the MBTI Kaggle dataset saved in CSV format.
Output: Classification of the input text into personality traits: ["I_E", "S-N", "F-T","J-P"]
Machine Learning Classifier: ["Naive Bayes"]
Stop-word List: [There, it, on, into, under, ...]
Start
#Inputting Snippet of Text
Assign the text of the post from the dataset to the variable "Text."
#Pre-processing steps.
#Tokenization/segmentation
Assign the tokenized version of "Text" to the variable "Token."
#Dropping stop words

© 2024 Department of Mathematics, Modibbo Adama University.

Create a new variable "Post_text" by dropping stop words from the "Token."
#Punctuation removal
#Split the dataset into training and test sets and assign them to variables
"X_train," "Y_train," "X_test," and "Y_test" using a 20% test size.
#Counter Vectorizer(Post_Text)
Apply counter vectorization to "Post_Text."
#Application of tf-idf
Implement the "Naive Bayes" machine learning classifier and fit it to the training data.
#SMOTE
Re-sampling techniques use to balance the in-balance dataset for better performance
#Classifier implementation
Assign the trained model to the variable "Model."
#Traits Prediction
Use the "Model" to predict personality traits for the test data and assign the predictions to
"Trait_Prediction."
#Accuracy
Calculate the accuracy of the predictions using "Trait_Prediction" and "Y_test" and store
it in the variable "Accuracy."
#Recall score
Calculate the recall score of the predictions using "Trait_Prediction" and "Y_test" and store
it in the variable "Recall.”
#Precision score
Calculate the precision score of the predictions using "Trait_Prediction" and "Y_test" and
store it in the variable "Precision."
#F1 score
Calculate the F1 score of the predictions using "Trait_Prediction" and "Y_test" and store it
in the variable "F1_score."
Assign the values of "Accuracy," "Recall," "Precision," and "F1_score" to their respective
personality traits: ["I_E", "S-N", "F-T", "J-P"].
Compatibility Rules and Classification
Return the values of the personality traits along with assigned room number.

3.2.5 Applying Naïve Bayes for personality classification

In this research, the Naive Bayes multinomial algorithm served as a pivotal tool for classifying personality traits. Leveraging its probabilistic framework, we trained the algorithm using labeled data to establish patterns and relationships between input features and personality traits. By analyzing the frequency distribution of words or features within the dataset, the Naive Bayes multinomial algorithm effectively categorized individuals into their respective personality types, enabling accurate predictions based on learned probabilities.

© 2024 Department of Mathematics, Modibbo Adama University.

4. Results

In this section, the results of the research are presented in Figure 2 to Figure 8.

4.1 Dataset visualization

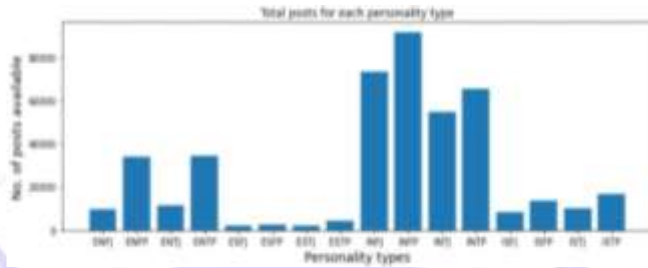

Figure 2: Distribution of posts based on personality traits within the dataset.

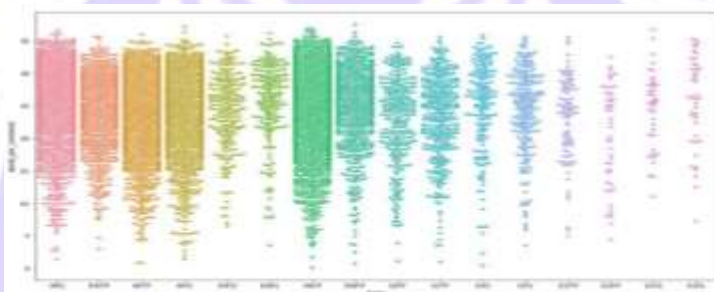

Figure 3: Relationship between posts and their associated personality types through a Swarm Plot.

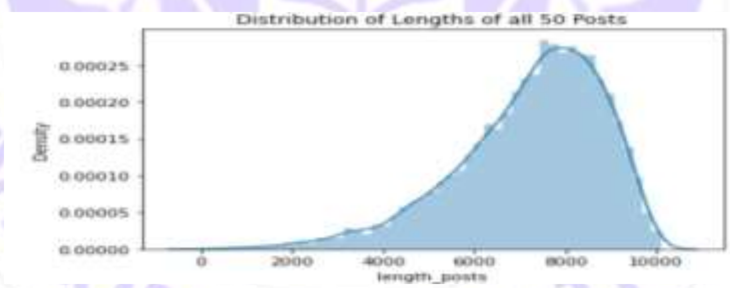

Figure 4: Histogram distribution of data related to post length through a Distance Plot.

© 2024 Department of Mathematics, Modibbo Adama University.

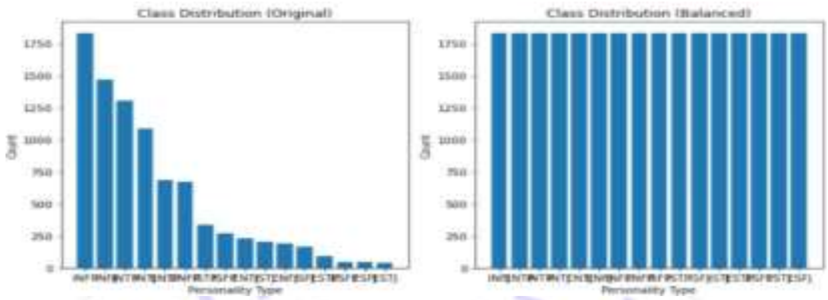

Figure 5 Implementation of Synthetic Minority Oversampling Technique (SMOTE).

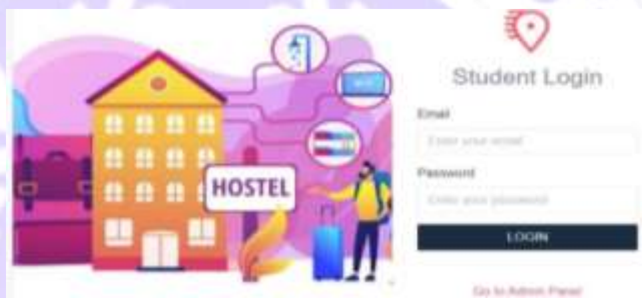

Figure 6: Student Login Page

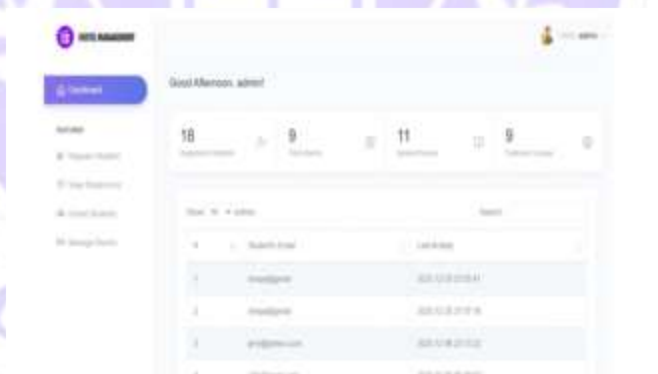

Figure 7: Admin Dashboard Page

© 2024 Department of Mathematics, Modibbo Adama University.

Calculate accuracy, precision , F1-score, and recall

```
In [15]: accuracy = accuracy_score(y_test, y_pred)
         precision = precision_score(y_test, y_pred, average='weighted')
         f1 = f1_score(y_test, y_pred, average='weighted')
         recall = recall_score(y_test, y_pred, average='weighted')

         print(f"Accuracy: {accuracy}")
         print(f"precision: {precision}")
         print(f"F1 Score: {f1}")
         print(f"Recall: {recall}")

         Accuracy: 0.694695548354085
         precision: 0.7244781937587788
         F1 Score: 0.6969274875978884
         Recall: 0.694695548354085
```

Figure 8: Evaluation Metrics

5. Discussion

One striking observation Figure 2 is the prevalence of posts linked to the INFP personality trait, which demonstrates the highest frequency among all traits. Conversely, posts associated with the ESTJ personality trait are notably less frequent, indicating a lower occurrence within the dataset. This distribution provides valuable insights into the prevalence and representation of different personality traits in the text data analysed. Moreover, the distribution of posts across personality traits can have implications for the development and refinement of classification models and this helps to inform a better decision and approach while using the data.

Figure 3 illustrates the relationship between posts and their associated personality types through a Swarm Plot, offers valuable insights into the distribution and characteristics of personality traits within the dataset. The Swarm Plot visually depicts the distribution of posts across different personality types, providing a nuanced understanding of how individuals with specific personality traits express themselves through text. Using the analysis of this Swarm Plot, certain personality types such as INFP, INFJ, INTJ, INTP, ENTP exhibit a higher frequency of posts and specific linguistic features compared to others amid the 16 personality traits. This information is instrumental in identify potential outliers or anomalies within the dataset, which offers valuable insights into individual variations in expression and behaviour.

The distance plot in Figure 4 visually represents the distribution of post lengths, allowing for a nuanced exploration of the range, frequency, and variability of post lengths within the dataset. Using analysis of the Distance Plot, the presence of distinct clusters and peaks corresponding to specific post length ranges was revealed, indicating common posting behaviours or communication patterns among individuals.

The graph in Figure 5 visually depicts the distribution of data before and after applying SMOTE, highlighting the transformation from an unbalanced dataset to a balanced one. The unbalanced data distribution often occurs when one class (in this case, a specific personality trait) is significantly more prevalent than others within the dataset. This imbalance can lead to biased model predictions, as the algorithm may prioritize the majority class and overlook minority classes. Consequently, the model's performance may suffer, particularly in accurately predicting and classifying minority traits. By employing SMOTE, we artificially generate synthetic samples for the minority class, thereby balancing the dataset and reducing skewness. This technique helps to rectify the imbalance in class distribution, ensuring that each personality trait is adequately represented in the training data. As a result, the trained model becomes more robust and capable of making accurate predictions across all personality types.

Different evaluation metrics were used to check the overall efficiency of predictive model as shown in Figure 8. The result of the performance metrics slated below has shown considerably satisfactory accuracy for the system solution carried out.

Accuracy: 0.694695548354085

© 2024 Department of Mathematics, Modibbo Adama University.

precision: 0.7244781937587788
F1 Score: 0.6969274875978884
Recall: 0.694695548354085

The result in Figure 7 depicts the total number of students registered, the available accommodation and the allocated rooms. The system would therefore smartly allocate the incoming students based on their predicted trait to a suitable room containing students of matching trait. The Figure shows that there are 18 students registered for accommodation, where we have total number of 9 rooms. From the 18 students registered for accommodation, 11 of them have been allocated rooms based on their predicted trait and are assigned rooms matching traits. Figure 8 depicts the evaluation metrics such as accuracy, precision, recall and f-measure, which describe the performance of a model.

6. Conclusion

The central theme of this work is the application of machine learning techniques on the benchmark, MBTI personality dataset namely MBTI_kaggle to predict from text different personality traits such as Introversion-Extroversion(I-E), iNtuition-Sensing(N-S), FeelingThinking(F-T) and Judging-Perceiving(J-P) and classify them into compatible group for efficient hostel allocation. The Mayers-Briggs Type Indicator (MBTI) model was used for text classification and personality traits recognition (Myers, 1962). After applying class balancing techniques (SMOTE) on the imbalanced classes, machine learning classifiers namely Naïve Bayes was experimented to identify the personality traits. Evaluation metrics, such as accuracy, precision, recall and Ƒ-score, were used to analyze and examine the overall efficiency of the predictive model. The obtained results show that score achieved by the classifier across all personality traits was good enough. It yielded 73% precision, 70% accuracy, F1 Score 70% and Recall 70%.

7. Recommendations

This system can be recommended for implementation in educational institutions, corporate housing facilities, co-living spaces, and residential complexes seeking efficient roommate pairing based on personality traits. Future work could focus on refining the model and exploring other machine learning algorithms to enhance personality prediction accuracy. The recommendation below can also improve the system.

i. The predictive performance of MBTI personality model needs to be compared with the Big Five Factor (BFF) model for better assessment of the traits.
ii. Multilingual textual content, can be examined for personality classification.
iii. More experiments on personality recognition may be conducted using Deep learning algorithms.

References

Acquah, J. D. G. (2016). Understanding Roommate Conflict among University of Cape Coast Students: A Poisson Regression Approach. *Journal of Social and Development Sciences*, 7(1), 73-81.

Agarwal, K., Ruth, V. M., Chaturvedula, K. & Jongoni, V. C. (2021). Analysis of human traits using machine learning. *IJCRT*, 9(6).

Alsadhan, N. & Skillicorn, D. (2017, November). Estimating personality from social media posts. *IEEE international conference on data mining workshops (ICDMW)* 350-356).

Celli, F. & Rossi, L. (2012, April). The role of emotional stability in Twitter conversations. *In Proceedings of the workshop on semantic analysis in social media,* 10-17.

Chishti, S., Li, X. & Sarrafzadeh, A. (2015). Identify Website Personality by Using Unsupervised Learning Based on

© 2024 Department of Mathematics, Modibbo Adama University.

Quantitative Website Elements. In Neural Information Processing: *22nd International Conference, ICONIP* 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings, Part I 22 (pp. 522-530). Springer International Publishing.

Farah, Agarwal, A., Gupta, A., Mittal, D. & Singhal, K. (2022). Analysis of problems faced by people living in shared accommodation. *International Research Journal of Engineering and Technology*, (IRJET), 9(5).

Katiyar, S., Kumar, S. & Walia, H. (2020). Personality prediction from stack overflow by using Naïve Bayes theorem in data mining. *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* 9(3).

Khan, A. S., Hussain, A., Asghar, M. Z., Saddozai, F. K., Arif, A. & Khalid, H. A. (2020). Personality classification from online text using machine learning approach. *International journal of advanced computer science and applications,* 11(3).

Lukito, L. C., Erwin, A., Purnama, J., & Danoekoesoemo, W. (2016). Social media user personality classification using computational linguistic. *8th international conference on information technology and electrical engineering (ICITEE)*, (pp. 1-6). IEEE.

Menon, V. M. & Rahulnath, H. A. (2016, September). A novel approach to evaluate and rank candidates in a recruitment process by estimating emotional intelligence through social media data. *International Conference on Next Generation Intelligent Systems (ICNGIS*) 1-6. IEEE.

Mohammad, M. I., Gambo, Y. L. & Omirin, M. M. (2012). Assessing facilities management service in postgraduate hostel using servqual technique. *Journal of Emerging Trends in Economics and Management Sciences,* 3(5), 485-490.

Ocho, L. & Okeke, B. S. (1997). Dynamics of Educational Administration and Management: The Nigerian Perspective. Meks Publishers.

Omonijo, D. O., Anyaegbunam, M. C., Nnedum, O. A. U., Chine, B. C. & Rotimi, O. A. (2015). Effects of college roommate relationships on student development at a private university, Southern Nigeria. *Mediterranean Journal of Social Sciences,* 6(6 S1), 506.

Sun, X., Liu, B., Meng, Q., Cao, J., Luo, J. & Yin, H. (2020). Group-level personality detection based on text generated networks. *World Wide Web*, 23, 1887-1906.

Varvel, T., Adams, S. G. & Pridie, S. J. (2003). A study of the effect of the Myers-Briggs Type Indicator on team effectiveness. In Proc. *ASEE Annual Conference and Exposition, Session*, 17(06).

Xue, D., Hong, Z., Guo, S., Gao, L., Wu, L., Zheng, J. & Zhao, N. (2017). Personality recognition on social media with label distribution learning. *IEEE access*, 5, 13478-13488.

© 2024 Department of Mathematics, Modibbo Adama University.