Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.2202.18

2022, 41(2) 185-196

A rule-based machine learning model for career selection through MBTI personality

Noureen Fatima *, Sana Gul, Javed Ahmed, Zahid Hussain Khand, Ghulam Mujtaba

Center of Excellence for Robotics, Artificial Intelligence & Blockchain, Department of Computer Science, Sukkur IBA University Pakistan

* Corresponding author: Noureen Fatima, Email: noureen.mscss19@iba-suk.edu.pk

Received: 03 November 2020, Accepted: 05 May 2021, Published: 01 April 2022

K E Y W O R D S

Machine Learning

NLP

Personality Type

MBTI Trait

SMOTE

A B S T R A C T

Career selection is one of the most important decisions in everyone's life. Being a student it's quite difficult to find the right career as the world is moving so fast and the competition level is too high so it is tough to choose the right job for anyone. According to the Council of Scientific and Industrial Research (CSIR) survey report, 40% of students are a bit confused about their career options. The productivity of human resources may reduce if someone chooses the wrong career. Therefore, we need an intelligent system that chooses your career based on a person's personality type. This study creates a personality profile of the person and suggests their best careers list according to their personality type after analysing the text written by the user such as a blog post, essay, or tweet. The main purpose of this project is to work on the text classification method, pre-processing of that dataset and convert it into main features and then train the model on a best-performed classification model. After applying various feature vector combinations and machine learning models that are detailed in this work, accuracies up to 93% were achieved.

1. Introduction

Career selection is one of the major decisions which people make for a better living. There are very few lucky people who choose the right career which satisfies them. In most cases, people end up choosing a career that is poles apart from their personality, generally because of the job security and monetary benefits. This career selection may provide them happiness at the start but may lead to less productivity, work stress, manual error, and shifting of job [1]. Therefore, choosing a career that is suitable for a candidate will solve the above-mentioned problem. There are several models available that identify the person's personality based on the available text data. The Myers–Briggs Type Indicator (MBTI) is considered the most reliable and popular method that assists the counsellors to choose the best career for the candidate [2].

MBTI is a theory that aims to highlight the difference between individual perceptions of the world using four classes. These four classes further create the combination of 16 personality types that are detailed in section 1.2. The four MBTI classes are as follows.

1. Introversion vs. Extraversion: This class shows whether a person is talkative or reserved. Introvert people are reserved while extrovert people are talkative.

2. Sensing vs. Intuition: This class shows the person's perception of the information. Someone who is good at sensing lives in today and enjoys the facts while the intuitive person tries to find the deeper meaning of the information.

3. Thinking vs. Feeling: This class shows the ability of the person for decision-making. The person can either be subjective or objective in making the decision.

4. Judging vs. Perceiving: This class reflects the person's attitude towards the world. People with judging preferences want everything perfect, neat, and ordered. While people with perceiving preference want things to be flexible and spontaneous [3].

Personality plays a very important role in life, especially in recruitment, team selection, educational field selection, and career selection. The productivity of the work can be improved by choosing the best person for the particular work [4]. As discussed above the person's personality can be identified by classifying the perception of the world through MBTI classes. There are 16 personality types shown in figure 1 that result from the interactions among the preferences of an individual. As a result, MBTI is used to indicate a person's personality type in real life for different purposes such as career selection and recruitment.

Natural language processing (NLP) is defined as the computerized approach, based on different models and theories to understand and analyse the human language [5]. There are various levels used in NLP to create emotional and semantic representations of the text, which are then conceded into the training modules. These levels are; (1) phonology, (2) morphology, (3) lexical, (4) syntactic, (5) semantic, (6) discourse, and (7) pragmatic [6].

The interpretation of speech sounds within and across words is dealt with by phonology level. Identification and analysis of the structure of the words are dealt with by morphology level. Identification of the word's position in a sentence, their meaning, and their relation to other words in that sentence are identified by lexical level. The syntactic level is used to analyse the words of a sentence to discover the grammatical structure of the sentence. Semantic analysis level is used to find interactions among word-level meanings in a sentence and the way that lexical meaning is combined morphologically and syntactically to form the meaning of the statement. The connections between sentences in a text and the properties of the whole statement in conveying meaning are dealt with by discourse level. Finally, the use of language in context, deriving the purposeful use of the language in different situations is determined by pragmatic level [5].

After analysing these levels, it was concluded that NLP should be used for automating the career prediction based on the personality of the person using the MBTI personality indicator. Researchers have proved the correlation between a person's personality and language. For Example, extroverts use more social and positive emotional words as compared to introverts [7]. Similarly, the people who use articles ('a', 'an' and the) are highly intellectual, assured, and open to experience [1]. Advances in machine learning have created several models using closed and open vocabulary language approaches to generate information from the language content available on blogs, microblogs (Twitter, Sina Weibo) [1], essays, and social media posts [8]. However, these proposed methods do not provide the comparative analysis of different text classification models and the accuracy of the models is less than the proposed method of this study. Furthermore, the existing work is contributing to prediction of the personality only.

Thus, to overcome the limitations of discussed studies, this study proposed an accurate rule-based model for career selection by detecting your MBTI Personality. This proposed model will predict the personality type of the person and suggests their best careers list according to their personality type after analysing the text written by the user such as a blog post, essay, or tweet. To achieve good accuracy and good features for models five supervised machine learning algorithms i.e. Decision Tree (DT), k-nearest neighbours (KNN), logistic regression (LR), Random Forest (RF), and Support vector machine(SVM) have been used. For measuring the performance of each text classifier model, Macro accuracy, Macro Precision, Macro Recall, and Macro F-measure are used. The major contribution is given below.

1. The proposed method will analyse the given social media posts or microblogs then it will create the personality profile of the person using the MBTI scale and finally suggest the best career according to the MBTI personality type.

2. We tested different Machine learning classifiers (LR, RF, SVM, KNN, and DT) and feature selection algorithms (Chi-squired and PCA) to maximize the performance of our proposed model by giving the most significant features to the proposed model. The SMOTE (The synthetic minority over-sampling technique) is the

method through which an unbalanced problem can be resolved by making the classes balanced. Thereafter, we compare the performance of each Machine learning classifier with basic features, with SMOTE with feature engineering and without SMOTE.

3. This study achieved an average accuracy of 93% by applying Natural Language Processing (NLP) techniques, feature engineering, and a machine learning classifier. Which is state of the art for predicting the career based on the analysis of the text. Surpassing previously reported highest average accuracy of 82%.

The rest of the article is organized as follows. In Section 2 we have discussed the related work/literature review for predicting the personality by analysing the text and their performance. The detail of the proposed model is given in section 3. In Section 4 the detailed discussion and results are specified. Significance along with the conclusion is given in section 5.

2. Related Work

In the research area of Natural language processing and social science, there is significant growth for automating the personality type prediction based on user data available online especially the data of social media. This data is significantly important for analysing the personalities, behaviours, and preferences of people [9].

Most of the studies have focused on the Big five model and MBTI, which are the two widely used models used for personality prediction. The big five models are mainly concerned with predicting big five traits i.e. extroversion, agreeableness, conscientiousness, neuroticism, and openness [10]. However, MBTI uses four classes discussed in section 1.1 for predicting personality types. According to the researchers, considering the reliability and validity MBTI has more applications in real life as compared to the Big five models. In the world's 500 super enterprises, such as IBM, Southwest Airlines, Disney, Pepsi, above 80% of the senior personnel managers are using MBTI [11].

Several studies deal with predicting career and personality based on social media data such as: In [12], the authors presented the personality prediction Model for career guidance by using Artificial Neural Networking. For data collection, the author used a web-based questionnaire designed using Google forms which comprised 36 questions as described by MBTI categorization. Evaluation matrices like sensitivity, specificity, precision, and accuracy for all the MTBI categorizations were evaluated by the model which showed values above 92% in all the cases

In [9], authors presented personality trait prediction for Facebook users using four machine learning models was investigated to examine the social network structures and linguistic features under personality interactions by use of personality project dataset. Results indicated that the XGBoost classifier outperformed the other three classifiers with 74.2% prediction accuracy.

Another model was designed to predict the personality trait based on MBTI data. They had analysed which features are predictive of which personality traits, using MBTI personality type and gender. The model for all the MBTI categorizations and accuracy showed values were following; I/E = 72.5%, S/I = 77.5%, T/F = 61.2%, J/P = 55.4% [13].

Personality can be understood as specific features of an individual which determine its preferences over things. The regression model was used to predict the personality types based on the MBTI dataset of Kaggle User's personality. 5-fold cross-validation was used to evaluate the classifiers and the overall accuracy of the model was 67% [3].

In [14], the model has been trained through different machine learning algorithms. TFIDF, and TSVD used for converting text into vector foam on the MBTI Kaggle dataset. The model for all the MBTI categorizations and accuracy values were following; I/E = 82.1%, S/I = 82.2%, T/F = 84.2%, J/P = 79.6%.

Exploring Twitter for open vocabulary personality prediction to analyse and compare three statistical models and find its correlation with personality traits and linguistic behaviour was presented. Naïve Bayes classifier outperformed the other statistical model with the highest accuracy of 80% for the I/E category and the rest for 60% for user classification [15].

The abovementioned work has achieved immense success in personality prediction however, no one yet has worked on carrier selection through MBTI personality and to improve the performance of existing methods using the imbalanced MBTI dataset and our major contribution is to suggest careers after predicting personality type and increase the performance of the model. Our model achieved above 97% performance in most of the cases which is greater than the above-mentioned studies. In addition, we applied the SMOTE method to deal with unbalanced class issues.

3. Methodology

This section gives an overview of data collection and classification techniques used for developing a predictive model for career selection.

3.1 Data Collection

The dataset consists of tweets with one labelled personality type of 16 MBTI types as shown in Fig. 2. These labelled tags are combinations of four characters and every character represents the first letter of the MBTI class. For example, if the personality type is ENFJ, it represents that this person has extroverted, intuitive, feeling, and judging personality traits. The dataset consists of 8675 rows.

Table 1

Detailed statistics of the dataset

Total	Per average
50	Number of tweets per user averagely
1311	Number of words per user averagely
6567.25	Number of characters per user averagely
4	Word length across the tweets per user

The distribution of MBTI traits in each class is given in Figure 2, Figure 3, Figure 4, and Figure 5. As you can see the data was quite unbalanced, so we divide the dataset into four classes and each class contains two features as shown in Table 1

3.2 Pre-Processing of Data

To explore the personality traits from Twitter tweets' text by individuals, a lot of pre-processing techniques were used through NLP. Following are NLP techniques that we have applied to remove the useless features. We have removed the URL, Links, stop words. After removing the useless word, we used the porter stemming technique and tokenizes it for enhancing the classification performance.

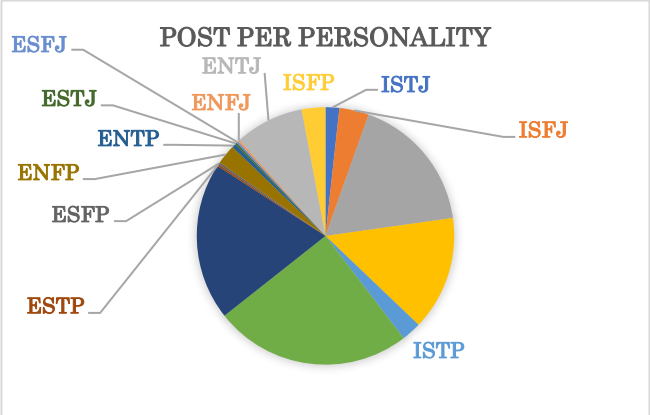

Fig. 1. Distribution of personality type per post

After the division of the dataset into four classes due to unbalanced data. We get one partially balanced class and three unbalanced classes. As shown in Fig. 2.

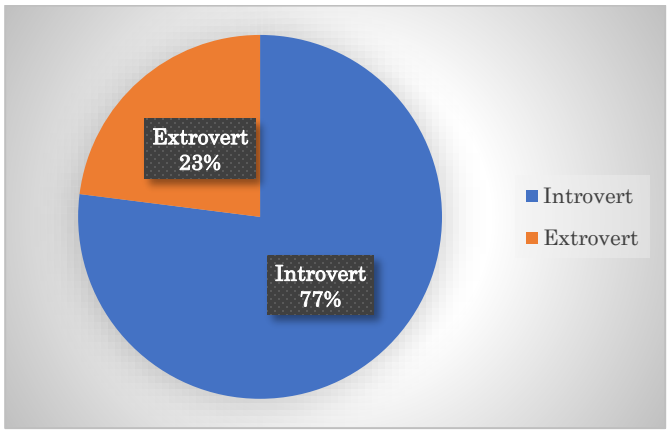

Fig. 2. Distribution of class A in a dataset

3.3 Feature Engineering

Feature engineering is the process of finding the best and most useful features that enhanced performance. Feature engineering is a fundamental task in text classification [16, 17-22] because the text contains a lot of features like words symbols and many others so is essential to find out which features of the text contribute more towards the improvement of the performance. Feature engineering consists of four steps; feature extraction, feature representation, feature selection, and feature reduction.

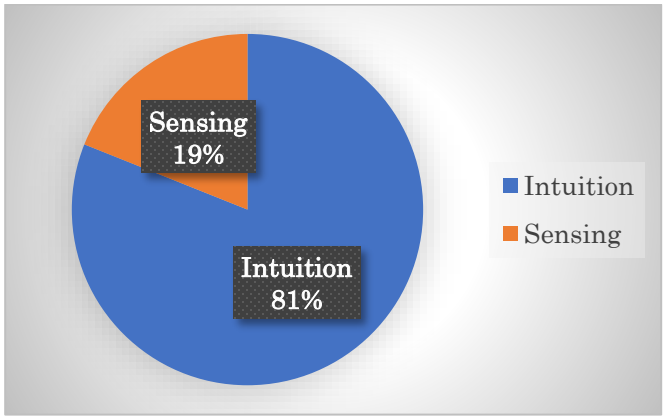

Fig. 3. Distribution of Class B in the dataset

3.3.1 Feature extraction

In feature extraction, we extracted the useful features from Kaggle MBTI dataset posts columns. We have used a fully automated feature extraction approach where no Human Interaction of expertise is needed. The content-based features were extracted from the MBTI dataset. The features include (bag of words) and n-gram techniques. Unique words were extracted through BOW techniques, each word represented as an independent

and distinctive feature. We have used a unigram technique that is a set of co-occurrences of the word within the given MBTI dataset.

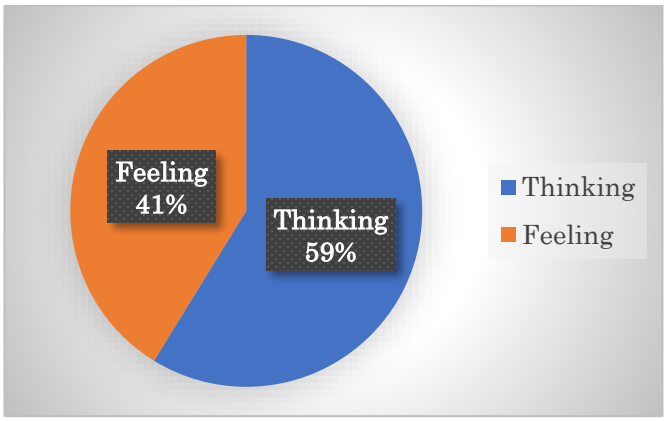

Fig. 4. Distribution of Class C in a dataset

3.3.2 Feature representation

Feature representation is converting the text values into numeric values for each feature extracted to learn the classification rule [23]. TFiDF (Term frequency-inverse document frequency) is used for the representation of the Personality type's features. TF (Term frequency) will show the occurrence of every single word in a particular document. IDF (Inverse of document Frequency) will show the occurrence of documents contained in a particular word. Machines are similar to logistic regression but when data is not linearly separable it is very useful. SVM works best for text mining or text classification. For predicting the MBTI trait, we have taken the advantage of the SVM classifier and applied it in our model.

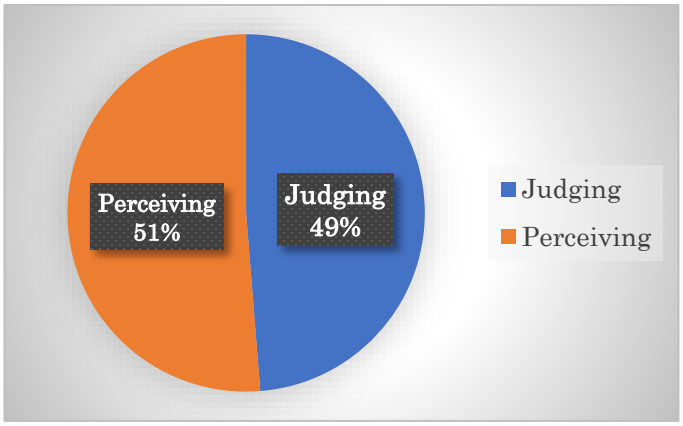

Fig. 5. Distribution of Class D in a dataset

3.3.3 Feature selection

Feature selection is the process of selecting the most relevant features in the entire feature list [23] and which features contribute the most accuracy. For personality traits, we have applied Chi-square and for the best 2000 features were extracted through select K-best library from Sklearn. In every class of Personality traits, we have 2000 features. The chi-square test determines whether there is a notable difference between the observed frequencies of the word and their expected frequencies [24-25]

3.3.4 Feature reduction

Feature reduction is the process of orthogonal transformation to transform a set of observations of correlated features into principal components. For that, we have applied for PCA (Principal Component Analysis). In general, the count of principal components is less than or equal to the original number of observations [22]. To improve the overall accuracy, we employed 0.9 PCA that reduces MBTI trait 2000 Features for each class into the following: 840 features for T/F class, 226 features for I /S class, 615 features for JP class, and in the last 364 features for I/E class.

3.4 Text Classification Techniques

There are many text classification techniques, but we have chosen the topmost practiced classification models in text classification as given in [25]. Support vector machine (SVM), K-nearest Neighbour (KNN), decision tree (DT), random forest (RF), logistic regression (LR), and stochastic gradient descent (SGD).

3.4.1 Support vector machine (SVM)

Support vector machines are a very popular and useful supervised machine learning classifier which is based on statistical learning theories [26]. It has proven accurate results and is extensively used in biomedical documents, image classification [27], and text classification [28-30]. SVMs are hyperplanes that separate the training examples by maximal margin [31]. SVM separates the classes by using a line/ hyper-plane. Support vectors were selected to opt for the best accuracy, the parameters for SVM are: one is linear SVM and second C values should be 1.0.

3.4.2 K-nearest Neighbour (KNN)

KNN is labelled as a lazy learning classifier because KNN memorizes the training dataset rather than learning from discriminative function from the training dataset. It is instance-based learning. New instances are classified by using Similarity measures, such as Euclidean distance or Jacquard Similarity by KNN [29, 32] shown in Eq. 1.

$$\sqrt{\sum_{i=1}^{k}(xi - yi)^2} \quad (1)$$

3.4.3 Decision tree (DT)

For the task of classification and prediction Decision Tree is the most commonly used classifier [33]. Decision tree structure looks like a flowchart in that each internal node represents a test on an element, each branch is the resultant of a test, each leaf node represents a class label, easily understood by humans, and entropy is used by DT classifier to compute the homogeneity among each class of Personality trait. Pruned DT or unpruned DT is created by a Decision tree classifier [34].

3.4.4 Random forest (RF)

Random is an ensemble supervised machine learning classifier that creates multiple decision trees using randomly selected features from training data into one forest [35]. Collectively a decision has been taken to improve the accuracy rather than relying on a single learning model. For the personality trait, we have created 100 random decision trees to opt for the best accuracy.

3.4.5 Logistic regression (LR)

LR is a statistical method that is used when we have a categorical dependent variable. Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

3.5 Rule-Based Classification Scheme

The rule-based classification scheme was designed by the usage of IF-Then rules for class prediction. Here for career prediction, we have defined 16 rules for each personality type according to the given list by MBTI career selection [36].

3.5.1 Evaluation metrics

For the performance evaluation of all five classification models, precision, recall, F-measure, and overall accuracy were calculated and compared. The reason behind the selection of these metrics was an imbalanced distribution of the classes because these metrics confirm equal weights for all the MBTI classes. These evaluation metrics will be discussed in the succeeding paragraphs.

3.5.1.1 Macro precision (precision): Precision (ranges from 0 to 1, higher is better) is the ratio of correctly predicted positive observations to the total predicted positive observations. Whereby, Precision_M is the average of each class's precision. Eq. 2 shows the mathematical representation of Precision_M.

$$\text{Precision}_M = \frac{\sum_{i=1}^{C} \frac{TPi}{TPi+FPi}}{C} \quad (2)$$

3.5.1.2 *Macro recall (*$Recall_m$*)*: Recall also termed as sensitivity (ranges from 0 to 1, higher is better) is the ratio of correctly predicted positive observations to all observations in the actual class. Whereby, Recall_M is the average of each class recall. Eq. 3 shows the mathematical representation of Recall_M.

$$\text{Precision}_M = \frac{\sum_{i=1}^{C} \frac{TPi}{TPi+FPi}}{C} \quad (3)$$

3.5.1.3 Macro F-measure ($F\text{-}Measure_M$*)*: F-measure is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Eq. 4 shows the mathematical representation of F-measure_M.

$$F - measure_M = \frac{(\beta^2+1)Recall_M \times \text{Precision}_M}{\beta^2\,(Recall_M + \text{Precision}_M)} \quad (4)$$

3.5.1.4 Overall accuracy: The ratio of classification results predicted correctly among all the classes is called overall accuracy. Eq. 5 shows the mathematical representation of accuracy.

$$Accuracy_{AVG} = \frac{\sum_{i=1}^{C} \frac{TPi+TNi}{TPi+FNi+TNi+FPi}}{C} \quad (5)$$

3.6 Handling with an Unbalanced Dataset

Imbalanced class problems are found in many classification problems [37]. Class imbalance problems occur when the number of instances from one or more classes is considerably greater than another class [38]. MBTI dataset shows an imbalanced classes distribution, instances of each class are: Intuition 373900, Sensing 59850, Judging 171700, Perceiving 262050, extrovert 99950, and introvert 333800 however, Thinking and Feeling class is balanced one. Such imbalanced class distribution can prevent the model from accurately classifying the instances in such cases, standard classifiers tend to be overwhelmed by the large classes and ignore the small ones. It usually produces a biased classifier that has higher predictive accuracy over majority classes, but poor predictive accuracy over minority classes as machine learning algorithms work best when there is an approximately equal number of instances in each class.

Several approaches for overcoming these problems have been proposed, such as a combination of

oversampling the minority (abnormal) class and under-sampling the majority (normal) class [39], as well as weight adjusting, approaches [40].

To solve imbalanced class problems, SMOTE (Synthetic Minority Oversampling technique) method is used in which class distribution is being modified in training by oversampling the minority class or under-sampling the majority class [39].

3.7 Experiments

We have created four modules; these operate separately on the eight classes of MBTI. Each module class contains two classes that are opposite either a person could be an introvert or extrovert. We have applied NLP for cleaning the data and extracting the useful features from Twitter data. TFiID is used to convert the text into feature vectors [41]. By utilizing chi-square and PCA for feature selection and reduction methods are employed. Once the features are extracted, five different text classification techniques (SVM, LR, KNN, DT, and RF) were applied for training and testing with and without SMOTE technique. 10-fold cross-validation was used for performing experiments on all the models. All experiments were based on a 10-fold cross-validation [42-43].

4. Results

This section contains all the details of the proposed model and the results of five text classifiers (SVM, LR, RF, KNN, and DT) along with the analysis of each classifier in terms of accuracy Macro: Precision, Recall, and F-measure and AUC. To overcome class imbalance problems, SMOTE method has been implied. As result, we have compared the two approaches with and without SMOTE method for each of the classifiers.

Table 2

Logistic regression result with basic feature

Class	Precision	Recall	Score	AUC	Accuracy
Extrovert	0.89	0.92	0.69	0.89	0.88
Introvert	0.98	0.92	0.93		
Judging	0.76	0.91	0.82	0.87	0.87
Perceiving	0.95	0.85	0.9		
Intuition	1	0.9	0.94	0.9	0.9
Sensing	0.31	0.91	0.46		
Thinking	0.93	0.91	0.92	0.926	0.91
Feeling	0.89	0.91	0.9		

4.1 Results Obtained by Using Basic Classifiers

All eight classifiers were run using all proposed features based on 10-fold cross-validation. **Error! Reference source not found.**, **Error! Reference source not found.** and **Error! Reference source not found.** show the result of each classifier. SVM perfumed well under the basic features with f-measure varying in between 0 to 0.92.

Table 3

Random Forest result with basic feature

Class	Precision	Recall	FScore	AUC	Accuracy
Extrovert	0	0	0	0.82	0.77
Introvert	1	0.77	0.87		
Judging	0.76	0.91	0.82	0.8	0.61
Perceiving	0.95	0.85	0.9		
Intuition	1	0.86	0.93	0.845	0.86
Sensing	0	0	0		
Thinking	0.93	0.91	0.92	0.92	0.91
Feeling	0.89	0.91	0.9		

Table 4

K-Nearest neighbour result with basic feature

Class	Precision	Recall	FScore	AUC	Accuracy
Extrovert	0.54	0.75	0.63	0.85	0.85
Introvert	0.95	0.87	0.91		
Judging	0.68	0.78	0.73	0.8	0.61
Perceiving	0.87	0.81	0.84		
Intuition	0.98	0.91	0.94	0.73	0.9
Sensing	0.43	0.75	0.55		
Thinking	0.9	0.77	0.83	0.707	0.8
Feeling	0.69	0.85	0.76		

4.2 Results Obtained by Classifiers with Feature Selection without SMOTE

We ran all five classifiers with feature selection to determine the most significant feature that may improve the performance of the classifier and reduce classification time. Feature selection algorithms, namely, the c2 test was tested in the experiment. And Feature reduction PCA method was tested in the experiment. **Error! Reference source not found.**, Table 6

Random Forest result with feature selection

, **Error! Reference source not found.**, **Error! Reference source not found.** and **Error! Reference source not found.** compare the five classifiers with each feature selection method without SMOTE. Compared with **Error! Reference source not found.**, **Error!**

Reference source not found. and **Error! Reference source not found.**, overall accuracy was increased and improved the F1-score and AUC. In summary, using the feature selection and feature detection techniques have only slightly improved the AUC and accuracy results compared with using all features to train the model **Error! Reference source not found.**, **Error! Reference source not found.** and **Error! Reference source not found.**.

Table 5

Logistic regression result with feature selection

Class	Precision	Recall	FScore	AUC	Accuracy
Extrovert	0.3	0.87	0.44	0.89	0.83
Introvert	0.99	0.82	0.9		
Judging	0.76	0.91	0.82	0.87	0.87
Perceiving	0.95	0.85	0.9		
Intuition	1	0.87	0.93	0.8	0.8
Sensing	0.08	0.75	0.15		
Thinking	0.89	0.89	0.89	0.8	0.88
Feeling	0.87	0.87	0.87		

Table 6

Random Forest result with feature selection

Class	Precision	Recall	FScore	AUC	Accuracy
Extrovert	0	0	0	0.82	0.77
Introvert	1	0.77	0.87		
Judging	0.93	0.91	0.92	0.92	0.91
Perceiving	0.89	0.91	0.9		
Intuition	1	0.86	0.93	0.845	0.86
Sensing	0	0	0		
Thinking	0.01	1	0.02	0.8	0.61
Feeling	1	0.6	0.75		

Table 7

K-Nearest neighbour result with feature selection

Class	Precision	Recall	FScore	AUC	Accuracy
Extrovert	0.18	0.9	0.29	0.73	0.8
Introvert	0.99	0.8	0.89		
Judging	0.32	0.92	0.48	0.758	0.72
Perceiving	0.98	0.69	0.81		
Intuition	0.99	0.91	0.95	0.73	0.8
Sensing	0.4	0.87	0.55		
Thinking	0.97	0.65	0.78	0.73	0.71
Feeling	0.4	0.91	0.55		

Table 8

Decision tree result with feature selection

Class	Precision	Recall	FScore	AUC	Accuracy
Extrovert	0.57	0.78	0.66	0.78	0..86
Introvert	0.95	0.88	0.91		
Judging	0.67	0.8	0.73	0.78	0.8
Perceiving	0.89	0.8	0.84		
Intuition	1	0.93	0.96	0.78	0.9
Sensing	0.51	0.95	0.66		
Thinking	0.9	0.82	0.86	0.87	0.83
Feeling	0.76	0.87	0.81		

Table 9

SVM with feature selection

Class	Precision	Recall	FScore	AUC	Accuracy
Extrovert	0.42	0.87	0.56	0.86	0.85
Introvert	0.98	0.85	0.91		
Judging	0.9	0.9	0.9	0.9	0.89
Perceiving	0.89	0.88	0.89		
Intuition	0.99	0.88	0.93	.93	0.87
Sensing	0.14	0.82	0.24		
Thinking	0.65	0.87	0.75	0.78	0.82
Feeling	0.94	0.8	0.86		

4.3 Results Obtained by the Classifier with Imbalanced Data Distribution

We applied over-sampled the minority class and majority class SMOTE to handle the unbalanced dataset distribution. **Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.**, and **Error! Reference source not found.** compare the five classifiers with each feature selection method SMOTE. **Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.**, and **Error! Reference source not found.** shows the result of each text classifier significantly improve the overall performance in term of precision, recall, F1-score, AUC, and Accuracy as compared to **Error! Reference source not found.**, **Error! Reference source not found.**, Table 6

Random Forest result with feature selection

, **Error! Reference source not found.**, **Error! Reference source not found.** and **Error! Reference source not found.**. Before using SMOTE some of the classes show biased data in **Error! Reference source not found.** Precision, recall, and F1-score of the extrovert class was 0 and the introvert class was 1 but after applying the SMOTE in table 10 shows the unbiased data. The precision, recall, and F1-score of the extrovert class were

between 0.69 and 0.93 and the introvert class was between 0.92 to 0.93.

Table 10

Logistic regression result with SMOTE

Class	Precision	Recall	FScore	AUC	Accuracy
Extrovert	0.89	0.92	0.69	0.91	0.88
Introvert	0.98	0.92	0.93		
Judging	0.76	0.91	0.82	0.87	0.87
Perceiving	0.95	0.85	0.9		
Intuition	0.85	0.88	0.86	0.9	0.9
Sensing	0.88	0.85	0.87		
Thinking	0.93	0.91	0.92	0.926	0.91
Feeling	0.89	0.91	0.9		

Error! Reference source not found., **Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.**, and **Error! Reference source not found.** show that the overall classifier performance was obtained using SMOTE, SVM, and Logistic Regression outperformed with average 91 AUC, F measure 0.92, Precision 0.89, and Accuracy 0.91. That implies that SMOTE performs well on this project.

Table 11

Random forest result with SMOTE

Class	Precision	Recall	FScore	AUC	Accuracy
Extrovert	0.88	0.87	0.88	0.92	0.88
Introvert	0.87	0.88	0.87		
Judging	0.88	0.84	0.86	0.92	0.91
Perceiving	0.83	0.87	0.85		
Intuition	0.84	0.88	0.86	0.92	0.87
Sensing	0.89	0.85	0.87		
Thinking	0.92	0.87	0.89	0.93	0.89
Feeling	0.86	0.91	0.89		

Table 12

K-Nearest neighbour result with feature selection

Class	Precision	Recall	FScore	AUC	Accuracy
Extrovert	0.54	0.75	0.63	0.853	0.9
Introvert	0.95	0.87	0.91		
Judging	0.9	0.77	0.83	0.758	0.8
Perceiving	0.69	0.85	0.76		
Intuition	0.89	1	0.94	0.96	0.94
Sensing	1	0.9	0.95		
Thinking	0.95	0.87	0.91	0.73	0.8
Feeling	0.43	0.75	0.55		

Table 13

Decision tree result with SMOTE

Class	Precision	Recall	FScore	AUC	Accuracy
Extrovert	0.87	0.75	0.81	0.83	0.79
Introvert	0.72	0.85	0.78		
Judging	0.79	0.8	0.8	0.82	0.8
Perceiving	0.8	0.79	0.8		
Intuition	0.88	0.86	0.87	0.89	0.87
Sensing	0.86	0.88	0.87		
Thinking	0.85	0.82	0.84	0.87	0.83
Feeling	0.82	0.84	0.83		

Table 14

SVM result with SMOTE

Class	Precision	Recall	FScore	AUC	Accuracy
Extrovert	0.88	0.87	0.88	92.6	0.88
Introvert	0.87	0.88	0.87		
Judging	0.88	0.84	0.86	90.6	0.85
Perceiving	0.83	0.87	0.85		
Intuition	0.84	0.88	0.86	92.8	0.87
Sensing	0.89	0.85	0.87		
Thinking	0.92	0.87	0.89	93.5	0.89
Feeling	0.86	0.91	0.89		

5. Discussion

The goal of the proposed approach was to predict the personality type and suggest a career based on MBTI traits. A comparative study of various classifiers to obtain the most accurate prediction of the type of personality and suggested the best fit career based on MBTI traits.

The accuracy of the classification mostly depends on the quality of the feature set. The irrelevant, incomplete and extraneous features generate less comprehensive and accurate results. Therefore, it is important to exclude non-discriminative irrelevant features from master features by using feature subset selection algorithms before applying classification algorithms [44]. The basic reason for using feature subset selection algorithms is to choose the best features for the classification and to remove the features which are irrelevant or not contributing to the performance. To evaluate this proposition, we aim to determine the best feature subset size for the classification of the personality of a person for improving the classification performance. After selecting the chi-squared with the best 2000 features were further decreased up to the point where no further improvement in performance was established. Furthermore, we also assessed the performance of all

five classifiers discussed above using 'different' features.

Agreeing to the "no free lunch" theorem [45], there is no single machine learning algorithm that performs best for all the application areas. Therefore, a variety of machine learning algorithms should be tested. Therefore, we evaluated the performance of five classifiers (SVM, LR, KNN, DT and RF). In most classes, Random Forest performs well due to its assembler nature. The choice of the kernel is a performance indicator of SVM [1]. The selection of the best kernel function parameters such as width or sigma parameter may advance the SVM performance [46].

In the experimental results of this research, logistic regression can identify the outliers that can reduce misclassification [23]. Furthermore, Logistic regression determines the relative impact on one or more predictor variables to the criterion value. Although the SVM classification model is suitable for both linear and nonlinear data. SVM is a versatile algorithm that provides the correct results by creating a novel kernel for decision function. In addition, SVM performs effectively on higher-dimensional nonlinear data. Finally, it is memory efficient and uses a subset of training points (support vectors) as decisive factors for classification [42].

DT classification algorithm may be a weak predictive performance due to it produces a weak or noisy classifier. It mostly grows large and needs pruning, which may cause a loss of information [23]. In addition, it is not suitable for multi-classification because it's used for binary classification. Thus, can show less performance than any other multiclass classifiers. Another reason for the low performance of DT classifiers is that it does not work well in data with class imbalanced. Finally, DT suffers because of its generic nature, a minute change may suffer the training set.

The proposed model, A Rule-based Model for Career Selection by detecting your MBTI Personality model achieved improved and reliable performance on different text classification classifiers. The experimental results on the publicly available dataset showed that it is the most reliable, and accurate model. We compared three different approaches with five text classifiers with different evolution matrices: simple baseline feature, feature selection with classifiers, and feature selection with SMOTE. The first baseline feature achieved the best result for LR but biased result. Second feature selection improved the overall performance in terms of accuracy and AUC but biased. In third, feature selection with SMOTE performed best among the previous two approaches and improved the performance of all classifiers.

6. Conclusion

The main purpose of this work was to build a career profile after predicting personality based on the tweets and text written by the user. The preliminary analysis focused on searching patterns in sentiments and analysing the distribution of emotion on the data. The text was pre-processed to remove punctuations, numbers, hyperlinks, and context-sensitive words. To represent features into dimensional vectors, TF-IDF was used. The experimental results on the publicly available dataset showed that it is the most reliable, and accurate model. We compared three different approaches with five text classifiers with different evolution matrices: simple baseline feature, feature selection with classifiers, and feature selection with SMOTE. The first baseline feature achieved the best.

7. References

[1] C. J. Burges, "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery, vol. 2, pp. 121-167, 1998.

[2] J. Passmore, M. Holloway and M. Rawle-Cope, "Using MBTI type to exploredifferences and the implications for practice for therapists and coaches: Are executive coaches really like counsellors", Counselling Psychology Quarterly, vol. 23, pp. 1-16, 2010.

[3] R. S. S. Chaudhary, S. T. Hasan and M. I. Kaur, "A Comparative Study of Different Classifiers for Myers-Brigg Personality", International Research Journal of Engineering and Technology, vol. 5, p. 21, 2018.

[4] L. R. B. Kollipara, G. Ghosh and N. Kasturi, "Selecting Project Team Members through MBTI Method: An Investigation with Homophily and Behavioural Analysis", Second International Conference on Advanced Computational and Communication Paradigms, Gangtok, India, 2019.

[5] H. B. K. M. H. Amirhosseini, K. Ouazzane and C. Chandler, "Natural Language Processing approach to NLP Meta model automation", International Joint Conference on Neural

Networks, Rio de Janeiro, 2018.

[6] E. D. Liddy, "Natural Language Processing. Encyclopedia of Library and Information Science", 2nd Edition, Marcel Decker Inc., 2021.

[7] J. W. Pennebaker and L. A. King, "Linguistic Styles: Language Use as an Individual Difference", Journal of Personality and Social Psychology, vol. 77, p. 1296, 1999.

[8] Z. H. D. Xue, S. Guo, L. Gao, L. Wu, J. Zheng and N. Zhao, "Personality recognition on social media with label distribution learning", IEEE Access, vol. 6, pp. 61959-61969, 2017.

[9] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Personality Predictions Based on User Behavior on the Facebook Social Media Platform", IEEE Access, vol. 6, pp. 61959-61969, 2018.

[10] A. Al Marouf, M. K. Hasan, and H. Mahmud, "Comparative Analysis of Feature Selection Algorithms for Computational Personality Prediction From Social Media", IEEE Transactions on Computational Social Systems, vol. 7, pp. 587-599, 2020.

[11] Z. Z. a. B. C. L. Wu, "The application of MBTI personality type theory in the Bank management", 6th IEEE Joint International Information Technology and Artificial Intelligence Conference, Chongqing, 2011.

[12] A. S. Rao, B. S. Kamath, R. Ramya, S. Chowdhury, A. Shreya, and R. K. K. Pattan, "Use of Artificial Neural Network in Developing a Personality Prediction Model for Career Guidance: A Boon for Career Counselors", International Journal of Control and Automation, vol. 13, pp. 391 - 400, 2020.

[13] B. Plank, & D. Hovy, " Personality traits on twitter—or—how to get 1,500 personality tests in a week", Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, September 2015, pp. 92-98.

[14] "Predicting your MBTI type using text data", Available: https://yix90.github.io/blog/2n 23 Feb 2018].

[15] L. C. Lukito, A. Erwin, J. Purnama, and W. Danoekoesoemo, "Social media user personality classification using computational linguistics", 8th International Conference on Information Technology and Electrical Engineering, 2016.

[16] C. C. Aggarwal, & C. Zhai, "Mining text data. Springer Science & Business Media", Springer, 2012, pp. 1-10.

[17] P. Domingos, "A few useful things to know about machine learning. Communications of the ACM", Communications of the ACM, vol. 55, pp. 78-87, 2012.

[18] J. Heer, Hellerstein, J. M., & S. Kandel, "Predictive Interaction for Data Transformation. In CIDR", in CIDR, 2015.

[19] J. Jiang, "Information extraction from text in Mining text data", Springer, 2012, pp. 11-41.

[20] A. C. Tantug, "Document categorization with modified statistical language models for agglutinati", vol. 3, pp. 632-645, 2010.

[21] I. H. Witten, Frank, E., Hall, M. A., & C. J. Pal, "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, 2016.

[22] S. Wold, Esbensen, K., & P. Geladi, " Principal component analysis: Chemometrics and Intelligent Laboratory Systems", in IEEE Conference on Emerging Technologies & Factory Automation Efta, 1987, pp. 704-706.

[23] F. Debole, & F. Sebastiani, "Supervised term weighting for automated text categorization. In Text mining and its applications", Springer, 2004 pp. 81-97.

[24] F. E. Witten IH, "An introduction to support vector machines and other kernel-based learning methods', Cambridge University Press, 2000.

[25] G. Mujtaba, L. Shuib, N. Idris, W. L. Hoo, R. G. Raj, K. Khowaja, et al., "Clinical text classification research trends: Systematic literature review and open issues", Expert 018/02/23/Predicting-Your-MBTI, [Accessed o Systems with Applications, vol. 116, pp. 495-520, 2019.

[26] A. M. Andrew, "An introduction to support vector machines and other kernel-based learning methods", Cambridge university press, 2000.

[27] W. L. Zhang Y, "Classification of fruits using computer vision and a multiclass support vector machine", Sensors, vol. 12, pp. 12489–12505,

2012.

[28] L. J. Spasić I, Keane JA, Nenadić G "Text mining of cancer-related information: review of current status and future directions", International Journal of Medical Informatics, vol. 83, pp. 956–965, 2014.

[29] K. Fukunaga, "Introduction to statistical pattern recognition", Academic press, 2013.

[30] Y. Zhang, Z. Dong, A. Liu, S. Wang, G. Ji, Z. Zhang, et al., "Magnetic resonance brain image classification via stationary wavelet transform and generalized eigenvalue proximal support vector machine", Journal of Medical Imaging and Health Informatics, vol. 5, pp. 1395–1403, 2015.

[31] M. R. Yeow WL, RG Raj, "An application of case-based reasoning with machine learning for forensic autopsy", Expert Systems with Applications, vol. 41, pp. 3497–3505, 2012.

[32] I. N. Bao Y, X Du, "Combining multiple k-nearest neighbor classifiers using different distance functions", Intelligent Data Engineering and Automated Learning–Ideal, Springer, 2004, pp. 634–641.

[33] Y. Zhang, S. Wang, P. Phillips, and G. Ji, "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection", Knowledge-Based Systems, vol. 64, pp. 22–31, 2014.

[34] Z. Zhao Y, "Comparison of decision tree methods for finding active objects", Advances in Space Research, vol. 41, pp. 1955–1959, 2008.

[35] W. M. Liaw A, "Classification and regression by randomForest", R News, vol. 2, no. 3, pp. 18–22, 2002.

[36] N. A. Schaubhut and R. C. Thompson, "MBTI type tables for occupations", CPP, 2008.

[37] M. Wasikowski and X.-w. Chen, "Combating the small sample class imbalance problem using feature selection", IEEE Transactions on Knowledge and Data Engineering, vol. 22, pp. 1388-1400, 2009.

[38] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, pp. 463-484, 2011.

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", Journal of artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[40] X.-Y. Liu and Z.-H. Zhou, "The influence of class imbalance on cost-sensitive learning: An empirical study", Sixth International Conference on Data Mining, 2006, pp. 970-974.

[41] A. I. Kadhim, Y.-N. Cheah, and N. H. Ahamed, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering", 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, Kota Kinabalu, 2014.

[42] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques", Emerging artificial Intelligence Aplications in Computer Engineering, vol. 160, pp. 3-24, 2007.

[43] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation", Encyclopedia of Database Systems, vol. 5, pp. 532-538, 2009.

[44] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning", Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1998.

[45] M. W. Wolpert DH., "No free lunch theorems for search", Technical Report, SFI-TR-95-02-010, Santa Fe Institute, 1995.

[46] R. A. Siddiqui MF, J. Kanesan, "An Automated and Intelligent Medical Decision Support System for Brain MRI Scans Classification", PLoS One, vol. 10, no. 8, pp. 1-16, 2015.