Personality Recognition for Candidate Screening

Piyush Patil, Saloni Goyal, Tanya Dwivedi and Suvarna Bhat

Computer Engineering Department, Vidyalankar Institute of Technology, Mumbai, India.

Corresponding author(s). E-mail(s): piyush.patil@vit.edu.in; saloni.goyal@vit.edu.in; tanya.dwivedi@vit.edu.in; suvarna.bhat@vit.edu.in;

Abstract

In this paper, we perform Personality Recognition for the development of a platform for analysis and examination of emotions and behavior of job candidates through personality traits recognition. Personality traits can be considered an important factor for working in a professional environment. Evaluation of such traits at a preliminary stage can prove to be beneficial in a working medium. We have decided to explore textual inputs for developing an ensemble model that gathers the information from text responses and integrates them with a provided standard and displays a clear and understandable way of assessing candidates' interest and enthusiasm. Hence, this research suggested an empirical technique to compare Machine Learning models such as Support Vector Machine, Na¨ıve Bayes, Decision Tree, Random Forest, Logistic Regression, Recurrent Neural Networks to discover the optimum personality recognition performance along with Natural Language Processing (NLP), and Affective Computing Methods and find better accuracy of classification algorithms than previous studies by merging Big Five dataset and MBTI dataset. Five human personality traits that are Extraversion (EXT), Neuroticism (NEU), Agreeableness (AGR), Conscientiousness (CON), and Openness (OPN) operated for problem analysis. The outcome revealed that Support Vector Machine and Logistic Regression performed better overall other models across all metrics with an average accuracy score 78.01% for EXT, 79.63% for AGR, 58.91% for NEU, 74.21% for CON, and 81.56% for OPN traits. However, the Naive Bayes algorithm resulted in overall lower performance.

Keywords: Personality traits recognition, Machine learning classification models, Natural language processing, Big five traits, Candidate screening

1 Introduction

Emotion recognition through text input is a demanding assignment that surpasses conventional sentiment analysis. Besides detecting basic responses such as neutral, positive, or negative, the objective is to pin down a set of emotions characterized by a higher gradient.Many subtleties are factored in to perform an accurate detection of human emotions where context-dependency is of prime importance.

There are different methods of tackling natural language processing problems, but they are majorly classified as rule-based and learning-based techniques. Rule-Based approaches target more on pattern-identification and are largely based on grammar analysis and sentence structure, whereas learning-based approaches prioritize probabilistic modelling and likelihood maximization. This study prominently focuses on learning-based methodologies. In this research, we have chosen text mining to streamline unstructured data and to find out personality traits which depends on the "Big Five personality" model and MBTI dataset.

Emotion identification and human personality classification are two distinct disciplines of study, each with its own theoretical underpinnings. However, they have quite same learning dependent methods. The main aim is to provide a broader assessment of the user's emotions, since it can get through the learning of a characteristics of the person, and personality traits analysis would provide a new point to understanding human emotion.

Psychology researchers mainly believe that there are five categories, or core factors, also called BIG 5 that determine one's personality. To mention this model, the acronym OCEAN (for openness, conscientiousness, Extraversion, agreeableness, and neuroticism) is generally used. Due to its popularity and clarity, we have chosen to use this precise model. Openness is a measure that describes an emotion, intellectual curiosity, willingness to try new things, general awareness, reaction to surroundings, etc. Conscientiousness is a scale that elaborates on how people control and regulate their senses and behavior. Extraversion indicates how people function and flourish in their general surroundings and react around others. Agreeableness reflects individual decisions in general concern for social harmony. Neuroticism measures negative emotions, mental stability, anger, anxiety, etc.

Judgment of personality is key when it comes to the prediction of characteristics of an individual which differentiate them from other individuals. Hence, getting a deeper knowledge of a person's personality types is extremely valuable for indicating their physical and mental well being. Myers-Briggs Type Indicator (MBTI) is an indicator of an individual's perception of the world and decision-making process. It classifies an individual's personality based on Sensing(S) or Intuition(N), Thinking(T) or Feeling(F), Extraversion(E) or Introversion(I) and Judging(J) or Perceiving(P). Information from these personality tests helps companies better comprehend the extent of their employees' strengths, weaknesses, and their capability to perceive and process information. As mentioned in Bajic [2015], Data obtained from MBTI

assessments helps businesses build stronger organizations as it guides them to assemble efficient teams, facilitate communication between the team and the manager, motivate employees, solve conflicts and develop leadership.

The main goal is to build a tool capable of recognizing the personality traits of a candidate, given a text containing his answers to pre-established personal questions with the support of statistical learning methods. Getting a general sense of a candidate's educational background, career history, and interest in the company should help you weed out unqualified candidates. Primary checks include looking at qualifications such as work experience, academic background, skills, knowledge base, traits and behaviors that indicate a person's behavior, as well as competency. With traditional recruiting methods, recruiters struggle to evaluate candidates accurately. The purpose of our model is to determine if the candidate has the basic skills, aptitude and enthusiasm to meet the requirements of the job. In candidate selection, various tools are used to assess a candidate's suitability for the job, including interviews, skills tests, psychometric tests, group discussions, and reference checks. It should be noted that the purpose of our model is to narrow down the most qualified candidates who should be treated to a traditional interview and remaining steps of candidate selection.

2 Related Works

Rahman et al. [2019], presents a great method to determine the way of detection of person's characteristics by making comparisons between several activation functions such as sigmoid, and leaky ReLU. Python 2.7 is used for coding of the model, along with Google's word2vec embeddings and Mairesse features. The experimental analysis uses big five personality traits that are EXT, NEU, AGR, CON, and OPN. Calculating the median F1-score of the functions sigmoid, tanh, and leaky ReLU gives 33.11%, 47.25%, and 49.07% respectively.

Pramodh and Vijayalata [2016], datasets stream-of-consciousness and MyPersonality are used. MyPersonality dataset has a mixture of 250 users who are updating around 10,000 Facebook Status Updates. Natural Language Toolkit has been used for their model and their F1-scores are given as 0.665, 0.632, 0.625, 0.624 and 0.637 for the traits OPN, CON, EXT, AGR and NEU respectively.

In Tareaf et al. [2019], proposes a model that can properly distinguish between a religious individual and a non-believer across 83% of circumstances, between individuals of Asian and European decent in 87% of situations, and between emotionally stable and emotionally unstable individual across 81% of circumstances. The presented analysis is a MyPersonality dataset containing 738,000 users who have granted their Facebook activities, data from other social networks, egocentric networks, demographic characteristics.

In Anatoli de Bradk e and ephane Reynal [2018], explored different art models including text, audio, and video in multimodal emotion recognition.

Under text input, the dataset was a study by Pennebaker and King consisting of 2,468 essays. Bag-of-words and Word2Vec embedding techniques were used for preprocessing. Personality scores were assessed by the Big Five Inventory. Different Classification models were tested against each other such as Multinomial Naive Bayes, Support Vector Machines, Recurrent Neural Networks, and LSTM.

In Majumder et al. [2017], deep Convolutional Neural Network or DCNN were used by the researchers to classify between personality traits on the basis of the Big Five model. The dataset used included the stream-of-consciousness essay dataset by James Pennebaker and Laura King. The model also utilizes Google's word2vec embeddings along with Mairesse features.

In Abyaa et al. [2018], supervised learning algorithms have been used for the classification of personality according the Big Five model. The algorithms used were SVM, RF, kNN (k-Nearest Neighbors), NB (Na¨ıve Bayes), J48, LR (logistic regression), and bagging. The dataset used was collected from 48 students. Data included is basically of educational data, survey responses, EMA, and sensor data. Positives results of 62.5%, 50%, 62.5% were achieved by Extraversion SVM, Random Forest, and logistic regression respectively. However only na¨ıve Bayes and bagging give great results: 57.14% for Openness. SVM and bagging gave uplifting outcome of 50% for agreeableness. In case of neuroticism, only SVM gave acceptable results of 57.14%.

In Tinwala and Rauniyar [2021], proposed a model for personality detection using deep convolutional neural networks. The dataset used is essays collated by James Pennebaker and Laura King which are based on the Big Five Model a.k.a the five-factor model or the OCEAN model. The document-level feature extraction was carried out using Google's word2vec embeddings and Mairesse features. The data processed is then fed to a deep convolutional network (DCN) and a binary classifier is used for classifying the availabily or non-availability of the personality trait. Function of tanh is best used for traits Extraversion, Neuroticism, and Agreeableness giving F1- scores of 61.2%, 66.33% and 62.67% respectively. Sigmoid is best used for Openness and Conscientiousness providing F1- scores of 69.71% and 67.46% respectively.

In Kumar and Gavrilova [2019], study proposes a personality traits classification system, which can incorporate the language-based features, that are formulated upon count-based vectorization and technique of Global Vectors word embedding, with a collection of predictive system that consists of decision trees and an Support Vector Machine classifier. The given mixture of results helps to reliably find out the personality traits by the most recent tweets from the given profile. The proposed system's performance gets validated on a giant, publicly available Twitter MBTI Personality Dataset and is compared favorably with other different state-of-the-art techniques.

3 Dataset

The dataset that we have chosen is from Kaggle [2017] that is based on the Myers Briggs Type Indicator Myers [1962] and the Neo Personality Inventory Costa Jr and McCrae [2008] also called the Big 5. The Myers Briggs Type Indicator (or MBTI for short) is a dataset of personality into 16 various types with 4 major divisions that is Introversion (I), Extraversion (E), Intuition (N), Sensing (S), Thinking (T), Feeling (F), Judging (J), Perceiving (P). Depending on these attributes' personalities can be coded in a four-letter term for example - ISTJ, ISTP, ESFJ. This dataset have more than 8500 rows of data, where each row is a MBTI type (4 letter MBTI code/type) of the person and a written entry of each of the last 50 things they have posted. The second dataset used is the MyPersonality Project dataset Kosinski et al. [2015] which consists of 2400 stream-of-consciousness texts labeled with a personality from Pennebaker and King [1999] and used by Mairesse et al. [2007] that examines five personality traits that are openness, conscientiousness, extraversion, agreeableness, and neuroticism. Major traits of both the personality indicators exhibit a correlation among them explained in Furnham [1996] and hence can be used together on the dataset for increased accuracy. Judging-Perceiving dimensions exhibit similarity with Conscientiousness; Thinking-Feeling dimensions for the MBTI is equivalent to Agreeableness; Introversion-Extraversion is correlates with Extraversion and Openness correlates with the Sensing Intuitive MBTI type. Only Neuroticism does not correlates with any of the type in MBTI and somewhat inconsistent across all of them. The dimension remarkably missing from the MBTI is Neuroticism.

4 Methodology

The following is the proposed system's operation procedure: (i) Merging dataset and re-sampling of the imbalanced date (ii) Pre-processing and Text-vectorization (iii) Personality Classification based on text data using Machine learning models (iv) Comparing the efficiency of the models with other classifiers (v) Various evaluation metrics

The publicly available dataset of Big five personality traits is acquired from the website. The dataset is of 2467 rows, in which each row is shown as individual user. Each user's essays are included along with that user's Big 5 personality type (e.g. openness, agreeableness). We merged this Big 5 dataset with the MBTI dataset to increase the efficiency of the Machine learning model. MBTI dataset is also publically available on Kaggle. This dataset consists of 8675 rows, where the user's past social media posts have been given with the MBTI personality type (e.g. ENTP, ISJF). We merged the data based on the correlation derived from the research paper of Adrian Furnham 1996 Furnham [1996]. As a result, a labeled dataset comprising a total of 11142 records.

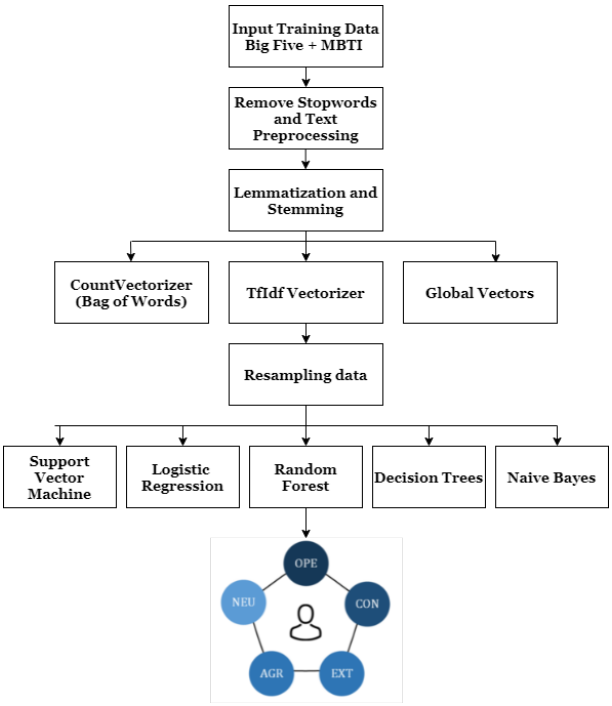


Fig. 1: Proposed text based personality recognition framework

4.1 Text Preprocessing

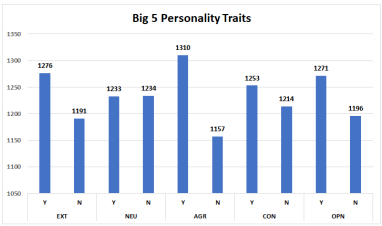
1. Removal of Stop words -

Stop words, such as articles, prepositions, pronouns, conjunctions, and others, are the most common words in any language and do not provide any significant information to the text. "the", "a", "an", "so", "what" are some of the stopwords in English. Elimination of these stopwords is the first step in text preprocessing. Stop words are present in abundance and hence can hamper the results while training the dataset. By keeping unwanted words out of our corpus we can focus more on the high-level information essential for training our dataset.

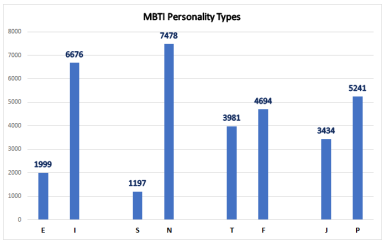
2. Handling imbalanced dataset -

As shown in Figure 1 and Figure 2, the dataset is unevenly distributed in all 5 types, mentioned as follows: S/N Trait: S=7466 and N=1194, T/F Trait: T=4685 and F=3975, I/E Trait: I=6664 and E=1996, J/P Trait: J=5231 and P=3249. The outcome of any algorithm applied on a skewed or unbalanced classified dataset always favors the large and smaller classes that are passed over for prediction. So, we used resampling methods such as Oversampling and Undersampling which basically make the dataset equally distributed.

A. Undersampling -



(a) Occurrence of Big 5 Personality Traits



(b) Occurrence of MBTI Dataset Traits

Fig. 2: Imbalanced dataset

The minority classes can present information that is pertinent to the outcome but the biased distribution of the classes can lead to ineffective results. Undersampling is the removal of random samples from the majority class. Under-sampling minimises the size of the data, requiring less time for learning. The drawback is that deleting majority classes may result in the majority class losing useful information.

B. Oversampling -

Another strategy used to add minority cases to the dataset in order to achieve a balance is over-sampling, which involves duplicating the current minority samples. This increases the data size but provides impetus to the minority classes.

4.2 Text Vectorization

1. CountVectorizer -

It's a technique for converting a given text to the vector or matrix. It counts the count of word and then convert it in vector format. It considers how many times a word appears in a text (multiplicity), ignoring grammatical subtleties and even word order. Countvectorizer creates the matrix which contain word with associated row for all the documents. The value of each cell is the number of words in that particular text sample. The frequency of a particular word in a text is proportional to the importance of the term in the text.

2. Term frequency–inverse document frequency (TFIDF) -

TFIDF technique is of two sections that are term frequency and inverse of document frequency. The term frequency is measured as a percentage of the total number of words in a single document. Term frequency does not consider the importance of words only the frequency. Some words can be most frequently present but are of little significance and hence can alter the results. Each word is given a weight based on its frequency in a corpus using inverse document frequency. This measure is generated by dividing the total number of documents divided by the number of documents which have that specific word.

4.3 Modelling

4.3.1 Support Vector Machine

SVMs are popular linear classifiers that essentially splits the data plotted on a multidimensional graph with a hyperplane. Ideally, the features we use in this classifier must correlate linearly with all the classes in a way that the classifier linearly segments the space to include all records with the same trait labels. For each label, we train a single-label One Vs All classifier that employs the subset of features that are most linearly related to the classes they are assigned to. Finding the right hyper-parameter can be difficult, but it can be done by experimenting with various combinations and observing which ones work best. The method involves the creation of a grid of hyperparameters and trying all of its various combinations, and hence is called Grid search. As mentioned in GeeksforGeeks [2019], GridSearchCV is a built-in feature that uses a dictionary to specify the variables that can be used to train a model. The parameter grid is type of dictionary, having keys as parameters and settings are the values.

4.3.2 Decision Tree

The decision tree classifier builds the classification model by making trees for taking decisions. Every node of the tree is defined as a test and every branch from that node is value as given in KDnuggets. By learning simple decision rules inferred from past data, the purpose of employing a Decision Tree is to develop a training model that can be used to predict the class or value of the target variable (training data). When utilising Decision Trees to forecast a record's class label, we start at the top of the tree. The record's attribute and the root attribute's values are compared. We jump to the next node depending on the comparison by the branch that to that value.

4.3.3 Naïve Bayes

Na¨ıve Bayes works fairly well for text based classification problems. The classifier makes predictions based on the learning of distribution of posterior probability. We train the Na¨ıve Bayes model with a subset of features listed in the previous section. It turns out that the Na¨ıve Bayes classifier yields the best result for one of the personality traits. This algorithm has a simple and intuitive design, and is a good benchmark for classification purposes.

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)}$$

4.3.4 Logistic Regression

Logistic regression is one of the classification method that has its roots in statistics but is now widely used in machine learning. It's a technique for calculating a data collection where one or more variables influence the outcome. It is used to create the best-fitting model and characterise the relationship of

the dependent variables and independent variables. In logistic regression Raj [2020], the Sigmoid function is used in logistic regression to show probability values. In converts number to the range of 0 and 1.

4.3.5 Random Forest

Random forest is a decision tree-based model, it predicts the result by averaging several results from multiple pre-built decision trees. Each decision tree has different structures learned from the training input. In our project, multiple features are in use, but we have do not know how to weigh the features to get the best results for the classification. So we experimented with the Random Forest model with the training set containing the features we choose to use and use the rest of the labeled data for validation.

5 Results and Discussion

We processed with the Big five dataset for personality traits classification for candidate screening. We initially used the CountVectorizer (Bag of Words) method for the vectorization the words on the five personality traits dataset and then trained using Support vector machines, Logistic Regression, Naive Bayes algorithm, Random Forest, and Decision Trees models. In addition, we applied hyperparameter tuning for support vector machines such as GridsearchCV to get better accuracy. We got the better accuracy of NEU traits by hyperparameter tuning. The comparison of the result is given below:

	SVM	DT	NB	LogR	RF	SVM(Hyperparameter Tuning)
cEXT	55.46	51.81	50.41	55.66	53.84	55.26
cNEU	51.61	54.45	51.82	53.23	56.27	58.91
cAGR	52.83	50.41	50.61	53.41	54.04	54.04
cCON	50.81	51.41	53.03	53.03	56.27	53.44
cOPN	57.08	52.42	53.03	56.03	59.11	58.90

Table 1: Data Essays:Vectorizer-Bag-of-words

Then, we used the TfIdf Vectorizer that takes not only the frequency but also the importance of words into account when analyzing a corpus. With the TfIdf vectorizer, we observed that accuracy increased for all the classification traits by around an increase of 2%. The results with Recurrent Neural networks are also not up to the mark.

Nevertheless, based on the Big 5 dataset, we could not achieve high accuracy. Following the research of personality traits in humans and getting a sense of the concept from the paper of Furnham [1996], we merged both five personality traits dataset and the MBTI dataset. By the help of CountVectorizer and TfIdf vectorizer, we convert words into vectors. Then, we trained the data on different machine learning models. By merging the MBTI dataset with the Big five personality, we achieved an increase of accuracy by around 20%.

	SVM	DT	NB	LogR	RF	SVM(Hyperparameter Tuning)
cEXT	53.84	55.66	49.79	55.06	53.64	55.66
cNEU	58.09	49.59	52.83	57.89	56.47	55.87
cAGR	56.07	51.61	50.61	55.87	49.79	56.07
cCON	54.64	51.01	52.22	55.66	52.83	53.44
cOPN	60.72	53.23	53.03	60.52	62.14	60.92

Table 2: Data Essays:Vectorizer-TfIdf

	SVM	DT	NB	LogR	RF	SVM(Hyperparameter Tuning)
cEXT	72.31	72.45	69.71	75.91	71.42	78.86
cNEU	N/A	N/A	N/A	N/A	N/A	N/A
cAGR	70.39	68.28	54.82	72.11	71.01	75.81
cCON	68.05	64.61	57.37	69.62	59.62	72.22
cOPN	77.47	76.08	75.63	79.97	78.66	81.15

Table 3: Data Essays+MBTI:Vectorizer-CountVectorizer

	SVM	DT	NB	LogR	RF	SVM(Hyperparameter Tuning)
cEXT	**78.01**	70.48	68.56	77.65	72.05	77.47
cNEU	N/A	N/A	N/A	N/A	N/A	N/A
cAGR	78.06	66.62	55.65	**79.63**	70.74	76.62
cCON	73.17	65.09	60.25	**74.21**	61.73	71.46
cOPN	**82.18**	76.08	74.18	81.56	79.47	81.29

Table 4: Data-Essays + MBTI :Vectorizer- TfIdf

Now, goal is to estimate likely performance of a model on out-of-sample data, so we applied kFold cross validation on the whole data to reduce the chances of overfitting of the data or high variance. We can get a more correct results of out-of-sample accuracy and effective data with kfold cross validation (every observation is given for both training and testing).With the cross-validation, we get results as minimum 55.03% and maximum cross validation score as 84.29% for the Extraversion trait with 10 folds. The average cross validation score is 77.08% for EXT, 81.47% for OPN, 78.80% for AGR, 74.22% for CON that we achieved with the kfold cross validation on the complete dataset. SVM and Logistic Regression are the most accurate and exact algorithms for our proposed research. Using all criteria, it produced great results for all attributes. SVM and Logistic Regression obtained maximum accuracy (82.18%) for cOPN trait. It has the best performance in all four aspects and all metrics. There are only four MBTI traits and five traits in the Big five personality traits dataset. So, we could not achieve better accuracy NEU trait. We achieved 58.91% accuracy for NEU traits based on the Big five personality traits dataset but better accuracy for the remaining four traits with the help of the MBTI dataset.

The comparison of present work with the previous research work Rahman et al. [2019] and Majumder et al. [2017] of personality traits classification

(Extraversion, Conciousness, Agreeableness, Neuroticism, Openness) is given below:

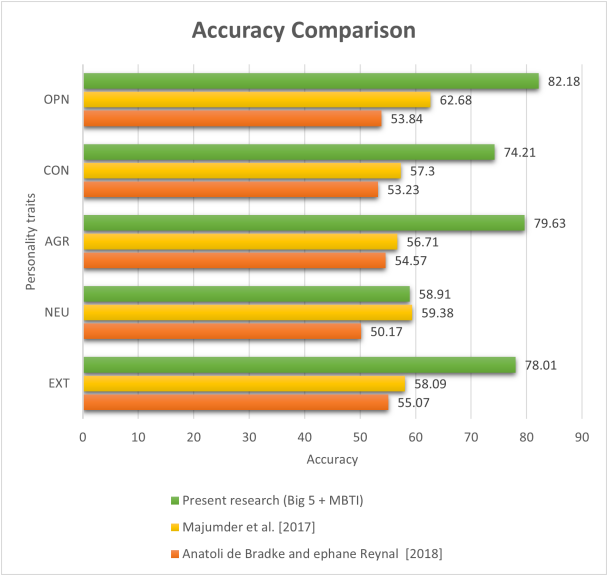


Fig. 3: Comparison of accuracy with present work

6 Conclusion

Using several machine learning models, this study provided a method for determining the optimum personality characteristic identification methodology. We are creating model which can help recruiters to analyse the emotions of candidates as candidate screening method which will eventually save a lot of time of recruiter and will get alternative to traditional interview process. This research will surely help interviewers and companies in the online interview process where company could not able analyse the emotions and personality traits of interviewee. This method consists of data filtering, data preprocessing, data merging, re-sampling of the data, and classification modeling. First, we merged the Big 5 dataset and the MBTI dataset. Following that, Countvectorizer and TfIdf vectorizer was used to vectorize the processed data and then trained with Support Vector Machine, Logistic Regression, Random Forest, Decision Trees, and Naive Bayes models. For comparative experimental analysis, five personality traits were used: EXT, NEU, AGR, CON, and OPN. Support Vector Machine and Logistic Regression outperformed other machine learning models, according to the comparative results. We observed Logistic Regression and Support Vector Machine model gives higher accuracy for the binary classification of the text data into personality traits with an accuracy

score 78.01% for EXT, 79.63% for AGR, 58.91% for NEU, 74.21% for CON, and 81.56% for OPN traits. However, the Naive Bayes algorithm resulted in overall lower performance.

For future work, we plan to investigate the impact of deep learning techniques with neural networks for better accuracy. The personality traits recognition can be extended to audio and video emotion recognition with help of Convolutional Neural Networks and Speech recognition methods.

References

Elena Bajic. How the mbti can help you build a stronger company—forbes, 2015. URL https://www.forbes.com/sites/elenabajic/2015/09/28/how-the-mbti-can-help-you-build-a-stronger-company/?sh=51754881d93c. [Online; accessed 1st-September-2021].

Md Abdur Rahman, Asif Al Faisal, Tayeba Khanam, Mahfida Amjad, and Md Saeed Siddik. Personality detection from text using convolutional neural network. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–6. IEEE, 2019.

Kasula Chaithanya Pramodh and Y Vijayalata. Automatic personality recognition of authors using big five factor model. In *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, pages 32–37. IEEE, 2016.

Raad Bin Tareaf, Seyed Ali Alhosseini, Philipp Berger, Patrick Hennig, and Christoph Meinel. Towards automatic personality prediction using facebook likes metadata. In *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 714–719. IEEE, 2019.

Rapha el Lederman Anatoli de Bradk e, Ma el Fabien and St ephane Reynal. Multimodal emotion recognition. *Projet Fil Rouge 2018-2019*, 2018. URL https://www.overleaf.com/project/5c06f9e12aee1927b458fc4a.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79, 2017.

Abir Abyaa, Mohammed Khalidi Idrissi, and Samir Bennani. Predicting the learner's personality from educational data using supervised learning. In *Proceedings of the 12th International Conference on Intelligent Systems: Theories and Applications*, pages 1–7, 2018.

Waiel Tinwala and Shristi Rauniyar. Big five personality detection using deep convolutional neural networks. 2021.

KN Pavan Kumar and Marina L Gavrilova. Personality traits classification on twitter. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

Kaggle. Myers-briggs personality type dataset, 2017. URL https://www.kaggle.com/datasnaek/mbti-type. [Online; accessed 16-August-2021].

Isabel Briggs Myers. The myers-briggs type indicator: Manual (1962). 1962.

Paul T Costa Jr and Robert R McCrae. *The Revised Neo Personality Inventory (neo-pi-r)*. Sage Publications, Inc, 2008.

Michal Kosinski, Sandra C Matz, Samuel D Gosling, Vesselin Popov, and David Stillwell. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American psychologist*, 70(6):543, 2015.

James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6): 1296, 1999.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500, 2007.

Adrian Furnham. The big five versus the big four: the relationship between the myers-briggs type indicator (mbti) and neo-pi five factor model of personality. *Personality and Individual Differences*, 21(2):303–307, 1996. ISSN 0191-8869. doi: https://doi.org/10.1016/0191-8869(96)00033-5. URL https://www.sciencedirect.com/science/article/pii/0191886996000335.

GeeksforGeeks. Svm hyperparameter tuning using gridsearchcv ml— geeksforgeeks, 2019. URL https://en.wikipedia.org/wiki/Support-vector_machine#Bayesian_SVM. [Online; accessed 2nd-September-2021].

KDnuggets. Algorithms, decision trees, explained. URL https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html. [Online; accessed 3rd-September-2021].

Ashwin Raj. Perfect recipe for classification using logistic regression, 2020. URL https://towardsdatascience.com/the-perfect-recipe-for-classification-using-logistic-regression-f8648e267592. [Online; accessed 3rd-September-2021].