IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PERSONALITY IDENTIFICATION BASED ON MBTI DIMENSIONS USING NATURAL LANGUAGE PROCESSING

[1]Kishan Das, [2]Himanshu Prajapati,

[1]M.E. Student, [2]Assistant Professor
[1,2]Department of Computer Engineering,
[1,2]Silver Oak College Of Engineering and Technology, Ahmedabad, India

Abstract: Personality means prevailing characteristics, habits, and behaviors that encourage a person to react, communicate, and think in a unique direction. It dictates, how a person acts, communicates, responds, and establishes his/her desires. Personality analysis makes a person unique in terms of his/her personality traits. Personality analysis can be beneficial in several real-life scenarios such as good or service and course recommendations, candidate screening, advertising, health advice, relationship counseling, and much more. In this research work, our main objective is to create a model that can analyze personality attributes of an individual which is based on named entity identification and a combination of feature engineering techniques including TF-IDF count vectorization and word2vec embedding. The most popular and open source Myers-Briggs Type Indicator (MBTI) dataset consisting of more than 8500 user-written posts has been used for evaluation purposes. We have implemented the most widely used and well-known ensemble learning models for experimentation consisting of Bagging, Boosting, and Stacking. A comparative analysis of the predictions made by the Ensemble Learning Algorithms has been carried out to determine the best overall performance. The results demonstrate that the stacking model delivers the best performance with an accuracy of 97%.

Index Terms - **Personality analysis, Natural language processing, named entity recognition, feature engineering, MBTI, ensemble learning models.**

I. INTRODUCTION

In psychology, personality is a psychological model that can be used to explain a range of individual behavior associated with the individual's characteristics. It includes natural and inherited psychological patterns that distinguish among people and the surroundings and the social community. In recent years, social media sites such as Facebook, LinkedIn, Twitter, etc. have been expanding exponentially and become influential and straightforward approach for both socialization and the transmission of relevant user information. This vast amount of useful information can be utilized to identify the personality traits of an individual. Since these different social media platforms emerge, users have significantly used them to convey their tastes and preferences, feelings, viewpoints and to exchange personal details like professional life without ever being judged [3]. These gestures demonstrate the characteristics of a person and his/her personality. It also tends to be a strong connection between a person's personality and their online social networking involvement. Every person has unique preferences, tastes, behaviors, and many other essential characteristics that distinguish them from another person. It enables us to distribute these different variants into separate categories, so that the effectiveness of advertising, promotion, market research, interview assessment, performance appraisal, and many other objectives can be improved [8]. From the survey, we observed that with the use of personality analysis frameworks such as BIG Five, MBTI, etc. a plethora of researchers pursues to automatically predict and classify personality traits using the posts published in social networking sites. Moreover, there are several approaches presently available and are implemented to evaluate the personality of a person, but many approaches have serious limitations and are relatively weak in identifying and analyzing characteristics. This research paper proposes a methodology to analyze and predict personality traits based on the four MBTI dimensions[7] and incorporates Natural language processing techniques and feature extraction methods to develop relevant feature vectors for training and validation with an ensemble learning models. Eventually, the performance of the proposed model is validated and compared with an open-source MBTI dataset and three popular machine learning ensemble methods.

The structure of this research paper is described as follows. Section II represents some important related work done for personality traits prediction. Section III represents the architecture and the methodology of the proposed model with the steps implemented for building a feature vector. Section IV represents the experimental results performed on the machine learning models. Ultimately, we sum up the conclusions and suggestions for future investigation in Section V.

II. Related work

The researchers of paper [1], created a framework for personality traits classification based on twitter posts that incorporates linguistic features and a combination of number-based encoding and word embeddings. They have used the MBTI dataset for training and validating the classification algorithms. For classification, they implemented XGBoost and SVM supervised machine learning algorithm. Before training and validation, the original text sentences are submitted to the preprocessing step where stopword filtration, email, URL, punctuation removal, and lemmatization methods are applied. A weighted summation of vector representation of each relevant text document is created by employing the TF-IDF vectorizer and GloVe word embedding and submitted to the classification algorithm to make a prediction. In paper [2], the author's Daniel Ricardo Jaimes Moreno et al. proposed a methodology that can predict an individual’s personality attributes with the use of latent features. For the evaluation and experimentation, they have employed the PAN CLEF dataset consisting of more than 14000 posts. In the preliminary step, each text document is subsequently broken up into five features based on contents such as phrases, urls, emoji, and mentions and then irrelevant terms such as punctuations, long or short words, stopwords are filtered out from the original sentence. Following this, the preprocessed texts are converted into document term matrix representation with the use of TF-IDF, and latent features are extracted by submitting this matrix representation to three distinct feature selection methods: NMF, PCA, and LDA. To evaluate the performance, a comparative analysis has been performed with TF-IDF representation as a core feature and latent features obtained from the dimensionality reduction techniques with three supervised classification algorithms named logistic regression (LR), support vector classifier (LSVC), and random forest (RF). The authors of research work [4], developed a system for predicting personality characteristics relying on semantic features. They have used myPersonality dataset for training and validation of the proposed system. The semantic characteristics have been used in this proposed approach to measure the similarity between the user text document induced vectors with the personality trait vectors. In the first step, they performed the fundamental preprocessing techniques on the user-generated text documents and removed the irrelevant entities from them such as removing meaningless or useless words, parts-of-speech tagging, etc. Besides, three vector representations (vec1, vec2, vec3) were formed to find the most important interpretation for identifying personality traits. For testing and validation, they evaluated the semantic similarity distance between the user text generated matrix and the personality attributes matrix by employing path-based and information content-based measures that are based on relatedness scores. Also, they assessed the performance of their proposed system on the basis of accuracy, precision, recall, and f-measuring metrics. The publishers of research work [8], proposed a framework that uses computational language characteristics to classify an individual’s personality traits. They have implemented three types of approach 1. Machine learning-based approach 2. Langauge based approach and 3. Grammatical based approach to provide the word assessment for each personality trait and to investigate the accurate identification of text-based personalities in the Indonesian language. Naive Bayes classifier algorithm is implemented in the first approach as it is relatively simple and easy to execute. In the second approach, the TF-IDF vectorization[12] technique has been accomplished to measure how relevant a word has occurred in each document while in the third approach, the unintended words are removed by analyzing the context of the words. Ultimately, personality characteristics are categorized and evaluated based on the three statistical approaches mentioned above.

The proposed approach in this research work is separated into four steps. 1) Pre-processing 2) Feature Engineering 3) Training and Validation 4) Evaluation. In the first step, the fundamental text pre-processing techniques such as similarity hashing, named entity recognition, lemmatization are implemented to filter out the irrelevant terms/ entities from the original text documents. In the second step, the TF-IDF vectorizer[12] and word2vec[9] embedding methods are employed to build a feature word vector consisting of important features extracted from the preprocessed text document. In the third step, the extracted features from the previous step are provided as an input to the supervised machine learning classification algorithms for training and validation. In the fourth or final step, the performance of the trained machine learning models is evaluated with evaluation metrics such as accuracy, recall, f1-score. Thereafter, a new or unseen user written text documents are submitted to the trained models to predict personality traits.

III. Methodology

3.1 Dataset Description

In this proposed methodology, we have used the popularly used and open source Merys-Briggs Type Indicator dataset for personality analysis. The MBTI dataset is premised on the psychologist Carl G. Jung’s personality traits hypothesis, and distinguish a person into sixteen personality labels among four dimensions. The dimensions are listed below:

- Introversion-Extroversion (I-E)
- Sensing-Intuition (S-I)
- Thinking-Feeling (T-F)
- Judging-Perceiving (J-P)

The dataset comprises of more than 8500 unique rows and is referred to as one user for each row. There are two columns in the MBTI dataset. The first column composed of 16 distinct labels and every label is a mixture of four MBTI dimensions and the second column includes the lastest 50 user written posts distinguished by ||| symbol[7].

3.2 Data Preprocessing

After analyzing the dataset structure, we have employed natural language processing methods to recognize the important patterns and features and removing the irrelevant and meaningless entities and keywords from the raw text document. The preprocessing step is further subcategorized into three phases: data deduplication, named entity identification, lexicon normalization. In the first phase, the duplicated or redundant text documents are removed from the dataset based on how two documents are contextually similar to each other. To achieve this, we have implemented a simhash algorithm that divides each document into several segments and generates a hash for each segment. The algorithm gives a similarity score for documents on the basis of the hashes. And thereafter, the named entity recognition component of natural language processing has been introduced to text documents to recognize and delete unreliable and useless entities and keywords from the user posts. These recognized entities can be used as additional patterns for building feature vectors. In the last phase of pre-processing, word normalization methods are applied to text documents to convert phrases into their stem/lemma form. Other pre-processing methods are used, such as stopword removal, tokenization, lowercasing.

3.3 Feature Engineering

Using the named entity recognizer, the identified entities or keywords from the user posts such as persons, places, locations, organization, products, events, etc. are attached to the preprocessed dataset as additional features. Apart from this, sentiment analysis is performed on each text document to add an extra sentiment feature to the preprocessed dataset. Sentiment analysis is the process of representing contextual thoughts of text information through values or labels.

The TF-IDF vectorizer[12] is then used to convert the derived important features into actual values. TF-IDF is referred to as Term Frequency-Inverse Document Frequency, which measures the significance of a term in the set of documents. In general, this determines a value for a word indicating the effectiveness of the word in the dataset. Following this, we have implemented the most popular and widely used word2vec embedding technique to transform the extracted feature into a document term matrix representation. Eventually, the vocabulary generated by TF-IDF vectorization[12] and word2vec[9] word vector representation is combined and concatenated to create a resultant feature vector.

3.4 Classification

For the classification of personality traits, we have implemented the three well-known ensemble learning methods[11] such as Boosting, Bagging, and Stacking models[14, 15]. In the preliminary step, we trained and validated the developed features with the XGBoost algorithm. XGBoost algorithm is based on decision tree gradient boosting techniques and is very fast and efficient for classification problems. It is a sequential ensemble learning algorithm. Besides, we used the random forest algorithm for personality prediction. Random forest is a bagging ensemble method composed of several decision trees. Unlike XGBoost, in the random forest, each decision tree works in parallel and made predictions for each subset and then the majority of the vote decides the outcome. The stacking model consists of two levels 1) base-level and 2) meta-level. The base-level consists of weak learning algorithms, while the meta-level consists of a strong learning algorithm. In the stacking process, the assumptions produced by weak learning algorithms are integrated and forwarded to the meta-level algorithm to create a new assumption that is more reliable. Finally, the assumptions generated by these ensemble learning models[11] are measured and compared to attain the optimum overall performance. Figure 1 shows the workflow of the Proposed approach.

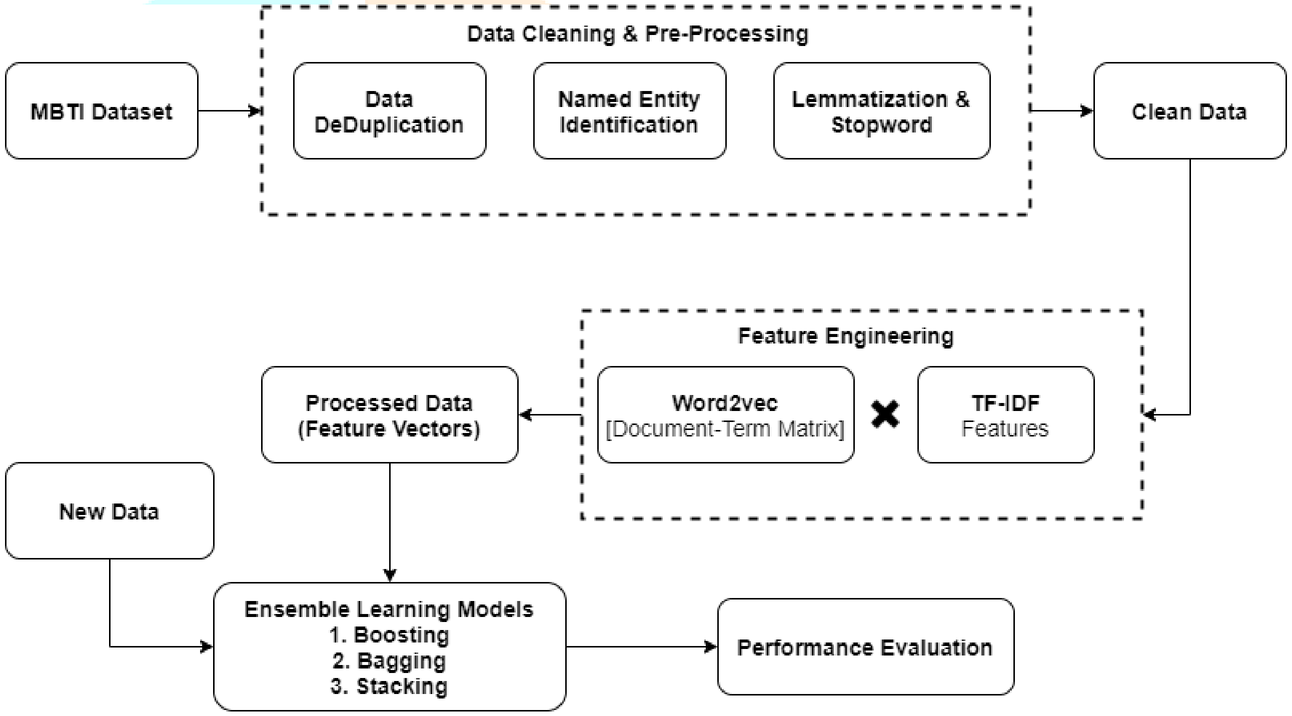

Figure 1 Personality Analysis Framework

IV. Results

For experimentation, we have used the four well-known evaluation metrics: Accuracy, Recall, Precision, and F1-score. For each of the four MBTI dimensions, such as introversion-extroversion, sensing-intuition, thinking-feeling, and judging-perception, these four metrics are evaluated. Table 1 shows the evaluation results of the boosting model. From the result, we can observe that the highest accuracy of 95.79% obtained for the Sensing-Intuition class while for Introversion-Extroversion, Thinking-Feeling, and Judging-Perceiving class the accuracies noted are 88.02%, 77.69%, and 71.96% respectively.

Table 1 Results of Boosting Model

MBTI Type	Accuracy	Precision	Recall	F1-score
Introversion/Extroversion	88.02%	87.26%	88.71%	87.98%
Sensing/Intuition	95.79%	95.07%	96.91%	95.98%
Thinking/Feeling	77.69%	75.45%	79.54%	77.44%
Judging/Perceiving	71.96%	71.93%	73.28%	72.60%

The performance evaluation of the bagging model is presented in Table 2. As a result, the accuracies obtained are 86.11%, 93.80%,74.93%, and 71.01% for the MBTI dimensions Introversion-Extroversion, Sensing-Intuition, Thinking-Feeling, and Judging-Perceiving. It shows a slightly lower performance compared to the Boosting model result.

Table 2 Results of Bagging Model

MBTI Type	Accuracy	Precision	Recall	F1-score
Introversion/Extroversion	86.11%	88.26%	83.00%	85.55%
Sensing/Intuition	93.80%	97.01%	90.59%	93.69%
Thinking/Feeling	74.93%	74.61%	75.85%	74.99%
Judging/Perceiving	71.01%	69.53%	75.86%	72.56%

Table 3 represents the performance of the stacking model. From the results, we can notice that the best performance achieved by the Sensing-Intuition class with an accuracy of 97.53% and precision at 98.03%. For the other three MBTI classes, the accuracy observed is 91.13%, 79.39, and 73.53%. Overall, the stacking model outperforms and shows the best performance for all the MBTI dimensions when compared to the boosting and bagging models.

Table 3 Results of Stacking Model

MBTI Type	Accuracy	Precision	Recall	F1-score
Introversion/Extroversion	91.13%	93.47%	88.33%	90.82%
Sensing/Intuition	97.53%	98.03%	96.81%	97.42%
Thinking/Feeling	79.39%	81.18%	75.84%	78.42%
Judging/Perceiving	73.53%	71.49%	77.22%	74.25%

V. Conclusion

The main objective of this research work is to build an automatic personality analysis model to simulate the personality attributes of an individual in a more efficient way. In the preliminary step, the original text documents from the MBTI dataset are preprocessed with the conventional text preprocessing techniques. Deeper feature engineering is carried out on the text documents through the use of natural language processing. Subsequently, the relevant feature vectors are generated by combining the TF-IDF features and the word2vec vector representation. With these feature vectors, three different ensemble learning methods are implemented and then a comparative analysis was carried out to find the best overall performance. From the experimental results, we can conclude that the stacking ensemble model outperforms the other two ensemble learning models and provides the best overall performance for all the MBTI dimensions. Earlier research work was conducted with an accuracy of 88%[7]. The MBTI data set was further analyzed in this research work and an accuracy of 97% was witnessed.

In future work, a detailed evaluation can be carried out to examine the original intent behind the use of phrases from the user-written information. Based on the nature of the words, for instance, the words 'grey,' 'sorrowful' and 'morose' can reflect a variety of depressive intensities which can create an immense improvement during the analysis. In text documents, the same phrase written differently can convey two different perceptions that can be identified by the word used alone.

REFERENCES

[1] K. N. P. Kumar and M. L. Gavrilova, "Personality Traits Classification on Twitter," 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-8, 2019.

[2] D. R. Jaimes Moreno, J. Carlos Gomez, D. Almanza-Ojeda, and M. Ibarra-Manzano, "Prediction of Personality Traits in Twitter Users with Latent Features," 2019 International Conference on Electronics, Communications, and Computers (CONIELECOMP), pp. 176-181, 2019.

[3] P. S. Dandannavar, S. R. Mangalwede, and P. M. Kulkarni, "Social Media Text - A Source for Personality Prediction," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), pp. 62-65, 2018.

[4] M. Hassanein, W. Hussein, S. Rady, and T. F. Gharib, "Predicting Personality Traits from Social Media using Text Semantics," 2018 13th International Conference on Computer Engineering and Systems (ICCES), pp. 184-189, 2018.

[5] I. B. Drexel, "Feature Engineering and Word Embedding Impacts for Automatic Personality Detection on Instant Message," 2019 International Conference on Information Management and Technology (ICIMTech), pp. 155-159, 2019.

[6] M. A. Rahman, A. Al Faisal, T. Khanam, M. Amjad, and M. S. Siddik, "Personality Detection from Text using Convolutional Neural Network," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), pp. 1-6, 2019.

[7] S. Bharadwaj, S. Sridhar, R. Choudhary, and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1076-1082, 2018.

[8] L. C. Lukito, A. Erwin, J. Purnama, and W. Danoekoesoemo, "Social media user personality classification using computational linguistic," 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 1-6, 2016.

[9] I. U. Ogul, C. Ozcan, and O. Hakdagli, "Keyword Extraction Based on word Synonyms Using WORD2VEC", 2019 27th Signal Processing and Communications Applications Conference (SIU), pp. 1-4, 2019.

[10] X. Jin, S. Zhang and J. Liu, "Word Semantic Similarity Calculation Based on Word2vec," 2018 International Conference on Control, Automation and Information Sciences (ICCAIS), pp. 12-16, 2018.

[11] Y. EMRE ISIK, Y. GÖRMEZ, O. KAYNAR, and Z. AYDIN, "NSEM: Novel Stacked Ensemble Method for Sentiment Analysis," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), pp. 1-4, 2018.

[12] C. Liu, Y. Sheng, Z. Wei and Y. Yang, "Research of Text Classification Based on Improved TF-IDF Algorithm," 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), pp. 218-222, 2018.

[13] J. Lilleberg, Y. Zhu and Y. Zhang, "Support vector machines and Word2vec for text classification with semantic features," 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), pp. 136-140, 2015.

[14] B. Pavlyshenko, "Using Stacking Approaches for Machine Learning Models," 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), pp. 255-258, 2018.

[15] S. Li et al., "Stacking-Based Ensemble Learning on Low Dimensional Features for Fake News Detection," 2019 IEEE 21st International Conference on High-Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pp. 2730-2735, 2019.