Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020)
IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2

Analysis of Personality Traits using Natural Language Processing and Deep Learning

Tejas Pradhan, Rashi Bhansali
Dept of Comp Engg, Vishwakarma Institute of Technology, Savitribai Phule Pune University, Pune, India
tejas.pradhan18@vit.edu, rashi.bhansali18@vit.edu

Dimple Chandnani, Aditya Pangaonkar
Dept of Comp Engg, Vishwakarma Institute of Technology, Savitribai Phule Pune University, Pune, India
dimple.chandnani18@vit.edu,aditya.pangaonkar18@vit.edu

***Abstract* — Personality test and analysis is an important factor in an individual's overall development. Most known personality test followed by people is Myer Briggs Test Indicator. While these tests are conducted by psychologists, it is very easy to deceive them in order to get the personality type of our choice, as these questions are very straight forward. This paper focuses on automating this task with the help of Neural Networks by using images instead of questions. A labelled dataset with user responses on social media along with their personality type is used for analysis. After cleaning, relevant response features are extracted using NLP followed by applying suitable classification algorithms. A model is trained to accurately predict the personality type of users based on their responses, which is then deployed on an interactive website in the form of a personality test.**

Keywords* — *Deep Learning, Machine Learning,* Myer Briggs Test Indicator, *Neural Networks, NLP

I. INTRODUCTION

Personality tests have gained popularity in the past few years since they are a pre-requisite for most institutions and companies. The main problem faced by individuals who want to take a personality test is the cost and the length of the test. Most personality tests have 50-70 questions, which becomes very cumbersome for the user. Our project aims to optimize this process by reducing the amount of text entered by the user and is completely free of cost. Our test is based on the Myer Briggs Test Indicator (used as MBTI in rest of the sections) which a very well-known form of assessment. According to this test, there are 16 types of personalities. It identifies one's 4-letter personality type, with each letter corresponding to a specific personality preference or tendency. They are Introversion (I) vs. Extraversion (E), Intuition (N) vs. Sensing(S), Thinking (T) vs. Feeling (F) and Judging (J) vs. Perceiving (P).

The test is available on an interactive website. The user needs to provide his interpretation of an abstract image in about 50-60 words. This idea is based on the fact that each personality thinks and expresses differently. Complete analysis is given at the end of the test along with the percentage of each trait. This could be very beneficial for psychological applications as well as the industry. Personality analysis not only helps in choosing the right career or academic courses but also helps in bringing out the best in oneself. Feedback is also taken to further train and optimize the model.

II. RELATED WORK

The MBTI is the questionnaire indicating different psychological aspects to identify how people perceive the world and make decisions. As in [1], they demonstrated SVM as classifier to classify different emotions mainly anger, happiness, sadness, neutral and fear. They used Berlin emotional database and LIBSVM for classification of emotions. They achieved 93.75% classification accuracy for gender independent case, 94.73% for male and 100% for female speech. As in [2], their approach uses Multilevel Perceptron (MLP) neural network and back propagation for learning. They have used 5 neural networks for each personality trait. Each position in the vector is forwarded to the classification module whose output is generates as either yes or no. The system tested 100 users and could predict personality based on OCEAN accurately.

As in [3], they used the 3-factor personality test and used a website for presenting the results. The system had 91% accuracy when tested with a real world questionnaire As in [4], they used SVM and CNN but achieved only 55% average accuracy. Using standalone MLP improved the results with accuracy at 62.68%. As in [7], they extracted keywords from paper by Latent Dirichlet allocation, done the clustering by KNN and classified using Term frequency-inverse document frequency (TF-IDF) values of each paper.

As in [9], they used graphology and trained a CNN model on handwriting images that highlighted features like curves and slant in letters, crossing of 't's and pressure applied while writing, for personality analysis. They conclude to have achieved an accuracy of 93.77%. As in [11], they put forth a recommender system for candidates applying for a job, by comparing the person's Curriculum Vitae (CV) with the job requirements using TF-IDF. Their system further analyses the personality of applicant through the CV.

As in [12], they used TensorFlow AI Engine and CNN to examine personality from video interviews in order to automate the recruitment process through Automatic Personality Recognition (APR). As in [13], they use a social media response data collected from YouTube and Twitter for personality evaluation. Through a detailed error study, they have plotted the trends for each personality trait against the

respective social networking site. As in [14], they have worked on presenting the computational abilities of Machine Learning models in extracting fine human behavioral traits from mobile sensing prints. As in [15], they use computer vision along with CNN to extract features from videos and thereby examine personality using a Deep Bimodal Regression.

III. METHODOLOGY

The implementation of this research project involves four main steps, depicted in figure 1.

a. Text mining that involves structuring the input data and parsing it followed by removal of unimportant elements and cleaning.
b. Machine Learning Models for personality classification.
c. Deep Learning Models for personality classification.
d. Development of an interactive website for the analysis based on user input.

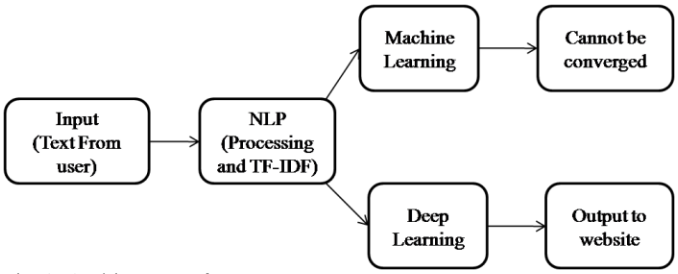

Fig. 1. Architecture of system

A. Text Mining using Natural Language Processing Techniques

For the purpose of this research, labelled data was collected from different people's posts on a personality test forum – personality cafe along with their personality types. A total of 8000 data points were used as a part of this research in the dataset. This included all the 16 MBTI types.

Data Preprocessing

As the dataset was collected from a huge social media data, it contained a lot of slang and words which could not be used as meaningful semantic features for identifying one's personality. The data also contained various punctuation marks and a lot of emoticons which had to be cleaned and removed from the dataset. The dataset went through rounds of data cleaning wherein all the redundant punctuations, emoticons, and "stop words" like 'a', 'the', etc. were removed. This was achieved using regular expressions in python and the NLTK text processing library. This approach was similar to that used in [8]. A clean data set was obtained with only relevant words which were important for analyzing the personality traits. According to MBTI, some personality types are more common than others. Hence, there is abundance of certain personality types in the dataset. Fig.2. shows the distribution of the personality types in the dataset. This imbalance in the data can be removed while training using oversampling the minority classed by using techniques like SMOTE (Synthetic Minority Oversampling Technique). However, our aim was to model the real-world scenario as much as possible. Hence, this paper focuses on optimizing the hyper parameters in the model while keeping the imbalance in the dataset constant to obtain greater performance. The text in the cleaned dataset was grouped by personality types to obtain a final dataset which consisted of personality classes and raw text entered by these personalities. Figure 2 depicts distribution of personality types.

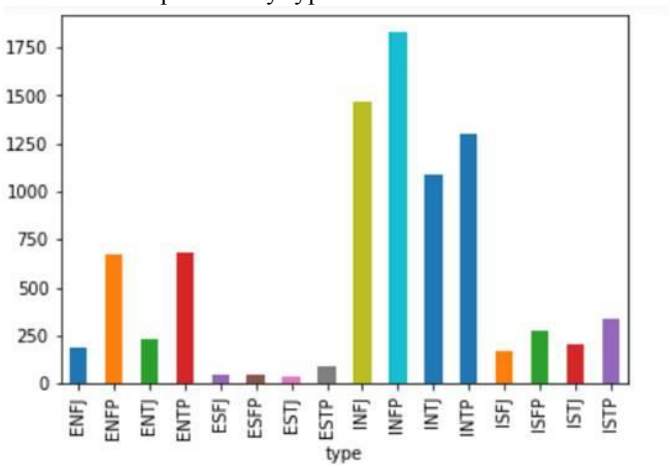

Fig. 2. Distribution of personality types

Feature Extraction

Raw text cannot be directly fed to the model. Meaningful features need to be extracted from this text in order to assess a person's personality types. The paper builds around the idea that specific personality types use specific words which is used to classify a person's personality trait. Therefore, in order to classify personality based on words, certain weight needs to be added to the words. This is done by a process called Vectorizing. The unstructured text was converted into vectors (mathematical values) based on their frequency and weight using the Count Vectorizer and Term Frequency- Inverse Document Frequency Vectorizer models. These are two of the most used vectorizing algorithms in Natural Language Processing. TF-IDF, as in [10] counts the frequency of the word used by the current data point in terms of the inverse of the number of times that particular word was used in the entire document. In the context of personality, TF-IDF measures how often a word was used by a given personality type as compared to all the personality types. TF-IDF performs the best in handling informal text data because it also takes into consideration the entire document and not a particular sentence from it. Hence, the error caused due to excessive usage of a particular word by a particular personality type is ignored and the only way that a word is assigned greater priority is if it is used by a particular personality and not by other personalities. In this process, all the raw text was converted to numerical features to obtain a completely numerical dataset. This was then fed to different machine learning and deep learning models for training purposes.

B. Machine Learning

The dataset obtained was trained on a total of 3

classification algorithms, namely Support Vector Machines, Naive Bayes' Classifier and Random Forests Classifier. 5600 samples (70%) were used for training and the rest 2400 were used for testing purposes. After tuning them for hyper parameters, these machine learning models could produce up to 60% accurate results as shown in the Table 1.

TABLE I. COMPARATIVE ANALYSIS OF ALL ALGORITHMS

Sr. No.	Algorithm	Classification Accuracy (%)
1	Naïve Bayes' Classifier	32.63
2.	Random Forest Classifier	36.03
3.	Support Vector Machines	57.90
4.	Convolutional Neural Networks	81.40

A basic Naïve Bayesian classifier was used to classify the data. The training parameter 'alpha' was tuned to 0.32 in order to fit the bias of the data. Fig. 3.1 shows the plot of change in the accuracy of classification with respect to the hyper parameter. This approach was similar to that in [5], however, the model gave only 32% accuracy on the validation set because of the high bias and multiclass classification. Further, a random forest classifier was trained on the dataset which gave an accuracy of 36.03%.

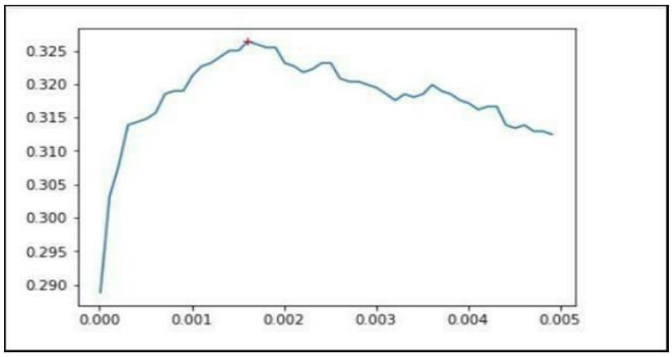

Fig. 3.1. Tuning of Naïve Bayes' Classifier

The next model used was a linear Support Vector Machines algorithm. The regularization parameter C was set to 0.16 for increasing the fitting towards the bias of the data. This approach was similar to that in [6]. The SVM model gave an accuracy of 57.9% on the validation set as shown in Fig. 3.2.

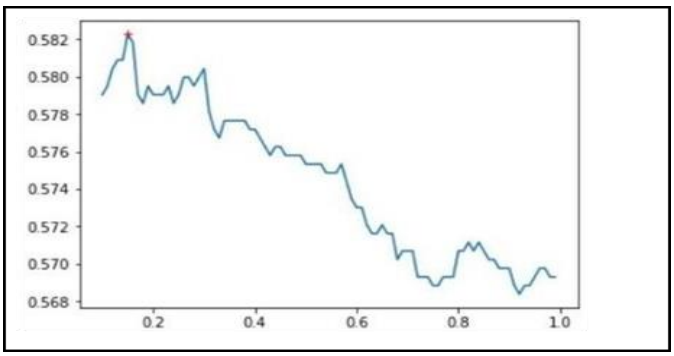

Fig. 3.2. Tuning of Support Vector Machines Classifier

C. Deep Learning – (Convolutional Neural Networks)

In order to improve the results, Convolutional Neural Networks, a deep learning approach conventionally used on images was used. In this implementation of CNN, 1D convolution of the word vectors is used. Each word will be converted to a 1-dimensional vector by an embedding layer. These vectors would be the inputs to the CNN layer. Unlike images, the input will be 1 dimensional. Hence, 1D vector CNN algorithm is used to take the input vectors, and assign learnable weights and biases based on various aspects of feature extraction and max pooling, leading to better differentiation between words. However, the accuracy could not improve because of the number of classes to predict. To solve this problem an approach called multilabel classification was used with the deep learning model. Instead of predicting 16 classes for 16 personality types, 4 binary classifiers were used in order to get better precision, recall and accuracy. Hence, the output layer consisted of 4 neurons with a sigmoid activation function instead of 1 neuron with soft-max activation. 1D convolution of word embeddings were created and fed as input to the neural network. The neural network further consisted of a max pooling layer and a dense layer which was followed by a sigmoid layer for obtaining multilabel classification results of the 4 binary classifiers. With detailed analysis, tuning of the number of hidden layers, activation functions at inner levels as well as the output, and max pooling, it can achieve around 81.4% accuracy in the multilabel classification. Using the TF-IDF vectorization followed by a meticulous hyper parameter tuning led to a much better accuracy as compared to that in [3] which used n-gram model with CNN. As a result of this training, Convolutional neural network model is used in the final system for the testing.

D. Development of the Website

Graphical looks and feel according to the most impressive way were the top priority for us. The proposed work used CSS and bootstrap for styling and making the website attractive. Graphical elements, images, navigation bars and use of different colors were carried out. The colors in the website were used because of a certain purpose as each color defines a specific meaning and adds more value to the website making it more evocative and expressive for the user, as shown in figure 4.

Fig. 4. Website homepage

Writing of content is a significant part in development of a web page and plays an important role in optimization. And hence this work is focused on using well-defined, pure, and accurate content which gave precise information regarding the different personality traits and our algorithm. Development of the web pages was done using HTML and CSS scripts. HTML was used for the text-based content on the website and to align and place text in the most efficient way possible. CSS was used to style the website to make it look user-friendly and attractive to the user. Figure 5 depicts a user input on abstract images.

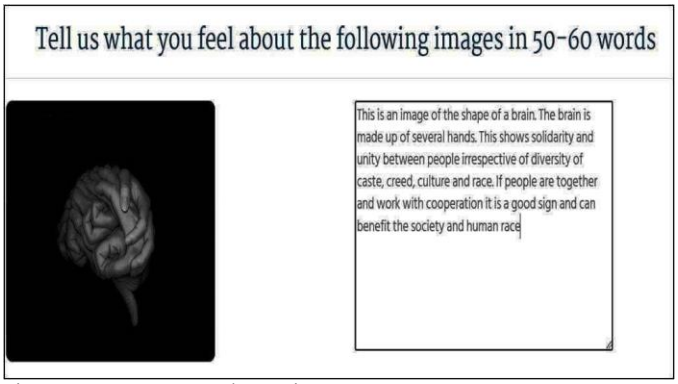

Fig. 5. User response on abstract images

Testing was done to check browser compatibility of different objects used in the website as well as to check any broken links. This research work has used Google Chrome as it supported all the styling and links of our website. Testing was also done to check the load time of images and web pages.

To deploy the CNN model for application in the real world, the model is integrated with the backend of our website. A pickle file containing the weights of the pre trained CNN model was used for real time prediction of the personalities at the backend of the website. For this Flask is used, which is also known as a micro web framework written in Python is generally used for small-scale websites. It is classified as a micro framework because it does not require particular tools or libraries. Figure 6 depicts sample personality prediction output.

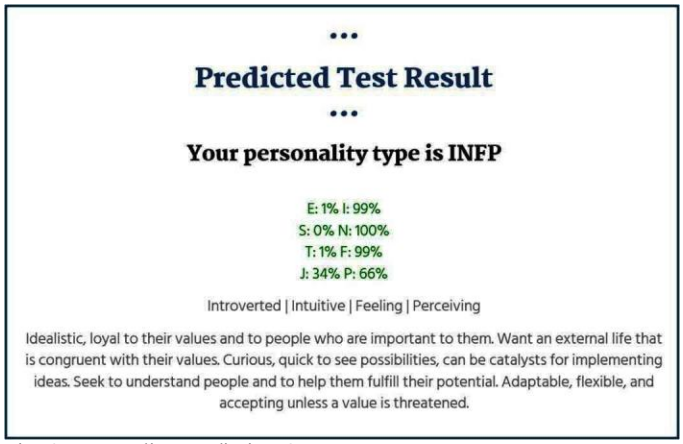

Fig. 6. Personality Prediction Output

IV. RESULT ANALYSIS

Table 2 shows the experimental analysis carried out in comparison with the previous work undertaken for the same in [2] and [3]. It also highlights the value addition by this methodology. Several machine learning and deep learning algorithms have been deployed by people before.

However, the proposed method not only performs better, but also yields the research into a tangible and practicable application in the form of a test that can be reliably used to successfully recognize personality of individuals.

TABLE II. COMPARATIVE WITH PREVIOUS WORK

Reference Paper	Methodology	Accuracy
Linguistic Features based Personality Recognition[2]	Naïve Bayes on social media data	91 %
Deep Learning- Based Document Modeling for Personality Detection from Text[3]	SVM CNN Multi-Layer Perceptron	55% 55% 62.6%
Proposed paper results	Naïve Bayes SVM CNN	32.6% 57.9% 81.4%

V. CONCLUSION AND FUTURE SCOPE

A personality test based on MBTI was created and was successfully deployed for its application in the real world through a fully responsive website. The test classifies the user's personality in as few as 60 words by using images instead of questions. The advantage of developing such a mechanism for classifying personalities is that it can identify personality traits with very less time and effort. The problem of attempting a minimum of 50-60 questions for a personality test result to be generated and which could be easily deceived was resolved. The outcome of this project was a definitive and efficient way of taking a personality test with the help of AI in a minimum amount of time, text with highly accurate results.

An API can be created for making this website lighter and the prediction can be done remotely through JSON request-response process. A chat bot can be deployed on the website

which will replace the text inputs with speech recognition and speech-to-text sequence models to enhance user experience. There is a bias in the model as some personality types are more common than the others. This can be tackled by scraping more data of the minority personality types. Additional models like recurrent neural networks can be used to improve the prediction accuracy by taking into consideration, the past results. The website becomes slow when there is too much traffic because the neural network model makes the back end heavy.

ACKNOWLEGEMENT

This work is an outcome of engineering design and innovation (EDI) which is an integral and unique component of our curriculum at Vishwakarma Institute of Technology, Pune, India. EDI is project centric learning where students can choose project area of their interest and faculty work as a mentor. We extend our sincere gratitude towards Dr. M L Dhore Sir for his consistent motivation and valuable guidance for proceeding from machine learning to deep learning. He thoroughly guided us in terms of mathematics behind the deep learning as well as at each step of writing this paper.

REFERENCES

[1] Yashpalsing Chavhan, M L Dhore, P Yesaware, "Speech Emotion Recognition Using Support Vector Machine", in International Journal of Computer Applications, vol.1, pp.6-9,February 2010
[2] Mayuiri Pundlik Kalghatgi, Mannjuli Rammanavar, Dr. Nandini Sindal "A Neural Network Approach to Personality Prediction based on the Big-Five Model" in International Journal of Innovative Research in Advanced Engineering, Issue 8, Volume 2, August 2015.
[3] Dilini Sewwandi, Kusal Perera, Sajith Sandaruwan, Oshani Lakchani, Anupiya Nugaliyadde, Samantha Thelijjagoda, "Linguistic features based personality recognition using social media data," in 6th National conference on Technology and Management, January 2017.
[4] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, Erik Cambria, "Deep learning-based document modeling for personality detection from text," published by the IEEE Computer Society , IEEE Intelligent Systems 2017.
[5] Manasi Ombhase, Prajakta Gogate, Tejas Patil, Karan Nair and Prof. Gayatri Hegde, "Automated Personality Classification using Data Mining Techniques" Pillai Institute of Information Technology.
[6] Joel Philip, "Machine Learning for Personality Analysis Based on Big Five Model," in IC Joel Philip, "Machine Learning for Personality Analysis Based on Big Five Model," in ICDMAI, Volume 2 in January 2019.
[7] Sang-Woon Kim and Joon Min Gil, "Classification systems based on TFIDF and LDA schemes," in Human Centric Computing and Information Sciences, 30, 2019.
[8] Dhore M L, Varpe K M, "Emotion Detection based on Images and Captions on Social Media" in International Conference on Analytical Innovation (ICAI) 2020, In press.
[9] Sandeep Dang and Prof. Mahesh Kumar, "Handwriting analysis of human behavior based on neural network," in International Journal of Advanced Research in Computer Science and Software Engineering, September 2014.
[10] Shahzad Qaiser, Ramsha Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents" in International Journal of Computer Applications 181(1), July 2018.
[11] Jayashree Rout, Sudhir Bagade, Pooja Yede, Nirmiti Patil, "Personality evaluation and CV analysis using machine learning algorithm" in IJCSE, May 2019.
[12] Hung-Yeu Suen, Chien-Liang Lin, Kuo-En Hung, "Tensor flow based Automatic Personality Recognition used in asynchronous video Interviews" in IEEE, March 2019.
[13] Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, Martine De Cock, "Computational Personality Recognition in Social Media" Unpublished.
[14] Clemens Stachl, Florian Pargent Sven, Hilbert Gabriella M. Harari, Ramona Schoedel, Sumer Vaid, Samuel D. Gosling, Markus Buhner, "Personality Research & Assesment in the era of machine learning" in European Journal of Personality, 28 May 2020.
[15] Zhang CL., Zhang H., Wei XS., Wu J. (2016) Deep Bimodal Regression for Apparent Personality Analysis. In Computer Vision – ECCV 2016 Workshops. ECCV 2016.