Personality Prediction Using Deep Learning

[1]Abhishek Jain, [2]Dr Sharda A. Chhabria

[1]Student, [2]Associate Professor
Artificial Intelligence Engineering Department
G.H. Raisoni Institute of Engineering & Technology Nagpur, India

***Abstract*: The attributes which characterize the person such as emotions, behaviour, mind and temperature define a personality of a person. The project aims to determine personality traits on the basis of Big Five Model and which is considered to be the multi-label classification problem. This project covered various machine learning algorithm and also deep learning techniques for training and testing of models. The steps for building the model are collection of data, pre-processing, feature extraction, splitting data, training, testing and implementing the model. The algorithm used in project are KNeighborsClassifier, Support Vector Machine, Gaussian Naïve Bayes, Long Short-Term Memory neural network. This model successfully classify the personality. The results indicate that LSTM performed best after that Gaussian Naïve Bayes, Support Vector Machine, KNeighborsClassifier. As per results, Gaussian Naïve Bayes perform well by achieving accuracy of 87% and precision of 82% before applying Long Short-Term Memory, However The performance of Support Vector Machine after applying Long Short-Term Memory is outstanding by achieving more than 83% precision and accuracy of 80% for different traits.**

***Index Terms*: K Neighbors Classifier (KNN), Support Vector Machine (SVM), Gaussian Naïve Bayes, Doc2vec, Long Short-Term Memory (LSTM).**

I. INTRODUCTION

Personality of an individual covers every aspect of life. It describes an individual's behavior and also influences daily life activities including emotions, preferences, motives and health [1]. Nowadays Recognition of personality from social networking has grabbed the attention of researchers and developers to develop the automated personality recognition systems. The ideology for developing such application is based on different personality models, like myPersonality model, Big Five Factor Personality Model [2], Myers Briggs Type Indicator (MBTI) [3], and DiSC Assessment [4].

The increasing use of Social Networking Platforms like Facebook and Twitter have shared ideas, sentiments, opinions with each other. As the comments, post, tweets made by users on this platform reflects their attitude, behavior and personality.[5] The existing work on Personality Prediction from social media text is based on supervised machine learning algorithm and deep learning techniques applied on benchmark datasets [6],[7],[8]. However, after study the major issue comes is the skewness of datasets, i.e., imbalanced dataset, which causes degradation of performance of personality recognition system.

In order to solve this issue different techniques are available for minimizing the skewness of the dataset, like Over-sampling, Under-sampling and hybrid-sampling [9]. When these techniques, are applied on the imbalanced datasets in different domain, have shown promising performance in terms of improved accuracy, recall, precision, and F1-score [10].

In this work, Machine Learning technique namely, Support Vector Machine, KNeighborsClassifier, Gaussian Naïve Bayes and deep learning technique like Doc2Vec, LSTM is implemented on benchmark dataset to classify the text into different personality traits such as Introversion-Extroversion(I-E), iNtuition-Sensing(N-S), Feeling-Thinking(F-T) and Judging-Perceiving(J-P). To improve the performance of the system, resampling technique [11] is also used for reducing the skewness of the dataset.

II. RELATED WORK

Predicting personality of an individual is a new and upcoming field. There is no extensive literature research on personality screening. Our article will be the first to provide readers with an overview of the latest trends and developments in this field.

A review of literature for personality prediction is shown in this section. The literature analysis of this work is categorized into two sub group, namely

1.supervised learning techniques.
2.deep learning techniques.

1.Supervised Learning Techniques:

A system was developed to predict personality of a user based on big five factor personality model from tweets posted in English and Indonesian language. In another work various classifiers are applied on myPersonality dataset. Result shows the accuracy achieved by Gaussian Naïve Bayes is 60%, which is better than SVM (59%) and KNN (58%). Even if the work didn't achieve the expected accuracy but it achieved the goal of predicting personality from twitter.

Automatic personality prediction system was proposed by using individual Facebook status text. Techniques like SMO for SVM, Logistic Regression, Multinomial Naïve Bayes. Results shows outstanding performance of multinomial naïve bayes over other. Using some other techniques may improves the result. Use of Feature selection techniques and more classifier will improve the overall performance.

A system is proposed which uses the social media tweets/post for predicting personality traits. The work includes data collection, data preprocessing, feature extraction and several machine learning algorithms for prediction. The feature vector is produces by

several feature extraction techniques like TF-IDF, LIWC, Emolex etc. this feature vector is then fed to several machine learning algorithm like SVM, Neural Net, Naïve Bayes for training and testing of these models. Enhancement of result can be made by implementing various State of the Art techniques.

The efficiency of different classifier is checked by using MBTI model to recognize the personality trait of individual from online text. Classifier namely, Naïve Bayes, Logistic Regression (LR), SVM and Random Forest are used in this work the result shows logistic regression with 66% accuracy for all MBTI traits. The improvement made by implementing XGBoost algorithm and some parameter tuning.

Table 1:Personality Prediction work based on Supervised Machine Learning Techniques

Sr. No	***Research***	***Goal and Objectives***	***Approach***	***Performance***	***Limitation and future work***
1	*Pratamo and sarno (2015) [16]*	*To predict personality from tweets posted in English and Indonesian language using Big Five Factor Personality Model.*	*Supervised* • *SVM* • *KNN* • *NB*	*Accuracy* *SVM=59%* *KNN=58%* *NB=60%*	*Semantic approach and extended dataset may improve the results.*
2	*Alam et al. (2013) [12]*	*Automatic identification of personality traits using status text from individual from Facebook.*	*Classification of personality prediction by SMO for SVM, Logistic Regression (LR), Multinomial Naïve Bayes (MNB).*	*MNB=61.79%* *LR=58.34%* *SMO=59.38%* *MNB performs better than other methods.*	*Use of different classifier and feature selection technique may improve the result.*
3	*Bharadwaj et al. (2018) [6]*	*Personality prediction from online text.*	*Neural Net, SVM, Naïve Bayes, TF-IDF, Emolex, LIWC.*	*SVM with all features performed best over others.*	*Neglect Word's gravity.* *Applying different state of the art technique may improve results.*
4	*Chaudhary et al. (2018) [13]*	*Predicting user's personality from online text using MBTI model.*	*Methodology namely, Naïve Bayes, Random Forest and LR are used for estimation.*	*Accuracy* *NB=55.89%* *LR=66.59%* *SVM=65.44%*	*Due to use of traditional methods accuracy is low but deep learning approach will definitely boost the accuracy.*

The table 1 represent the review of above cited studies for personality prediction and classification using supervised machine learning techniques.

2 Deep Learning Techniques:

Deep Learning is a sub branch of Machine Learning under the umbrella of Artificial Intelligence (A.I), which is acquiring knowledge from the experience from training data. Deep Learning perform the task regressively and improving the accuracy after every iteration.

A personality classification was made [15] which uses AttRCNN model to identify the personality traits from text. The model is capable of identifying the hidden features. Outcomes of this work shows that deep learning and semantic feature approach is better than any other techniques.

A deep learning Convolution Neural Network was made to classify the traits from big five factor personality model. This work was based on essay dataset. This model achieved the accuracy of more than 62.68%. Although the result was not improved, yet in future if LSTM approach may be applied for better efficiency.

A model proposed by B. Cui and C. Qi that use parts of text/post as input and classify personality. Various classifier like SVM, softmax as baseline and deep learning models are used. Results shows SVM outperformed with 34% train and 33% test accuracy and Naïve Bayes and softmax with 34% train and 33% test accuracy, while Deep learning model gives 40% train and 38% test accuracy, Although the accuracy is very low as it doesn't reach the boundary mark of 50%

Table 2 Personality Prediction work based on Deep Learning Techniques

Sr. No	*Research*	*Goal and Objectives*	*Approach*	*Outcomes*	*Limitation and Future Work*
1	*Majumder et al. (2017) [1]*	*Predicting personality from Big Five Personality Model.*	*CNN, Deep learning Approach*	Accuracy *NEU=59.38% EXT=58.09% AGR=56.71% CON=56.73% OPN=62.68%*	*In future more techniques would make improve results.*
2	*Xue et al. (2018) [15]*	*To predict personality traits from online text using deep learning methods.*	*Deep learning technique namely, AttCNN Approach.*	*MAE OPN= 0.3577 CON= 0.4251 EXT= 0.4776 AGR= 0.3864 NEU= 0.4273*	*In future other deep learning techniques will improve the results*
3	*Cui, and Qi (2018) [33]*	*A model that takes a part of post or text as input and classify it into different personality traits.*	*Multi-layer LSTM model*	*Overall accuracy=38% I/E=89.51% T/F=69.09% J/P=69.37% S/N=89.84%*	*In future word embedding technique will improve the result.*

The table 2 present the review of above cited studies for personality prediction and classification using Deep Learning techniques.

III. METHODS

The procedure of proposed system starts with data collection, preprocessing, feature extraction, training and testing with different machine learning and deep learning algorithm and lastly evaluation of models by several measuring parameters.

1.Data Collection:

The Benchmark dataset used in entire work is myPersonality dataset. It is the contribution of study made on psychological traits by university of Cambridge, which is collection of several examples of Facebook user attributes such as like, comment, post, status etc. Dataset is based on Big Five Personality Traits namely, Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism

2. Data Preprocessing:

Preprocessing plays an important role in any project, as it explores the data and useful for converting the data to desired format for further implementation. There is various preprocessing technique to validate the data like checking for null value, removing of unwanted row or column, etc.

3 Feature Extraction:

Feature extraction reduce the number of resources that defines the data. In general, there are many features in dataset which takes too much time and space computation. It also creates the problem of overfitting for classification problem, in order to avoid this feature engineering techniques like engineering techniques are used.

4 Training and testing with different Machine Learning and Deep Learning Algorithm:

After preprocessing and feature extraction the data is ready for implementation of different machine learning and deep learning algorithms. Using different Classifier like KNeighborsClassifier, SVM, Gaussian Naïve Bayes for the classification of various personality traits using Facebook text status data. The LSTM recurrent neural network model will help to boost the performance of algorithms and Doc2vec model of Gensim library of python convert the document into binary vector which is used to train the data using LSTM.

The algorithms covered in this work are as follows:

1.KNeighborsClassifier:

KNN was developed by Evelyn Fix and Joseph Hodge in 1951.It is one of the most fundamental classification algorithm and most commonly used algorithm. It is instant-based and non-parametric learning method. In this method classifier learns from the train data and then classify the new input by previously measured scores.

2.Support Vector Machine:

SVM is the supervised algorithm which was developed by Vladimir Vapnik, Boser, and Guyon in 1997 at bell laboratories. In general, if we have to classify the new data point to some classes then, SVM make the hyperplane in high-dimension space to classify the data point to the respected class in which they would be in. There may be many hyperplanes to classify the data points, but the best hyperplane to choose is decided by maximum separation between classes. So, the distance from the nearest point of each class from hyperplane should be maximum.

3.Gaussian Naïve Bayes:

Naïve bayes is probabilistic model. It based on bayes theorem, which name after Reverend Thomas Bayes. It holds the assumption that features are independent. while dealing with continuous data, it treats the data in classes as is distributed by Gaussian Distribution.

4.Doc2vec:

Doc2vec is the model of Gensim library which is open-source library which is used to convert the document to binary vector. It was developed by Mikilov and Le in 2014. The Doc2Vec model use following steps is as follows:

For Training data, Document is required A word vector **W** is generated for each word, and a document vector **D** is generated for each document. The model also trains weights for a softmax hidden layer. In the inference stage, a new document may be presented, and all weights are fixed to calculate the document vector.

5.Long Short-Term Memory:

Long Short-Term Memory is a recurrent neural network which is used in deep learning technique. An LSTM unit is composed of cell, an input gate, an output gate, and a forget gate. LSTM has feedback connections. The feedback network reduces the loss and increase the accuracy of overall model.

6.Evaluation of models by several measuring parameters:

Evaluation of models can't measure by only one parameter i.e., Accuracy. As only accuracy is not enough to represent the efficiency, so various evaluation technique like Confusion Matrix, F1 score, Precision and Recall parameters are used in this work to give overall evaluation.

IV. IMPLEMENTATION

This project is divided into 5 sub parts, so it is easy to analyze and work on it. The overall implementation of project started with a brief study related to the topic, Research papers and methodology helped us to analyzed the topic and understand the use of algorithms and their importance in this project. Algorithms like KNeighborsClassifier, Support Vector Machine, Gaussian Naïve Bayes, doc2vec, LSTM.

After selecting the algorithm, This paper is worked on implementation of such models. As this project is sub divided into parts, which is easy to work on time. The dataset is taken from open-source which is preprocessed with Keras and TensorFlow module. Algorithms are applied on train and test data. After complete evaluation integration of all the models in one jupyter notebook were done.

The overall process will a personality prediction system that will predict the personality on the basis of the text status of Facebook.

V. RESULTS

In this work we split the data into 70% training data and 30% testing data. For training of LSTM 20000 maximum features,100 epochs and 128 units are given. The results represent the efficiency of different classifier for predicting personality from text data. This work underlines the concept of machine learning and deep learning. Here are the Results of different classifier as follows.

1.KNeighborsClassifier:

In this work the result presents the accuracy score of KNeighborsClassifier to be 72%, while precision is 75% and recall is 74%. F1 score is 71%. After applying Doc2vec accuracy falls to 57%, precision turn to be 62%, recall is 61% and F1 score is 61.9%.

2.Support Vector Machine:

In this work the result shows the accuracy score of SVM to be 74.06%, while precision is 75% and recall is 74.2%. F1 score is 73%. After applying Doc2vec accuracy improves and becomes 79%, precision turn to be 83%, recall is 81% and F1 score is 82%.

3.Gaussian Naïve Bayes:

In this work the result shows the accuracy score of Gaussian Naïve Bayes to be 87%, while precision is 86.7% and recall is 86%. F1 score is 86%. After applying Doc2vec accuracy down to 53%, precision turn to be about 50%, recall is 49% and F1 score is 49%.

4.LSTM:

The accuracy score of LSTM is above 90%. which proves that using deep learning approach will definitely boost the result.

VI. SUMMARY

The motive of this paper is" Personality Prediction Using Deep Learning" by the help of Facebook text status collected in myPersonality dataset. This paper is results of the studies done on various machine learning and deep learning approaches. One can easily find the working of this algorithms by the help of this paper. The results show the applicability of this model in various field such as Psychology, Corporate Organization, colleges and academics etc. This work will help others in field of research and advancement in technology.

VII. ACKNOWLEDGMENT

Special Thank to Dr Sharda A. Chhabria Bot In charge (Centre of Excellence) for guiding throughout the whole process of building the system. For discussion on various algorithm and techniques for working of project.

REFERENCES

[1] N. Majumder, S. Poria, A. Gelbukh and E. Cambria, "Deep Learning-Based Document Modeling for Personality Detection from Text," in IEEE Intelligent Systems, vol. 32, no. 2, pp. 74-79, Mar.-Apr. 2017.

[2] L. R. Goldberg, L. R.," An alternative" description of personality": the big-five factor structure," Journal of personality and social psychology, vol. 59, no. 6, p.1216, 1990.

[3] I. B. Myers, "The Myers-Briggs Type Indicator: Manual" ,1962.

[4] D. Shaffer, M. Schwab-Stone and P. Fisher, "Preparation, field testing, interrater reliability and acceptability of the DIS-C," J Am Acad Child Adolesc Psychiatry, vol. 32, pp. 643-648, 1993.

[5] D. Xue et al., "Personality Recognition on Social Media with Label Distribution Learning," in IEEE Access, vol. 5, pp. 13478-13488, 2017.

[6] 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), Bangalore, 2018, pp. 1076-1082.

[7] M. Gjurković and J. Šnajder, "Reddit: A Gold Mine for Personality Prediction," In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pp. 87-97, 2018.

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

[9] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa and M. García-Borroto, "Study of the impact of resampling methods for contrast pattern-based classifiers in imbalanced databases," Neurocomputing, 175, pp. 935-947, 2016.

[10] [A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," 2016, arXiv preprint arXiv:1608.06048

[11] P. Kaur and A. Gosain, "Comparing the Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise," In ICT Based Innovations, pp. 23-30, Springer, Singapore, 2018.

[12] F. Alam, E. A. Stepanov and G. Riccardi, "Personality traits recognition on social network-facebook," WCPR (ICWSM-13), Cambridge, MA, USA, 2013

[13] S. Chaudhary, R. Sing, S. T. Hasan and I. Kaur, "A comparative Study of Different Classifiers for Myers-Brigg Personality Prediction Model," IRJET, vol.05, pp.1410-1413, 2018.

[14] B. Cui and C. Qi, "Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction".

[15] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao et al, "Deep learning-based personality recognition from text posts of online social networks," *Applied Intelligence*, vol. 48, no. 11, pp. 4232-4246, 2018

[16] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," 2015 International Conference on Data and Software Engineering (ICoDSE), Yogyakarta, 2015, pp. 170-174.