Proceedings of the International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT 2023)
IEEE Xplore Part Number: CFP23CV1-ART; ISBN: 978-1-6654-7451-1

Framework for Implementation of Personality Inventory Model on Natural Language Processing with Personality Traits Analysis

*Dr. P. William
Department of Information Technology
Sanjivani College of Engineering,
Savitribai Phule Pune University,
Pune, India
william160891@gmail.com

Yogeesh N
Depatment of Mathematics
Government First Grade College,Tumkur
Karnataka, India
yogeesh.r@gmail.com

Dr. Vishal M Tidake
Depatment of MBA
Sanjivani College of Engineering,
Savitribai Phule Pune University,
Pune, India
tidkevishal@gmail.com

Snehal Sumit Gondkar
Department of Electrical Engineering
Sanjivani College of Engineering,
Savitribai Phule Pune University,
Pune, India
gondkarssnehal@gmail.com

Chetana. R
Department of Mathematics
Siddaganga, Institute of Technology,Tumkur
Karnataka, India
cr@sit.ac.in

Dr. K. Vengatesan
Department of Computer Engineering,
Sanjivani College of Engineering,
Savitribai Phule Pune University,
Pune, India
vengicse2005@gmail.com

***Abstract*—The phrase "personality" refers to an individual's distinct mode of thought, action, and behaviour Personality is a collection of feelings, thoughts, and aspirations that may be seen in the way people interact with one another. Behavioural features that separate one person from another and may be clearly seen when interacting with individuals in one's immediate surroundings and social group are included in this category of traits. To improve good healthy discourse, a variety of ways for evaluating candidate personalities based on the meaning of their textual message have been developed. According to the research, the textual content of interview responses to conventional interview questions is an effective measure for predicting a person's personality attribute. Nowadays, personality prediction has garnered considerable interest. It analyses user activity and displays their ideas, feelings, and so on. Historically, defining a personality trait was a laborious process. Thus, automated prediction is required for a big number of users. Different algorithms, data sources, and feature sets are used in various techniques. As a way to gauge someone's personality, personality prediction has evolved into an important topic of research in both psychology and computer science. Candidate personality traits may be classified using a word embedding model, which is the subject of this article.**

Keywords*: *Personality Prediction, Word Embedding Model, Machine Learning, Interview Answers, Personality Traits

I. INTRODUCTION

Personality is a term that describes a person's unique collection of qualities, such as their behaviour or feelings, that are influenced by environmental or biological factors. Reflects an individual's variety of thought and behaviour, as well as their feelings at the end of it. Constantly fluctuating traits of an individual's personality are expressed in their cumulative nature, rather than in discrete qualities. The Latin word "persona," which meaning "mask," is the origin of the English "personality".

Numerous studies have shown the importance of a person's personality in various kinds of relationships. Predicting work happiness, interpersonal performance, and preferences for various user interfaces have all been shown to be effective with this model. We are who we are because of the unique features we possess. Every individual has a distinct set of lasting characteristics and a particular style of relating to others and the environment. Personalities are widely accepted to be flexible, stable, and prone to change. [1-2]

The Big Five [2-6] and MBTI (Myers-Briggs Type Indicator) personality models [7-10] have been the most often used to predict personality. The DISC structure (Dominance, Influence, Compliance, and Stability) has been used by several researches [11-15].

Fig. 1 HEXACO Personality Inventory Model [10]

For the most part, the HEXACO model is used to depict people's personalities (Figure 1). A set of lexical tests led to the development of the HEXACO personality model, which has a six-factor framework. This model is also known as the OCEAN model (N). "Lexical Hypothesis" lies at the heart of the Big Five and HEXACO models. This hypothesis claims language contains our unique personality qualities and that

the models' key parameters may be derived by factoring together descriptions of human behaviour [16-18], is the model's unique sixth component. In comparison to the Big 5, the HEXACO model has been argued to be a better model since it takes into consideration a wide range of psychological phenomena, including gender differences in personality traits. H-factor traits include honesty, justice, integrity, and humility, all of which are important in the workplace. According to a wide range of studies, the H-factor may be used to identify and predict a wide range of workplace deviance and misconduct. Several beneficial characteristics have been linked to it, as well.

As a result of interview responses, it is possible to avoid long and contentious personality tests while still obtaining reliable results [16]. Personality-related themes have been hypothesised to be present in applicants' answers to queries on earlier behaviour and situational judgement. In order to make a forecast about someone's personality based on a given set of data, a person must perform the task of personality prediction. Predicting someone's personality using publicly available social media posts has been the subject of many studies to far. Using normal PRT approaches, researchers used status updates from online social networking accounts to project people' identities. The major objective of this research is to conduct a review of the many experiments reported where there are a variety of personality prediction systems available, and this article compares and contrasts them.

II. RELATED WORKS

Using Facebook interaction data, Tandera and colleagues [17-19] built a prediction engine that can estimate a user's personality. The Big Five personalities are used in this investigation. This analysis relies on a 150-customer dataset compiled by hand. The Facebook API Graph is used to obtain the data. After that, the user addresses are manually entered into the Apply Magic Sauce programme to customise the application.

Algorithms for personality prediction have been created and employed that are more specific and accurate than previous attempts. Inter-human contracts may be compared to this method. Big Five personality characteristics (three) and Myers-Briggs personality types (one) are used to classify the dataset's four unique corpora.

Weibo data and survey modes were analysed. Textual analysis of user input was used to determine their personality. Data was obtained using correlation and principal component analysis and then analysed using different correlation models [17].

It was shown that the Big Five Personality scores could be accurately predicted using Linear Regression (LR) and Support Vector Regression (SVR). Regression models were also compared to those based on the Linguistic Inquiry and Word Count (LIWC) tool for determining the qualities of respondents with Facebook statuses [15].

An n-grams bag, POS ID, and word vector were all used to collect data using this special language instrument. Characters' vocabulary usage was discovered via the assessment of these elements [20].

III. FRAMEWORK FOR PERSONALITY PREDICTION

The HEXACO Model's personality ratings are calculated using the methods described in this section. The main framework for personality prediction is shown in Figure 2.

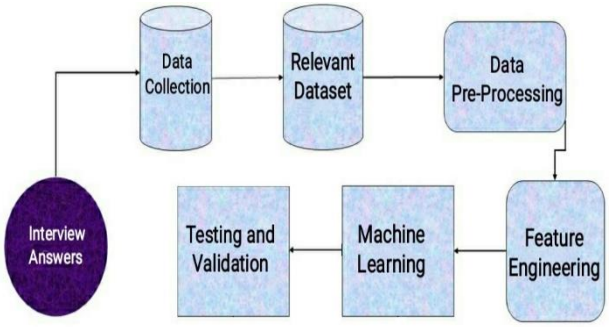

Fig. 2. General Framework for Personality Prediction

A. Data Collection

The data set may be created using the interview questions and responses from a survey of students. Interview answers from students will form the basis of the data sets (in the form of text). The answers of students will be collected and analysed in order to produce a database for the prediction of personality traits. The open-ended questions must be answered by each student.

B. Data Pre-processing

During this stage, the dataset will be pre-processed to remove any extra letters and words.

- *Hyperlink removal:* Using a regular expression, hyperlinks are removed since we are just dealing with textual material.
- *Emotions handing:* As a result of the emoticon text conversion, our training dataset now has a higher quality.
- *Unwanted character removal:* Punctuation, numerals, multiple spaces, and symbols that don't add anything to the message are discarded.
- *Stopword removal:* The nltk stopword module is used to remove stopwords from the dataset, such as a, for, and so on, which do not offer meaningful information during training.
- *Lemmatization:* Many inflected versions of one word might be grouped together as a "lemme," or "lemma group," since they all convey the same meaning.
- *Stemming:* Stemming is a rudimentary heuristic approach for removing affixes of derivation from phrases. A snowball stemmer was employed to accomplish stemming.

One-hot encoders may be used to encode the six HEXACO personality traits into a corpus matrix for classification tasks.

C. Feature Extraction

Raw or annotated text has to be transformed into features before it can be utilised by a machine learning model to better understand the meaning of the text. Fig. 3 depicts the current models in use.

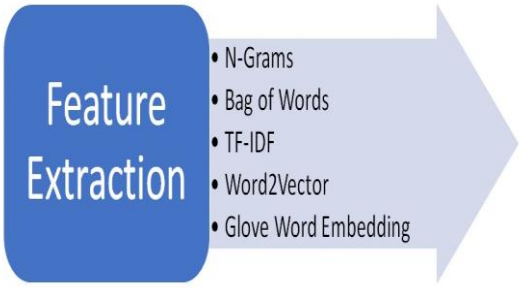

Fig. 3. Models for Feature Extraction

- *N-grams*: The term "n-gram" refers to a sentence consisting of n-item sequences. A variety of grammars may be given additional properties and semantic information can be extracted from the text.
- *Bag-of-Words (BOW):* It is possible to discover personality characteristics at the word level using the BOW technique, which extracts numerous aspects from words that are connected with distinct types of personalities.
- *Term Frequency and Inverse Document Frequency (TF-IDF):* Product of term frequency and document frequency is called TF-IDF, or term Frequency and Inverse Document Frequency, shortened. The lack of terminology for certain personality types is due to the fact that our dataset is skewed in favour of some personality types. As a result, low-level TF-IDF properties may be advantageously extracted.

D. Implementation of Word Embedding Model

Phrases that occur in the same context are represented identically by a word embedding, which is a kind of learned text representation [21, 22]. Each phrase is mapped to a specific vector space using word embeddings, which are real-valued matrices. There is just one vector for each word, and the vector values are taught similarly to neural networks, which is why the technique is often linked with the topic of deep learning. For each sentence, a real-valued vector is used to express it mathematically. Contrarily, the number of measurements required for sparse word representations such as a one-hot encoding might be in the hundreds of millions.

Term usage illustrates the distributed representation of the issue. As a result, statements repeated in similar contexts take on the same meaning, and their essence is instantly captured. This is in contrast to the bag of words paradigm, where each word has a distinct representation regardless of its use unless specifically addressed. Other applications, including document categorization, rely on unsupervised learning based on document data rather than a neural network model. NLP analyses word embeddings in combination with neural network models. Each word is encoded one at a time throughout the cleaning and preparation process. Vector space dimensions, such as 50, 100, or 300 are specified in the model. The vectors should be injected with tiny random numbers. Using the Backpropagation approach, an embedding layer is fitted to a neural network's front end.

Different matrix factorization algorithms such as LSA in word2vec may be combined with local context-based learning and GloVe (Global Vectors for Word Representation). By combining data from the whole text corpus with statistics from the frame, GloVe creates an abstract word context. An improved word embedding model was created as a consequence.

Consequently, the construction of more accurate word embeddings is made possible by the creation of a learning model. For unsupervised word representation learning, GloVe is a unique global log-bilinear regression model that outperforms current models in tasks such as word comparison, word similarity, and named item recognition.

E. Training

Supervised Machine Learning relies heavily on text classification. 80% of the data is used for training and 20% is used for testing on unknown data in an 80:20 split for training and testing reasons. Interview replies from a variety of pupils have been categorised. Following feature engineering, the data is categorised and processed using a number of machine learning approaches that take use of the features acquired.

F. Testing and Validation

A dataset comprising the previously trained model is required in order to analyse its performance. The confusion matrix, which comprises a matrix of correctly recognised classes and is particularly valuable for detecting if our model is under- or over-fitted to the testing dataset, was also emphasised as an important consideration in making the best model selection. Thus, throughout the testing and validation phases, each critical component was considered to ensure that the resultant model was the most efficient.

G. Web Development

Finally, a top-notch model was selected after completing all the previous steps, which included gathering and processing data and conducting critical model development, evaluation, and validation activities. We chose this one because of its speed and ease of use. The personality test was made easier to administer with the use of a web application. A great framework for building a web application that interacts easily with Python is Flask, which was utilised in this case.

IV. RESULTS AND DISCUSSIONS

Using textual information to predict a person's personality is an important study subject. Some research has been done, but there is still room for improvement in forecasting accuracy. Thus, several aspects of the approach must be enhanced, including the algorithms, feature extraction, and data collection. An algorithm was used that was capable of

analyzing text-based responses and generating the relevant results. It demonstrates how the terminology employed in routine interview replies reflects an individual's personality. Using a variety of data sets, the algorithm's validity was tested. Computers are now capable of making reliable decisions based on the user's personality type.

V. CONCLUSION

It was the goal of this study to examine the current body of knowledge on personality detection using online written text and to provide recommendations for future research in this area. Following the results of our research, we examined numerous methods or approaches for predicting personality types. Using a number of personality theories, the behaviour connected with applicants' replies may assist in forecasting their traits. Previously, questionnaires were employed, but this was a laborious process. An interviewer's personality may be predicted by their answers to interview questions, and this can be done successfully if they use a framework described in the research. The study reveals the various techniques and templates employed. By concentrating on possible routes, prediction accuracy may be increased while still accommodating certain tailored programmes and other suggestions.

VI. FUTURE ENHANCEMENT

Automatic character recognition from online textual information is attracting increased study interest due to its broad variety of computational applications. According on the findings of many experiments, there are a large number of possible directions for attitude prediction to go in. A custom-designed regression algorithm may be fed with a large number of linguistic features to increase the accuracy of personality assessment systems. Open-ended questions and eligible candidates should be added to the dataset's collection. Even the criminal justice system might use this strategy. Gender and age detection may be added to the personality analysis system that has been built.

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] P. William, A. Badholia, B. Patel and M. Nigam, "Hybrid Machine Learning Technique for Personality Classification from Online Text using HEXACO Model," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 253-259, doi: 10.1109/ICSCDS53736.2022.9760970.

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

[4] 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

[5] P William, Dr. Abhishek Badholia (2020) Evaluating Efficacy of Classification Algorithms on Personality Prediction Dataset. Elementary Education Online, 19 (4), 3400-3413. doi:10.17051/ilkonline.2020.04.764728

[6] 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.

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

[8] P. William and A. Badholia, "Analysis of Personality Traits from Text Based Answers using HEXACO Model," 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2021, pp. 1-10, doi: 10.1109/ICSES52305.2021.9633794.

[9] 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.

[10] P. William, A. Shrivastava, H. Chauhan, P. Nagpal, V. K. T. N and P. Singh, "Framework for Intelligent Smart City Deployment via Artificial Intelligence Software Networking," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), 2022, pp. 455-460, doi: 10.1109/ICIEM54221.2022.9853119.

[11] Pawar, A.B., Jawale, M.A., William, P., Sonawane, B.S. (2022). Efficacy of TCP/IP Over ATM Architecture Using Network Slicing in 5G Environment. In: Asokan, R., Ruiz, D.P., Baig, Z.A., Piramuthu, S. (eds) Smart Data Intelligence. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-19-3311-0_8

[12] Rawat, Romil and Yagya Nath Rimal, P. William, Snehil Dahima, Sonali Gupta, and K. Sakthidasan Sankaran. "Malware Threat Affecting Financial Organization Analysis Using Machine Learning Approach," *International Journal of Information Technology and Web Engineering (IJITWE)* 17, no.1: 1-20. http://doi.org/10.4018/IJITWE.304051

[13] P. William, A.B. Pawar, M.A. Jawale, Abhishek Badholia, Vijayant Verma, Energy efficient framework to implement next generation network protocol using ATM technology, Measurement: Sensors, Volume 24, 2022, 100477, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2022.100477.

[14] P William, Dr. Abhishek Badholia (2020) Evaluating Efficacy of Classification Algorithms on Personality Prediction Dataset. Elementary Education Online, 19 (4), 3400-3413. doi:10.17051/ilkonline.2020.04.764728

[15] 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.

[16] 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.

[17] P William, Dr. Abhishek Badholia 2021. Assessment of Personality from Interview Answers using Machine Learning Approach. International Journal of Advanced Science and Technology. 29, 08 (Jul. 2021), 6301-6312.

[18] P. William, Dr. Abhishek Badholia."A Review on Prediction of Personality Traits Considering Interview Answers with Personality Models", Volume 9, Issue V, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 1611-1616, ISSN : 2321-9653.

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

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

[21] K. Buraya, A. Farseev, A. Filchenkov and T. S. Chua, "Towards User Personality Profiling from Multiple Social Networks," In AAAI, pp. 4909-4910, 2017.

[22] Valanarasu, Mr R. "Comparative Analysis for Personality Prediction by Digital Footprints in Social Media." Journal of Information Technology 3, no. 02 (2021): 77- 91.