



## **A Cross-Bayesian Deep Learning Method for Text-Based Personality Trait Classification**

**Mrs. Revathi Pemmaraju**

Assistant Professor, Department of IT  
Sridevi Women's Engineering  
College, Telangana  
swecrevathi@gmail.com

**Duddilla. Ramya Sri**

BTech Student, Department of IT  
Sridevi Women's Engineering  
College, Telangana  
ramyasriduddilla03@gmail.com

**Seetha.Vijaya Lakshmi**

BTech Student, Department of IT  
Sridevi Women's Engineering  
College, Telangana  
seethavijaya2002@gmail.com

**Kuntamukkula. Roshitha**

BTech Student, Department of IT  
Sridevi Women's Engineering  
College, Telangana  
roshitha3006@gmail.com

**ABSTRACT**— Personality is an individual's unique combination of traits that determines how they think, feel, and behave. There is a chance to automatically identify a person's personality qualities from the text they provide on social networking sites. The purpose of this study is to utilize the XGBoost classifier, a machine learning technique, to predict four personality qualities from input text based on the Myers- Myers Type Indicator (MBTI) model: introversion, extroversion, intuition, and sensing-thinking, respectively. Experiments make use of the Kaggle benchmark dataset, which is freely accessible to the public. The key problem with the previous work is the bias of the dataset, which is reduced by using the Re-sampling approach, also known as random over-sampling, leading to improved performance. Pre-processing methods including tokenization, word stemming from, stop words deletion, and feature selection utilizing TF IDF are also used to get more insight about the author's character from the text. This research lays the groundwork for designing an individual identification system that might help businesses better understand their consumers and attract the most qualified employees. All classifiers provide respectable results when applied to the full set of personality characteristics, but the XGBoost classifier's performance stands out because to its ability to consistently achieve above 99% precision and accuracy.

---

### **INTRODUCTION**

Everything a person does is colored by their personality. Every aspect of one's day-to-day existence, from their feelings to their

preferences to their motivations to their physical well-being, may be affected by one's mindset [1].

The popularity of social media networks like Twitter and Facebook has encouraged people in the virtual world to open up and reveal their true selves via the expression of their thoughts, feelings, and emotions.

It's no secret that people's personalities reflect in the comments they leave or the tweets they send out on social media [2].

Recently, academics have been interested in automating the process of personality detection using social networking sites. The Big Five Factor Model [3], the MBTI (Myers-Briggs Type [4], and the DiSC Assessment [5] are just a few of the personality theories upon which the underlying concept of such apps is founded. Currently available methods for extracting a user's personality from their social media posts rely on supervised approaches to machine learning applied to a collection of benchmark datasets [6, 7, 8]. However, the main problem with these research is the skewness of the records, which means that there are uneven classes with regards to certain personality characteristics. This problem mostly affects the efficiency of facial recognition software.

The aforementioned problem may be addressed with the use of over-sampling, under-sampling, or a hybrid-sampling strategy [9].

Applying such methods to unbalanced datasets across domains has resulted in significant gains in recall, precision, recall, and F1-score [10].

This study use XGBoost, a machine learning approach, to identify text according to a variety of personality qualities, including extraversion and introversion, intuition and

sensing, emotion and logic, and judgment and perception, using the reference personality recognition dataset.

In addition, resampling method [11] is used to reduce skewness in the dataset and boost system performance.

#### A. Description of the Issue

Researchers are increasingly interested in determining a user's character based on their online discourse. Personality prediction using just the input text has been researched and developed enough [6, 7, 8].

However, additional effort is needed to enhance the efficiency of the current personality detection system, a problem that often emerges from the existence of unbalanced classes of personality characteristics. As part of the planned effort. Personality identification performance may be enhanced by using resampling, a method for balancing datasets.

#### Research Questions, Part B

To categorize character qualities from the provided text, RQ.1 asks how to implement a supervised machine learning algorithm, in this case the XGBoost classifier.

To what extent can performance be improved by using a class balancing strategy on unbalanced classes of personality characteristics, and how effective is the suggested methodology in comparison to existing machine learning methods?

Question 3: How does the suggested approach compare to existing standards in terms of its effectiveness?

#### Goals and Purposes (Part C)

1) Aim: The goal of this study is to use a supervised machine learning algorithm, the XGBoost classifier, on the MBTI

personality benchmark dataset in order to categorize a user's personality characteristics based on the input text. The improvements made here build upon those made in [6].

a) Recognizing personality characteristics from the provided text using a machine learning approach called XGBoost classifier.

b) Using a resampling approach on the unbalanced groups of personality characteristics to boost the suggested system's accuracy.

Comparing the effectiveness of the proposed model against that of other machine-learning strategies and baseline procedures is step c.

#### D. Importance of the Research

People have unique personalities because of the unique ways they think, act, and feel. A person's personality is a major factor in determining their tastes in media such as books, websites, songs, and films.

[12].

The suggested study on personality identification improves upon previous efforts in this area by [6]. The proposed work is important because (i) it addresses the inefficiency of the existing study's performance due to skewness by applying a re-sampling technique to the imbalanced dataset, and (ii) it lays the groundwork for the development of cutting-edge applications for personality recognition that could help businesses with things like hiring the right people and expanding.

#### **RELATED WORK**

##### **Use of Supervised Machine Learning in Recognizing Individuals**

One of the most common methods used by academics to delve into written and spoken material is the Unsupervised machine

learning methodology, often known to be the Corpus based method (CBA). Benefits include the ability to analyze word choice, frequency, collocation, and concordance, among others. One of the main limitations of such methods is that the Unsupervised or Library-Based approach (CBA) need an annotated text for classifier testing and training purposes. a database-free method for recognizing emotions using an inference system based on neurofuzzy logic. The suggested system is able to learn from and generalize to new situations with high efficiency. The experimental results demonstrate the effectiveness of the suggested method in comparison to SVM.

##### **Using unattended Machine Learning for Recognizing Individual Characteristics**

This method is used when determining the veracity of the annotated data presents a challenge. Keyword lists pertaining to the various classes are used to accomplish the categorization of input text. In the case of examining domain-dependent data, the unsupervised method is simple to implement. Following is a summary of some of the research done on various machine learning methods. developed a method to extract and categorize user personality qualities from social media sites such Friend Feed utilizing an unsupervised technique and five characteristics of personality (Big Five). A personality model was developed by mining a variety of linguistic traits thought to be associated with character. The technology is able to successfully calculate personality ratings from a given text. However, the system's lack of attention to user-system interaction was glaring. The

most influential groups in a Twitter network graph are identified using a system called Twitter Personality based Important Group Extraction (T-PICE). By combining data showing additional facets of user behavior based on the use of machine learning techniques with the previous methods, they were able to identify users' personality characteristics. A pre-processing phase that eliminates network edges based on individuals' personalities is included into a pre-existing modularity-based community discovery method to achieve this goal. However, while thinking about a massive graph, scalability issues must be resolved.

### **Hybrid and Semi-Supervised Methods for Recognizing Individual Characteristics**

The characteristics of supervised methods and lexicon-based methods that use labeled data records are incorporated into semi-supervised and hybrid procedures. The following works are relevant to the topic of hybrid methods: created an algorithm that uses still photos with a focus on neutral expressions to compare the personalities of humans and chimps. The findings demonstrate that human beings have a more refined sense of perception than chimps. To learn more about the similarities and differences between humans and chimpanzees, additional research is needed.

### **Using Deep Learning Methods for Recognizing People**

The algorithms that make up deep learning are meant to be emulations of the brain in both form and function. In layman's words, it has neurons that take in information and neurons that send out information. The recognition of speech, computer vision,

NLP, and handwriting creation are just few of the areas where deep learning-based models might be useful. An 8-fold reduction in data need model for predicting the Big Five human personality characteristics has been developed. Underlying tweets posted by users is an embedding layer called GloVe, which is utilized for word extraction. Using the supplied Twitter data, we train and evaluate the model. Moving ahead, data is tested across three fusions: (i) LIWC + GP, (ii) 3-Gram + GP, & (iii) GloVe + RR. The suggested model had a higher mean correlation (0.33 over the Big Five attributes) than the state-of-the-art did. The current effort only used Twitter material in English; this could be expanded to include other languages. In addition, a larger sample of tweets may be used to assess the effectiveness of the suggested model. Using a network of deep neural networks, as suggested in [10], a method for identifying personalities is applied to textual data. AttRCNN, a hierarchical system utilized in this study, can remember semantic characteristics at a more fundamental level than previous methods. As shown in the data, the suggested characteristics significantly outperform their counterparts.

### **METHODOLOGY**

This article proposes using a supervised learning method for character profiling. The model will take a text or post fragment as input and use it to make predictions about the writer's personality based on the words it finds. Classification and forecasting are two of the many applications of the Mayers-Briggs Type Indicator [4]. Based on these

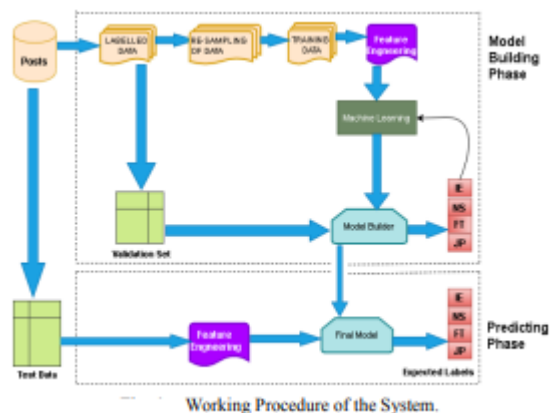
four factors, this model classifies people into one of sixteen distinct personality types:

(i) Attitude—Extraversion vs. Introversion: This dimension characterizes the source of an individual's motivation, whether it comes from the approval of others' opinions and assessments or from inside themselves. (ii) Information Sensing versus iNtuition (S/N): this shows how people receive information and observant(S), using their five senses and careful observation, whereas intuitive types value originality and changeability above steadfastness and reason. (iii) Thinking vs. Feeling (T/F): Someone with a strong Thinking trait will always act in a rational manner when making a choice, whereas those with a strong Feeling trait will put their empathy and emotional needs ahead of reasoning. (iv) Methodology The Judging vs. Perceiving (J/P) dichotomy explains how a person goes about their day-to-day tasks, decisions, and long-term plans. Judgmental people have very well-organized minds.

They're not big fans of winging it. People that are perceptive are naturally intuitive and unplanned. They are skilled at improvising and keeping their alternatives open [40].

#### D. The Predictive System for Characteristics of Individuals' Characteristics

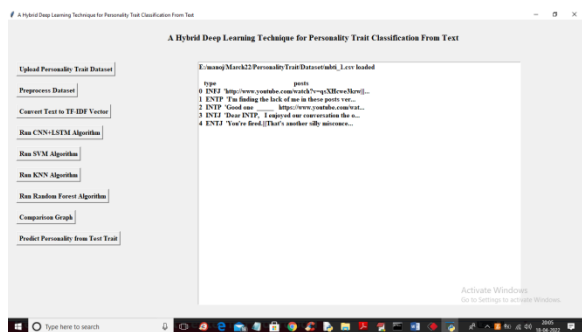
First, the model that is suggested is trained with both data with labels (MBTI type) and unlabeled data (tweets), as shown in Fig. 4. The effectiveness of the trained model is next assessed. The information set will be divided into a training phase, a validating phase, and a testing phase so that predictions may be improved. Data overfitting is mitigated through the validation process.



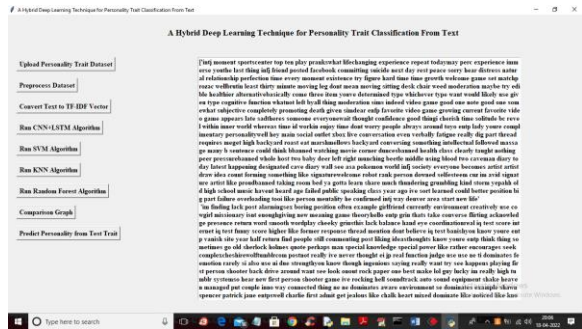
## RESULT AND DISCUSSION

Author uses social media material to categorize personality traits including introversion and extroversion, intuition and sensing, thought and emotion, and judgment and perception in proposed article. As a result of analyzing postings written by humans, we may learn about their character quirks. Unfortunately, the classification accuracy of the numerous recent machine learning algorithms that are trained on datasets for personality trait categorization is low.

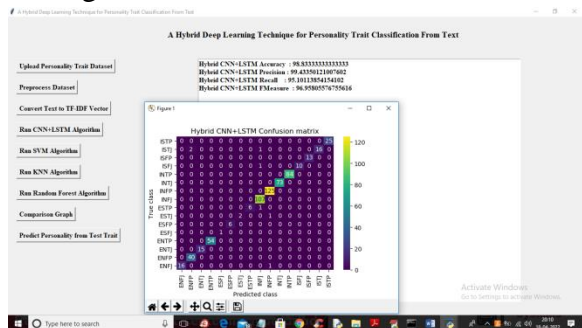
The author proposes a solution by combining two algorithms for deep learning into a single one; they term it Hybrid CNN+LSTM. The proposed hybrid approach first trains a CNN model on the dataset, then extracts features from the learned CNN model and retrains the LSTM algorithm on those features to improve the LSTM's prediction accuracy.



Input data loaded



create a TF-IDF matrix using the above post messages



seen that we achieved 98% accuracy with CNN+LSTM



On the x-axis are the names of the algorithms, and on the y-axis are the various

metrics they achieved, with the proposed Hybrid CNN+LSTM achieving the highest accuracy.

CONCLUSION

In this work, we investigated the task of personality trait classification from textual content. To accomplish the research task, we proposed applying a deep learning model, namely CNN+LSTM. The proposed study includes the following modules: (i) acquiring data, (ii) pre-processing of data, and (iii) implementing the deep neural network. The proposed CNN+LSTM model for personality trait classification is a merger of CNN and LSTM that assists in classifying the input text into different personality traits like I-E, N-S, T-F, and J-P. The main emphasis of the CNN model is to extract and retain the local features using a convolutional and max-pooling layer. CNN acts as a robust tool for choosing the best features that enhance the prediction accuracy. The LSTM model preserves the prior information regarding context, which helps to exploit significant context information at the start of a sentence. Its benefit is that it takes sequential information through the examination of prior data. After receiving the final representation of an input sentence, it is classified among the different personality traits. The experiments with different machine learning and deep learning models are also conducted and their results are recorded on the personality trait dataset. The results show that the proposed CNN+LSTM model for personality trait classification produced improved results in terms of improved accuracy (88% for I-E, 2509

91% for N-S, 85% for T-F, and 80% for J-P, precision (88% for I-E, 91% for N-S, 85% for T-F, and 80% for J-P, and f1-score (88% for I-E, 91% for N-S, 85% for T-F, and 80% for J-P, and the proposed CNN+LSTM model for personality trait classification (88% for The information obtained from this research acts as best practices for the selection, management, and optimization of their policies, services, and products. A. LIMITATIONS The possible limitations of the proposed work are given as follows: (i) In this study, we performed a personality trait classification on limited personality traits pertaining to only textual content in the English language. (ii) The work is limited to random word embedding without the exploitation of several word representation models like Glove, Fasttext, and word2vec, (iii) The attention mechanism was not introduced with CNN+LSTM that assists in extracting relative significant features, (iv) Other combinations of deep learning models like CNN+Bi-LSTM, CNN+GRU, Bi-LSTM+CNN, and CNN+RNN are not applied for personality trait classification. (v) The focus of the current research content is only on the MBTI dataset for personality trait classification. (vi) The limited number of machine learning classifiers is used for experimentation that needs to be further extended with other machine learning classifiers. B. FUTURE DIRECTIONS The possible future directions of the work are as follows: (i) The work can be extended by conducting experiments in other languages on an extended set of personality traits using non-textual content such as images and videos, (ii) In future work, we can exploit

other pre-trained word representation schemes like Glove, word2vec, and Fasttext regarding word embedding layer, (iii) Introducing an attention mechanism regarding personality classification may enhance the system's performance, (iv) We will explore different combinations of deep neural networks like CNN+Bi-LSTM, CNN+GRU, Bi-LSTM+CNN, and CNN+RNN for personality trait classification, and CNN+Bi-LSTM+CNN for facial recognition. (v) In addition to the MBTI dataset, other different datasets regarding personality trait classification tasks can be exploited. It is planned to collect additional data and apply the proposed methods to some larger corpora in order to test the effectiveness of the proposed approaches on massive data records, (vi) The combination of other deep neural network frameworks will better handle the problem of personality trait classification tasks. Thus, in the future, we will apply other neural networks, (vii) In the future, we will focus on applying ensemble methods to enhance system performance, and (viii) We will work on including further base models and by searching for other parameters that may assist in enhancing the overall accuracy of the proposed work.

## **REFERENCES**

- [1] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and S. Khan, "Classification of poetry text into the emotional states using deep learning technique," *IEEE Access*, vol. 8, pp. 73865–73878, 2020.
- [2] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection

from text,” *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 74–79, Mar. 2017.

[3] W. A. H. W. Azizan, A. A. A. Rahim, S. L. M. Hassan, I. S. A. Halim, and N. E. Abdullah, “A comparative study of two machine learning algorithms for heart disease prediction system,” in *Proc. IEEE 12th Control Syst. Graduate Res. Colloq. (ICSGRC)*, Shah Alam, Malaysia, Aug. 2021, pp. 132–137, doi: 10.1109/ICSGRC53186.2021.9515250.

[4] A. Hassan and A. Mahmood, “Convolutional recurrent deep learning model for sentence classification,” *IEEE Access*, vol. 6, pp. 13949–13957, 2018.

[5] N. Chen and P. Wang, “Advanced combined LSTM-CNN model for Twitter sentiment analysis,” in *Proc. 5th IEEE Int. Conf. Cloud Comput. Intell. Syst. (CCIS)*, Nov. 2018, pp. 684–687.

[6] A. S. Khan, H. Ahmad, M. Zubair, F. Khan, A. Arif, and H. Ali, “Personality classification from online text using machine learning approach,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, pp. 1–5, 2020, doi: 10.14569/IJACSA.2020.0110358.

[7] L. Liu, D. Preotiuc-Pietro, Z. R. Samani, and M. E. Ungar, “Analyzing personality through social media profile picture choice,” in *Proc. Int. AAAI Conf. Social Media (ICWSM)*, 2016, pp. 211–220.

[8] All things Statista. Number of Monthly Active Twitter Users Worldwide From 1st Quarter 2010 to 1st Quarter 2019 Retrieved From. Accessed: Jan. 18, 2020. [Online]. Available: <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>

[9] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, “Deep learning-based document modeling for personality detection from text,” *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 74–79, Mar. 2017.

[10] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao, Z. Wu, X. Zhong, and J. Sun, “Deep learning-based personality recognition from text posts of online social networks,” *Int. J. Speech Technol.*, vol. 48, no. 11, pp. 4232–4246, Nov. 2018.

[11] M. Osama and S. R. El-Beltagy, “A transfer learning approach for emotion intensity prediction in microblog text,” in *Proc. Int. Conf. Adv. Intell. Syst. Inform. Cham, Switzerland: Springer*, Oct. 2019 pp. 512–522.

[12] A. Khattak, M. Z. Asghar, Z. Ishaq, W. H. Bangyal, and I. A. Hameed, “Enhanced concept-level sentiment analysis system with expanded ontological relations for efficient classification of user reviews,” *Egyptian Informat. J.*, vol. 3, pp. 1–17, Apr. 2021, doi: 10.1016/j.eij.2021.03.001.

[13] M. Z. Asghar, A. Sattar, A. Khan, A. Ali, F. Masud Kundi, and S. Ahmad, “Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language,” *Expert Syst.*, vol. 36, no. 3, Jun. 2019, Art. no. e12397.

[14] F. M. Alotaibi, M. Z. Asghar, and S. Ahmad, “A hybrid CNN-LSTM model for psychopathic class detection from tweeter users,” *Cognit. Comput.*, vol. 13, no. 3, pp. 709–723, Mar. 2021.

[15] A. Khattak, A. Habib, M. Z. Asghar, F. Subhan, and I. A. R. Habib, “Applying deep neural networks for user intention



identification,” *Soft Comput.*, vol. 25, no. 3,  
pp. 2191–2220, Feb. 2021.