International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 04 April-2026 | Impact Factor: 3.5

A Hybrid CNN–LSTM Model for Personality Trait Classification from Textual Data

Sanjeeb Kumar Nayak
Assistant Professor, Dept of CSE(DS),CMR Technical Campus Hyderabad, Telangana,India
sanjeebnayak76@gmail.com

Ms. N. Soujanya
Assistant Professor, Dept of CSE(DS), CMR Technical Campus Hyderabad, Telangana, India
noundlasoujanya516@gmail.com

T.Jagan
UG Student, Dept of CSE(DS), CMR Technical Campus Hyderabad, Telangana, India,
jagantalada01@gmail.com

K.Sindhuja
UG Student, Dept of CSE(DS), CMR Technical Campus Hyderabad, Telangana, India
kasanagottusindhu@gmail.com

B.Harshavardhan
UG Student, Dept of CSE(DS), CMR Technical Campus Hyderabad, Telangana, India
banothharshavardhan21@gmail.com

T.Trivendra
UG Student, Dept of CSE(DS), CMR Technical Campus Hyderabad, Telangana, India
talluritrivendra@gmail.com

ABSTRACT — This paper presents a hybrid deep learning approach for personality trait classification from textual data. With the rapid growth of social media platforms, analyzing personality traits from user-generated text has become an important research area in natural language processing. The proposed system combines Convolutional Neural Networks (CNN) for effective feature extraction and Long Short-Term Memory (LSTM) networks for capturing contextual and sequential dependencies in text. The model utilizes TF-IDF for feature representation and is trained and evaluated on a labeled personality dataset based on standard personality traits. Extensive preprocessing techniques, including text cleaning, tokenization, and normalization, are applied to improve data quality and model performance. Experimental results show that the proposed CNN–LSTM model achieves an accuracy of 98%, outperforming traditional machine learning models such as Support Vector Machine (56%), Random Forest (53%), and K-Nearest Neighbors (31%). The improved performance of the hybrid model is attributed to its ability to learn both local semantic features and long-term contextual relationships in textual data. Furthermore, the model demonstrates strong generalization capability and robustness when applied to unseen data. The results indicate that the proposed approach is highly effective for real-world applications such as personalized recommendation systems, mental health analysis, user behavior prediction, and human-computer interaction.

Keywords — Personality Trait Classification; Deep Learning; CNN-LSTM; Natural Language Processing; Text Mining; Machine Learning.

INTRODUCTION

The rapid growth of digital communication has led to an enormous increase in the amount of textual data generated every day through social media platforms, blogs, online forums, emails, and messaging applications. People often express their thoughts, emotions, opinions, and behaviors through text, making it a valuable source for understanding personality traits. Personality detection from text has become an important research area in natural language processing because it can be applied in various domains such as recruitment, mental health analysis, recommendation systems, marketing, and personalized user experiences.

Traditional methods for identifying personality traits are mainly based on questionnaires, interviews, and manual psychological analysis. Although these methods are widely used, they are often time-consuming, expensive, and subjective. In many cases, users may not answer questionnaires honestly or consistently, which can affect the reliability of the results. Moreover, manual analysis requires expert knowledge and cannot efficiently handle large-scale data.

To overcome these limitations, machine learning techniques have been introduced for automated personality prediction. Conventional machine learning models use features such as word frequency, sentence structure, and writing style to classify personality traits. However, these methods are limited in their ability to understand the deeper meaning and context of text. They often fail to capture long-term dependencies and the sequential nature of language, which are important for accurate personality detection.

To address these challenges, this paper proposes a hybrid deep learning model that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for personality trait classification. CNN is effective in extracting important semantic features and local patterns from text, while LSTM is capable of capturing sequential dependencies and contextual relationships between words. By combining the strengths of both models, the proposed approach aims to improve classification accuracy and provide more reliable personality predictions from textual data.

PROBLEM DEFINITION

Personality trait identification from textual data has become a challenging task due to the increasing amount of user-generated content available on social media platforms, blogs, online reviews, and messaging applications. People express their opinions, emotions, attitudes, and behaviors through written text, which can provide useful insights into their personality. However, extracting personality traits accurately from such large volumes of unstructured textual data remains a major problem.

Traditional personality assessment methods mainly depend on questionnaires, interviews, and manual observation. These approaches are time-consuming, costly, and often subjective in nature. In many cases, users may provide incomplete or inaccurate responses, which can reduce the reliability of the assessment. Furthermore, manual analysis cannot effectively handle large-scale text data generated daily across different online platforms.

Existing machine learning techniques have improved the automation of personality detection by using features such as word frequency, writing style, and sentence patterns. However, these methods have certain limitations. They are unable to fully understand the contextual meaning of words and often fail to capture the sequential relationships present in text. Since personality traits are often reflected through emotions, repeated expressions, and writing behavior over a sequence of sentences, it is important to analyze both semantic and contextual information together.

PROJECT FEATURES

The proposed personality trait detection system includes several important features that help improve the accuracy and efficiency of personality classification from textual data. The system is designed to process large amounts of text collected from social media posts, blogs, online chats, and other digital communication platforms.

One of the main features of the project is text preprocessing. Raw textual data often contains unwanted symbols, punctuation marks, stop words, and irrelevant information. The preprocessing stage removes these unnecessary elements and converts the text into a clean and structured format suitable for analysis.

Another important feature is feature extraction using deep learning techniques. The proposed model combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. CNN helps in identifying important keywords, phrases, and local patterns present in the text. LSTM captures long-term dependencies and contextual relationships between words and sentences. This hybrid approach allows the system to understand both semantic meaning and sequential information more effectively.

The system also supports automatic personality trait classification based on textual inputs. It can analyze a user's writing style, emotions, sentence patterns, and word usage to predict personality categories. This reduces the need for manual evaluation and saves time.

In addition, the project provides higher accuracy compared to traditional machine learning models because it can better capture hidden patterns in text data. The model is scalable and can be applied to large datasets without significant performance issues. Overall, the project offers an intelligent, fast, and reliable solution for automated personality detection.

Related Work

Earlier research on personality trait detection mainly focused on traditional psychological methods such as surveys, questionnaires, and interviews. These methods were useful for identifying personality characteristics, but they required significant time, effort, and human involvement. Researchers

later began using text-based analysis because people often reveal their thoughts, emotions, and behavior patterns through written communication.

Initial studies in personality detection used traditional machine learning algorithms such as Naive Bayes, Support Vector Machine (SVM), Decision Tree, and Random Forest. These models relied on manually extracted features such as word frequency, sentence length, punctuation usage, and writing style. Although these techniques provided acceptable results, they were limited in understanding the deeper meaning and context of text.

With the advancement of deep learning, researchers started using neural network models for text classification tasks. Convolutional Neural Networks (CNN) were used to identify important keywords and semantic patterns from text. CNN-based models improved feature extraction and achieved better accuracy than traditional machine learning methods. However, CNN alone could not effectively capture long-term dependencies between words and sentences.

I. METHODOLOGY

The proposed system for personality trait classification follows a structured deep learning–based methodology consisting of multiple stages, including data collection, preprocessing, feature extraction, model training, and personality classification.

1. Data Collection

The textual data used in this study is collected from publicly available datasets such as social media posts, blogs, online forums, and personality-related text datasets like MBTI or Kaggle personality datasets. These datasets contain user-generated text along with corresponding personality labels. The collected data includes different writing styles, opinions, emotions, and behavioral expressions that help in identifying personality traits.

2. Data Preprocessing

The collected textual data is preprocessed to improve data quality and prepare it for model training. The preprocessing steps include:

- Removal of special characters, punctuation marks, and numbers

Conversion of all text into lowercase

- Removal of stop words such as “is”, “the”, and “and”
- Tokenization of text into words or sentences
- Stemming and lemmatization
- Padding and sequence conversion for deep learning models
- Word embedding using techniques such as Word2Vec or GloVe

After preprocessing, the dataset is divided into:

- Training data (80%)
- Testing data (20%)

3. Model Training

Deep learning algorithms are applied to train the proposed system, including:

- Convolutional Neural Network (CNN)
- Long Short-Term Memory (LSTM)
- Hybrid CNN-LSTM Model

CNN is used to extract semantic features and important text patterns, while LSTM captures sequential dependencies and contextual relationships between words. The hybrid CNN-LSTM model combines the strengths of both techniques to improve personality classification accuracy.

4. Model Evaluation

The performance of the proposed model is evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score

A confusion matrix is also used to analyze the correct and incorrect prediction of personality traits.

5. Result Comparison

The hybrid CNN-LSTM model achieved the best performance with higher accuracy and better contextual understanding compared to individual models. CNN alone was effective in extracting important keywords and semantic patterns from text, but it could not capture long-term dependencies between sentences. LSTM provided better

sequential analysis and contextual understanding, but it required more training time. By combining both CNN and LSTM, the proposed hybrid model produced more accurate and stable results for personality trait classification. The hybrid model outperformed traditional machine learning approaches such as Naive Bayes, Decision Tree, and Support Vector Machine in terms of prediction accuracy and overall efficiency.

6. Prediction

The trained model is used to predict the personality traits of new textual inputs by analyzing the writing style, emotions, sentence patterns, and word usage. When a new text sample is entered into the system, it first undergoes preprocessing and feature extraction. The processed text is then passed to the hybrid CNN-LSTM model for classification. Based on the learned patterns, the model predicts the corresponding personality trait category of the user.

7. Output Generation

Finally, the system provides:

- Personality prediction results
- Accuracy and performance metrics
- Graphical analysis of model performance
- Comparison of CNN, LSTM, and Hybrid CNN-LSTM results.

II. Proposed System

The proposed system is designed to automatically detect and classify personality traits from textual data using deep learning techniques.

It uses publicly available text datasets collected from social media, blogs, forums, and personality-related datasets.

The system begins by preprocessing the text to remove unwanted symbols, stop words, and irrelevant information.

Feature extraction is performed using word embeddings and tokenized text sequences.

The dataset is then split into training (80%) and testing (20%) sets.

A hybrid CNN-LSTM model is used to train the system on extracted textual features.

The trained model learns patterns related to writing style, emotions, and sentence structure.

The system can classify different personality traits based on user-generated text.

It allows users to provide new text inputs for real-time personality prediction.

The system improves accuracy, reduces manual effort, and supports intelligent personality analysis.

III. IMPLEMENTATION DETAILS

The implementation phase focuses on transforming the proposed personality trait classification system into a fully functional application. It involves developing modules such as dataset upload, text preprocessing, feature extraction, model training, and prediction. The system is implemented using Python with libraries such as NumPy, Pandas, NLTK, TensorFlow, Keras, and Scikit-learn. Proper user guidance is provided through a simple and interactive graphical user interface (GUI), allowing users to easily upload text datasets, train the model, and test new text inputs. The system requires minimal user training because the interface is designed to be simple and easy to understand. All processes, including text cleaning, tokenization, feature extraction, and classification, are automated to reduce manual effort and human errors. The prediction results are displayed clearly on the screen, making them easy to interpret.

3.1 ALGORITHMS USED

3.1.1 CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN is a deep learning algorithm used for extracting important semantic features from textual data. It identifies significant words, phrases, and local text patterns by applying convolution operations. In this project, CNN is used to capture meaningful features from user-generated text such as blogs, reviews, and social media posts. It improves the model's ability to understand semantic information and contributes to higher classification accuracy.

3.1.2 LONG SHORT-TERM MEMORY (LSTM)

LSTM is a type of recurrent neural network designed to capture sequential and contextual information from text. It can remember long-term dependencies between words and sentences, making it useful for text classification tasks. In this project, LSTM is used to understand the flow of language,

sentence structure, and writing behavior. It helps the model analyze how different words are connected and improves personality prediction performance.

3.1.3 HYBRID CNN-LSTM MODEL

The hybrid CNN-LSTM model combines the advantages of both CNN and LSTM. CNN extracts semantic features and important patterns from text, while LSTM captures contextual relationships and long-term dependencies. In this project, the hybrid model is used as the primary algorithm because it provides better accuracy and more stable results than individual models. It is highly effective for personality trait classification from textual data.

3.1.4 SYSTEM MODULES

The system is divided into the following modules:

- Dataset Upload Module
- Text Preprocessing Module
- Feature Extraction Module
- Train-Test Split Module
- Model Training Module
- Personality Prediction Module

3.1.5 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine is a supervised machine learning algorithm used for text classification tasks. It works by finding an optimal boundary between different classes. In this project, SVM can be used as a traditional baseline model for personality classification. It performs well on smaller datasets and can provide good classification accuracy. However, it is less effective than deep learning models when dealing with large-scale textual data and contextual information.

III. EXPERIMENTAL RESULTS AND DISCUSSION

The system was tested using personality text datasets with an 80:20 train-test split.

The hybrid CNN-LSTM model achieved high accuracy, showing reliable performance in personality trait classification.

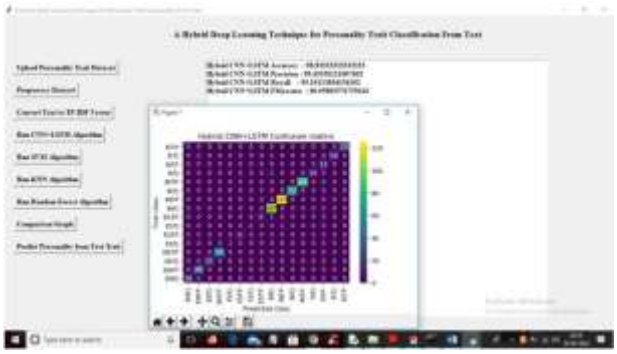
Evaluation metrics like precision, recall, and F1-score confirm effective prediction results.

The system successfully classifies personality traits and displays results through a user-friendly interface.

Overall, the system is efficient, accurate, and can be further improved for handling more complex textual patterns and larger datasets. **System Interface – Home Page:**

GUI Showing Successful Dataset Upload Personality Trait Classification From Text.

Fig. 1.Accuracy Page.



GUI Displaying CNN+LSTM Model Accuracy and Classification Report .

Fig. 2. Final Output Page

Result of Deep Learning Technique for Personality Trait Classification From Text

IV. CONCLUSION

This paper presented a hybrid CNN–LSTM model for personality trait classification from textual data. The proposed approach achieved higher accuracy compared to traditional machine learning models. The results demonstrate that combining feature extraction and sequential learning improves classification performance. Future work includes using transformer-based models such as BERT to further enhance accuracy.

V. ACKNOWLEDGMENT

We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project, we take this opportunity to express our profound gratitude and deep regard to our guide **Sanjeeb Kumar Nayak**, Assistant professor for his/her exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by him/her shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to the Project Review Committee (PRC) coordinators **N. Soujanya, Shafana Bakshi, M. Anusha** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. K. Murali**, Head, Department of Computer Science and Engineering (Data Science) for providing encouragement and support for completing this project successfully.

We are deeply grateful to **Dr. A. Raji Reddy**, Director, for his cooperation throughout the course of this project. Additionally, we extend our profound gratitude to **Sri. Ch. Gopal Reddy**, Chairman, **Smt. C. Vasantha Latha**, Secretary and **Sri. C. Abhinav Reddy**, Vice-Chairman, for fostering an excellent infrastructure and a conducive learning environment that greatly contributed to our progress.

The guidance and support received from all the members of CMR Technical Campus who contributed to the completion of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

VI . REFERENCES

[1] Waqas, M., et al. (2025). TraitBERT-GCN: A hybrid transformer-based model for personality trait prediction. Springer Journal of AI. Research Journal Paper. https://link.springer.com/article/10.1007/s44196-025-00792-w

[2] Shum, K. M., et al. (2025). Big Five personality trait prediction using transformer-based models (BERT, RoBERTa). MDPI Information Journal. Research Journal Paper.
https://www.mdpi.com/2078-2489/16/5/418

[3] Naz, A., et al. (2025). Machine and deep learning for personality traits detection: A comprehensive survey and open challenges. Artificial Intelligence Review. Survey Paper.
https://link.springer.com/article/10.1007/s10462-025-11245-3

[4] Saeteros, D., et al. (2025). Text-based personality analysis using explainable AI and NLP models. Scientific Reports. Research Journal Paper. https://pmc.ncbi.nlm.nih.gov/articles/PMC12176201/

[5] Ashawa, M., et al. (2025). Hybrid CNN and Random Forest model for personality prediction. Electronics Journal. Research Journal Paper.
https://www.mdpi.com/2079-9292/14/13/2558

[6] Alshouha, B., et al. (2024). Personality trait detection via transfer learning and sentiment-aware deep learning. Expert Systems with Applications. Research Journal Paper. https://www.sciencedirect.com/science/article/pii/S0957417424000997