IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 September 2024, accepted 22 October 2024, date of publication 25 October 2024, date of current version 6 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3486578

RESEARCH ARTICLE

AI Knows You: Deep Learning Model for Prediction of Extroversion Personality Trait

ANAM NAZ[1], HIKMAT ULLAH KHAN[2], SAMI ALESAWI[3], OMAR IBRAHIM ABOUOLA[4], ALI DAUD[5], AND MUHAMMAD RAMZAN[6]

[1]Department of Computer Science, University of Sargodha, Punjab 40162, Pakistan
[2]Department of Information Technology, University of Sargodha, Punjab 40162, Pakistan
[3]Department of Computer Science, King Abdulaziz University, Rabigh 21911, Saudi Arabia
[4]Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah 21959, Saudi Arabia
[5]Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates
[6]Department of Software Engineering, University of Sargodha, Punjab 40162, Pakistan

Corresponding authors: Ali Daud (alimsdb@gmail.com) and Hikmat Ullah Khan (dr.hikmat.niazi@gmail.com)

ABSTRACT The recent rise of Artificial Intelligence (AI) has already revolutionized human lives and is improving the quality of human life in many ways. The field of AI, Natural Language Processing (NLP), helps to understand, comprehend and even generate new content. NLP is used for various content analysis tasks such as sentiment analysis, fake news detection, etc. However, human personality traits detection is a new research domain. Analyzing users generated content has a significant role in identifying and understanding users' views and behaviors. Users' traits detection can be helpful in analysis of consumers' personalization, finding top candidates for recruitment, career counselling, etc. In this research study, our aim is to predict personality of extroversion behaviors using machine and deep learning approaches. Extroversion means whether a person is an introvert or extrovert as this trait is relevant to certain jobs like team management, social ties etc. For empirical analysis, we investigate MBTI dataset with various feature engineering techniques including textual features like Term Frequency-Inverse Document Frequency (TF-IDF), Parts of Speech (PoS) tagging, as well as deep word embeddings ok word2vec, GloVe. The state-of-the-art shallow machine learning, ensemble modelling and deep learning models are applied. The main novelty is the exploration of latest sentence embeddings which captures semantic information of content in a better manner. Thus, the comprehensive results analysis reveals sentence embeddings as features to Bi-LSTM achieves highest accuracy of 92.52% and outperforms the existing studies in the relevant literature.

INDEX TERMS Cognitive science, deep embeddings, deep learning, feature engineering, machine learning, psychology.

I. INTRODUCTION

In the era of social media, where users post huge volume of textual content on daily basis, from status updates to posting comments and tweets. The world has transformed into a social web where users are engaging in users' generated content (UGC) on a regular basis in this digital age. Through posts, interactions, and content patterns including text, photos, audios, and videos, users reveal their attributes of behaviors in front of social platforms [1]. This significance of social mining not only highlights the technological advancements but also the fundamental relationship between users' attitudes, personalities, and growth of the social web [2]. Social media platforms are a central focus of users rely on 4C: collaboration, communication, and community engagement with others, constructing their personalities, to explicit their thoughts and emotions in real-time [3]. Recognizing the role between usage of various social media platforms and personality traits can provide deeper analysis into users' online behaviors by analyzing language [4], to generate semantic

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

context and relevant responses [5], leads to a huge source of information for predicting individuals' personality and psychological tendencies [6].

Personality characteristics are enduring patterns of thoughts, feelings, and behaviors that define individuals from one another. The characteristics represent consistent cognition, beliefs, emotions, decisions, and behavior styles, often shared by genetic and environmental impact, impacting various aspects of life including relationships, views, opinions, new ideas, and career choice [7]. Through the lens of psychology, personality serves as a baseline framework for comprehending individual's responses to distinct situations. Furthermore, personality influences individuals emotional experience, shaping their response to stress, their sentimental analysis from different platforms like twitter and Facebook [8], certain personality traits such as neuroticism may lead individuals to higher level of anxiety, depression, and other mood disorders whereas traits like resilience and optimism can symbolize against the negative impact of stressors and promote resilience.

Personality trait detection from textual data significantly impacts various fields by opening new research directions and applications. In psychological research, it enables deeper understanding and automated assessment of traits, enhancing personalized experiences and mental health support. In marketing, it offers insights into a consumer behavior, enabling targeted and effective campaigns and advertisings [9]. In career counselling and recruitment, it aids in matching candidates with suitable job roles and improving hiring processes [10]. It also has applications in human computer interaction, improving user experience through personalized interfaces and responses. Furthermore, in fields like social media analysis, personality trait detection helps to evaluate user behavior, preferences and sentiments, that offers insights into public opinions, decision-making and trends. This study laid a foundation for future advancements such as cross-cultural studies and multilingual analysis of individual analysis, broadening the scope and utility of personality detection technologies [11].

Personality traits can be broadly categorized into 5 dimensions: openness to experience, conscientiousness, extroversion, agreeableness, and neuroticism. Each dimension represents a spectrum of characteristics, ranging from adventurous to cautiousness, from sociability to introversion, from nobility to pessimism, and from emotional stability to sensitivity [12].

Extroversion and introversion are defined axes of the Myers-Briggs Type Indicator (MBTI) dataset, psychometric tool used to evaluate personality of ones into one of well-defined personality types [13]. Extroversion/Introversion reflects where individuals focus their energy and attention. Extraverts are socially engaged with others, and energized also, while introverts are more prefer solitary activities. People who are extroversion preferences in their behaviors, with an attitude of outward focus on the external world, socially expanded and assertive, talkative, and energized by social environment. They enjoy spending their time around others, love crowd, thrive in social gatherings, and often search out for experiencing new challenges and activities [14]. Extroverts power the roles of seeking collaboration, networking, and public speaking to build and expand relationships as well as influence others. In contrast, introversion is associated with attributes of inward focus on the internal world, exhibited by introspection, reflection, and a preference to stay alone or solitary. Introverted individuals are typically reserved, thoughtful, and inclined towards solitary activities. They may prefer quiet environments where they can engage in deep contemplation, creativity, decision-making processes, and introspective exploration. Introverts excel in roles that require focused attention, analytical thinking, and independent work, leveraging their introspective nature to delve deeply into complex problems and ideas [15].

Within the MBTI framework, extroversion is associated with the Extroversion (E), while introversion is associated with the Introversion (I) preference as shown in table 1. These axes represent broader patterns of behavior and cognitive orientation, influencing how individuals interact with their environment, process information, and make decisions.

Personality detection is a new research area in this regard. As the advancements of AI for text data mining have introduced research domain of Natural Language Processing (NLP) which has gained a lot of attention globally thanks to progressions in Large language models like ChatGPT, Bard, etc. The increasing growth and power of social media through various channels including discussion, Question answering forums [16], and blogs etc. [17] has increased the importance of detecting personality traits and allows us to examine the sentimental analysis of individual [12] with use of language which is an important fact for indication of personality traits from text. This symbiotic relationship between UGC and AI-driven NLP not only enhances the utility of social web but also enriches the overall digital experience, reflecting the power of technology and human expression in the virtual realm.

In this research study, our aim is to detect extroversion personality trait of human personality: For empirical analysis, the real-world dataset MBTI has been used. For feature extraction, we used standard textual features like TF-IDF and POS have been used and for deep learning, word embeddings and sentence embeddings have also been used. Shallow Machine Learning algorithms such as Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN), Naïve Bayes (NB), and Ensemble models like Random Forest (RF), Gradient Boost (GB), Adaboost, and Extreme Gradient Boosting (XGB) used textual features. Whereas, Deep Learning algorithms Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (Bi-LSTM) use deep features of word2vec, GloVe, and sentence embeddings. In addition, the state-of-the-art Transformers-based model Bidirectional

TABLE 1. Sample sentence of IE traits.

Sentences	Trait
"I prefer to spend my whole evenings time with books rather than going outside that is full of crowd or social events."	Introversion - I
"I failed a public speaking class and lose confidence in front of stage, whatever I could do better to achieve this position again."	
"I just cherish the time of solitude within my inner world… just enjoy the me time while you can."	
"I have posted my thoughts, and I hope now community follow my trend."	Extroversion - E
"Even when I am alone, I enjoy calling or texting friend to catch up and stay connected."	
"I am not concerned with getting my feelings hurt by others who post things straight forward, instead I involved with them."	

Encoder Representations from Transformers (BERT) model is also applied. The results are evaluated using standard performance evaluation measures of accuracy, recall, precision, f-measure, AUC, and ROC. From these results, the proposed features to be effective in term of increasing classifier accuracy for the researchers to consider AI applications in personality traits is an active area of research.

Our main research contributions in this research study are as follows:

- Exploration of diverse features for personality trait detection including textual features of TF-IDF, POS tagging and deep features such as word embeddings of word2vec, Glove and sentence embeddings.
- Application of diverse shallow ML models, and ensemble models like and Adaboost for prediction of personality traits of extroversion.
- Application of state-of-the-art DL algorithms like LSTM and Bi-LSTM and transformer-based model BERT also applied.
- Conducting a comprehensive analysis to predict highest accuracy of 92.52% with deep learning algorithm Bi-LSTM integrated with advance deep feature sentence embeddings that shows top results as compared to existing studies using standard performance evaluation metrics.

The rest of the paper is organized as follows: Section II reviews the existing studies in the domain, Section III presents the proposed research framework sharing steps of the methodology, Section IV presents experimental setup sharing datasets and performance evaluation measures. The results are discussed in a comprehensive way, before concluding the manuscript and sharing potential future work.

II. RELATED WORK

The related work section of the paper encompasses two primary domains: firstly, personality detection through the application of ML techniques, and secondly, the exploration of personality detection utilizing DL methodologies.

A. EXISTING STUDIES OF MACHINE LEARNING

Predicting personality traits using ML involves depicting various models to analyze patterns and relationships within textual data. ML focuses on the use of data and algorithms to imitate learning from data and perform tasks without explicit instruction. Machine learning methods demonstrate the significant correlation between behavioral tendencies among social media text data exploring additional personality parameters and integrating experts at logical insights for deeper analysis [18], utilizing embeddings techniques and learning models to analyze language text and compute personality traits [19]. Analysis of existing studies as shown in table 2.

Personality model-based trait on introversion – extroversion demonstrates predicting approach with a reduced set of predictors by employed ML modes and class imbalances methods, such as synthetic minority over-sampling technique (SMOTE) and adaptive synthetic sampling, accuracy of 84% [20], another extending study carried out with same methodology following hybrid approach of SMOTE-ENN, results with 72% [21]. The potential of Electro Encephalo Gram (EEG) signals in personality IE trait prediction employed with genetic programming classifier, achieving an accuracy of 73.54% [22]. SVM model utilized to classify human personalities as extroverts or introverts using social media posts on Kaggle dataset, focusing on extracting extrovert or introvert label features. They employed preprocessing, feature extraction with the tf-idf method, and classification with SVM. The study found the accuracy of 84% with additional information on emoticons and mention counts [23].

Personality classification using ML algorithms on the Apache Spark platform employs SVM, LR and NB to Train and forecast personality pairings from a tweet using PySpark in personality categorization tasks based on word usage with 79.3% achieving interdependency varying degrees of success across different personality dimensions of online users [24]. The ML classifier XGBoost is used to anticipate MBTI personality traits from the text, employing preprocessing and feature selection using TF-IDF techniques laying a foundation for IE traits potential applications with accuracy of 79.01%, in organizational settings for employee selection and client servicing [25]. Ensemble model with feature selection PCA and Chi-square for prediction from social media conversations, employing supervised ML algorithms to assess traits performance of 67% with RF and 73% with LR, utilized feature extraction methods like TF-IDF and POS tagging [26], Stacked ensemble learning utilized meta model based on deep neural network with combined features of BERT to ML and DL models for the prediction of traits with tf-idf features 72.69% accuracy [27].

B. EXISTING STUDIES OF DEEP LEARNING

The subset of ML that deals with neural networks with multiple layers, capable of dealing with complex patterns and representations of data, excels in processing large volume

of data and extract hierarchical features, but often require substantial computational resources and large dataset for training. Models include convolutional networks, artificial networks, deep belief networks, and recurrent networks, and have attained state-of-the-art-performance in different domains. It emphasizes the successful use of DL models like CNN, LSTM, and RNN in the context of social media analysis [28] for predicting personality traits by using social media profiles with psychological analysis and provides intuition into current trends and future directions in text-based personality trait classification. Here is an analysis of existing studies used for personality traits detection. Classifications of YouTube personality data set with a wide range of supervised algorithms such as neural networks, proposed Bi-LSTM, with other models like Convolutional Neural Network (CNN), re-current neural networks, gated and bidirectional gated recurrent unit, and LSTM, are utilized to predict the features rates with f-score 87% [29]. The psychopathic study was investigated to classify the traits from social media text using a hybrid approach CNN+LSTM, utilized word embeddings with various configurations including changes in model layers, units, batch size, and vocabulary size were tested to achieve accuracy at 88%. The study compared the model's performance with ML classifiers comparison like DT, SVM, KNN, NB, XGBoost, RF, and LR, finding the most effective among them [30].

DeepLSTM with EEG approach was introduced using a deep-embedded clustering model, aimed at extracting personality traits, learning features and representations, and assigning groups combined with text analysis of tweets posting. Model classification is carried out on ASCERTIAN dataset with 90% accuracy to extract deep semantic features tested on data, existing shallow ML classifiers, like MLP, DT, SVM, LibSVM, KNN, and Hybrid Genetic Programming yielding the lowest prediction error [31]. Also, the Neural network model incorporates Word2Vec with CNN-LSTM layers to analyze textual data and identify personality for recommendation system enhanced with emotional factors from blog post for improved accuracy [32]. To optimize DL with limited data by using data fusion techniques and source mapping the MBTI data was mapped into the traits and integrated with the Essays and mypersonality dataset by using the BERT approach demonstrating its efficacy in automated personality detection by using social media texts, with performance of 79.7% for IE trait [33]. Multi-label semi-supervised learning techniques for personality trait text interpret the emotional information into the type indicators personality estimation model by focusing on Twitter data, emotional context in social media analytics is achieved by comparing models trained with and without emotional data emphasize the impact of preprocessing steps like stemming, lemmatization, and stop-word removal on prediction accuracy of 78.84% extroversion trait by comparing with ML based models [34]. All Weight Shared Electra for Personality prediction with 83.5% explores a novel multi-task learning model to predict traits from social media posts simultaneously. The study stands out for its use of shared weights between the two personality models, leveraging the ELECTRA transformer for text encoding, allows for a more integrated analysis of personality traits, demonstrating the correlation between traits models that outperform earlier classical approaches in both MBTI and pandora dataset predictions [35]. The task involves using a Transformer-MD model performed 66.08% on MBTI datasets to aggregate scattered social media posts, creating detailed personality profiles for users, showing marked improvements over earlier methods concate with SVM+XGB, Glove-LSTM, GRU-MLA(BERT), and BERT encoder [36].

Leveraging ML and NLP classifications, improving text representation, and feature generation using bidirectional encoder BERT by using pre-trained model ULMfit, and Elmo to predict personality axes by using online social media collected dataset based on essays + myPersonality combined robust scaling data that integrate attributes of sentiment, grammatical, augmentation, imbalanced data for better performance at 61.15% extroversion trait with proposed BERT model [37]. Sentence-BERT with comparison of ML Models is utilized to predict the personality based on question-and-answer session, combined with word2vec-sbert achieving accuracy of 73.4% for IE trait on Kaggle dataset to analyze user behavior [38]. A fuzzy logic-based approach to personality recognition using the type indicators enhances the traditional MBTI by incorporating fuzzy logic reasoning, allowing for analysis of personality types that can effectively capture the subtleties and complexities of human personality, providing a more detailed and accurate classification than traditional methods [39].

III. PROPOSED RESEARCH METHODOLOGY

The proposed research methodology is based on a few steps which include many steps. From data collection to its deployment, also employ shallow machine learning, ensemble learning, deep learning and transformer-based model techniques leveraging the rich array of features extracted from text to word embeddings to build accurate personality prediction model. The proposed framework defines the steps towards personality traits detection shown in Fig 1. After collection of dataset pre-processing is performed. Afterwards, features are computed and then features become an input for selected algorithms, evaluated using standard performance parameters.

A. DATA ACQUISITION AND SELECTION

This section discusses the source of the dataset. We utilized two datasets for the experiment, MBTI and Friends Personality dataset. We chose both data sets from online repository Kaggle and GitHub. Kaggle is a well-liked website for machine learning competitions and dataset sharing and discovery. The authenticity and dependability of the dataset were guaranteed by obtaining it from source. The focus is on label introversion (I) – Extroversion (E) to predict personality extroversion.

TABLE 2. Summary of existing studies related to extroversion.

Ref	Year	Technique	Model	Features	Dataset	Results (%)
[17]	2020	ML	GBM, KNN, RF	-	MIES	73.8
[20]	2020	ML	SVM	TF-IDF	MBTI	84
[22]	2020	Ensemble	XGB	TF-IDF	MBTI	79
[21]	2021	ML	SVM, NB, LR	count vectorization	Twitter	84
[27]	2021	DL	CNN-LSTM	BoW	MBTI	85
[31]	2021	Transformer based model	BERT	Sentence Embedding	MBTI	72
[34]	2022	Transformer based model	Elmo, ULMfit, BERT	Word embeddings	essays	59
[23]	2023	ML	LR, SVM, NB, RF	POS tagging	MBTI	67
[26]	2023	DL	CNN, LSTM, Bi-LSTM	Fast-Text	YouTube, Personality	88
[28]	2023	DL	DeepLSTM	Word embeddings	ASCERTAIN	90.3
[30]	2023	Transformer	BERT	Word embeddings	Essays, mypersonality	87
[24]	2024	Ensemble	LR, SVM, CNN, Bi-LSTM	BERT, GloVe, TF-IDF	mypersonality	72

B. DATA PREPROCESSING

Data preprocessing is performed on both datasets to make it ready for analysis and to apply ML and DL techniques. The following are stages are followed in preparing data. Initially, text is cleaned by means of elimination of raw text, irrelevant characters, digits, punctuation, HTML tags, URLs and unique symbols. Character normalization gives the data in a standardized format by assuring the uniformity of textual data, including converting characters to lowercase, and simplifying textual data. Tokenization is the process of breaking a text into individual words, which is acknowledged for further analysis. Tokenization techniques such as lemmatization or stemming are used to lessen the words to their base form known as lemma, also stemming to reduce the word from root form by eliminating the suffixes, which used to enhance consistency and lowering dimensionality. Additionally, stop phrases – taking place phrases with non-informative or irrelevant repetition that cause imbalance in textual data.

C. FEATURE ENGINEERING

The domain of personality trait shows the following method to find traits with dataset. Feature Extraction is a significant step in text data analysis, aiming to transform raw text into a format appropriate for algorithms. Various vectorization techniques according to the selected features including words, symbols, and others to be followed. It is important to identify which text features take part most to performance improvement with vectorization for converting text into numeric that algorithms can understand and process. Features are categorized into two parts, text features and word embeddings.

1) TEXTUAL FEATURE

TF (Term Frequency) indicates the repetition of each word in a document, while IDF (Inverse Document Frequency) manifests the presence of documents containing a specific word. Due to textual data, for training of ML model, we need to convert it into numerical format. For this we apply vectorizing method, called TF IDF by using scikit-learn library for statistical analysis. TF (Term Frequency) indicates the repetition of each word in a document, while IDF (Inverse Document Frequency) manifests the presence of documents containing a specific word. This is utilized to determine the significance of words within a corpus by analyzing its frequency. In this, individual posts by applying a lambda function that joins all split text into a unit string. Then combined posts are processed using tfidf_vectorizer( ), which transforms the text into numerical features. We computed these features by the following procedure. All parameters are defined in table 3 for understanding.

$P = \{P'_1, P'_2, \ldots\ldots, P'_N\}$, set of all combined posts, then set the term-frequency, as in (1).

$$TF = \left[tf_{i,j}\right] \text{ where } tf_{i,j} = \frac{count\ of\ term\ tj\ in\ post\ P'_i}{total\ term\ in\ post\ P'_i} \quad (1)$$

and inverse document frequency matrix from (2).

$$idf_j = \log\left(\frac{N}{|\{P'_i \in P : t_j \in P'_i\}|}\right) \quad (2)$$

TF-IDF score is calculated for each term *tj* in each post as in (3).

$$tf_i df_{i,j} = tf_{i,j}.idf_j \quad (3)$$

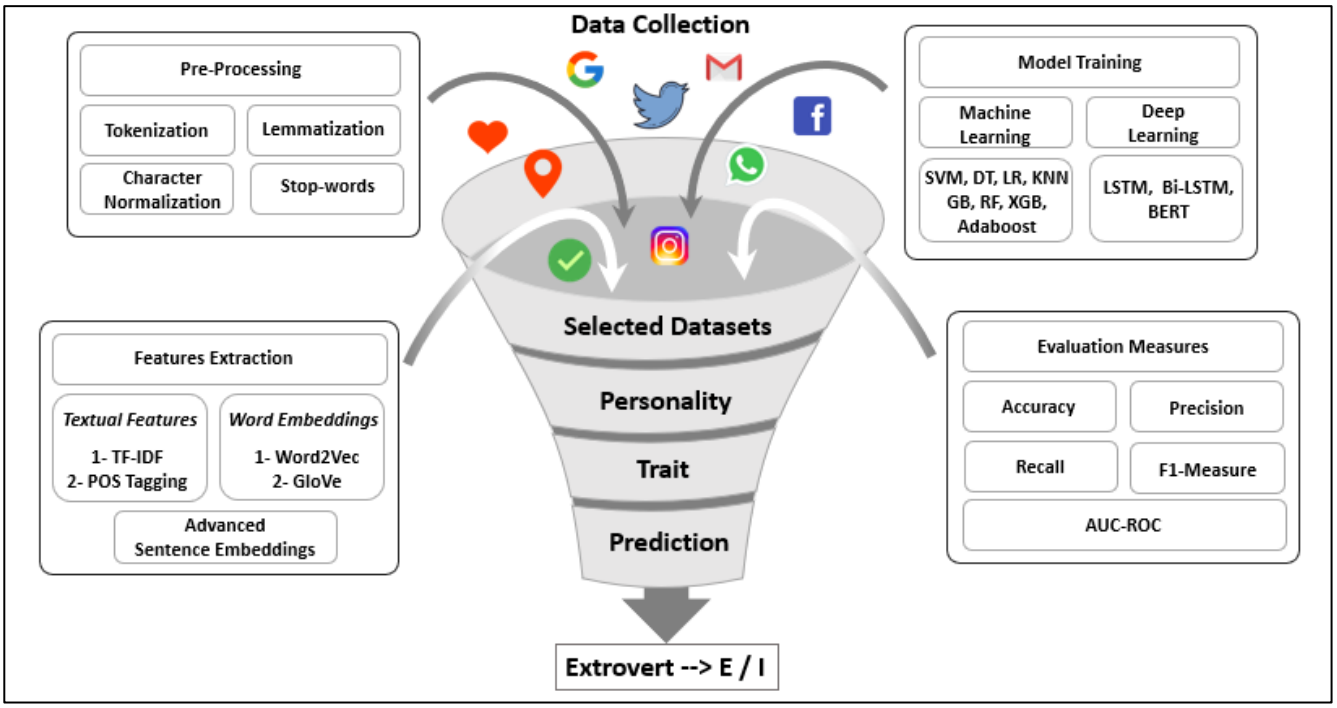

FIGURE 1. Proposed framework showing steps of research methodology.

Moreover, POS tagging is a fundamental task in NLP that involves assigning grammatical categories such as noun, verb, adjective, etc. to words in a text data corpus [40]. By tagging each word with its corresponding POS, which provides valuable linguistic information that can be used for various NLP tasks. By using the following equational function POS tags access for each token in the document with spaCy model then joined into a single string in (4) and (5).

$$POS\ (T) = token_i.pos\_tagging|\forall\ token_i \in\ doc \quad (4)$$

$$POS\_tags\ (T) = join\ (POS\ (T)) \quad (5)$$

This function is applied to each post in the dataset as in (6) resulting in a new feature column containing pos_tagged text, then used as feature set for X_{IE} with labels Y_{IE} corresponding to IE traits.

$$POS\_tags\ (P_i) = pos\_tagging\ (P_i) \quad (6)$$

Store the POS text for each post P_i in the dataset as new feature as in (7).

$$data[{}'pos\_tags{}']_i = pos\_tags\ (P_i) \quad (7)$$

Define the feature set X_{IE} as pos_text in (8).

$$X_{IE} = \{data[{}'pos\_tags{}']_i | i \in \{1, 2, \ldots., N\}\} \quad (8)$$

Define the labels y_{IE} as the corresponding labels for I/E trait from the dataset, in (9).

$$y_{IE} = \{data[{}'IE{}']_i | i \in \{1, 2, \ldots., N\}\} \quad (9)$$

Finally, these features function as the premise for the next degrees, allowing tasks that consist of sentiment analysis, text classification, or information retrieval.

2) WORD EMBEDDINGS FEATURE

For classification of data, we undertake a sturdy methodology leveraging superior phrase embeddings strategies, which includes Word2Vec, Glove, and Sentence Embeddings, entails several systematic steps. By following preprocessing, follow Word2Vec chosen word embeddings strategies to change the text data map into dense vector representations, capturing semantic relationship among words and sentences based on their co-occurrences in massive textual content corpora. The objective function computed by using (10) to maximize the average log probability.

$$J = \frac{1}{T} \bigcup_{t=1}^{T} \bigcup_{c \le j \le c, j \ne 0} \log P\left(w_{t+j|w_t}\right) \quad (10)$$

The probability of a context word given the target word is computed using the function in (11).

$$P\left(w_o|w_I\right) = \frac{\exp\left(v_{wo}^T v_{wI}\right)}{\sum_{w \in V} \exp\left(v_w^T v_{wI}\right)} \quad (11)$$

Additionally, we also applied GloVe embeddings, using international phrase-phrase co-incidence records, generates embeddings that emphasize each local and global context facts by leveraging the global word-word co-occurrence statistics from a corpus. To compute the co-occurrence probabilities for encoding meaningful relationships between words

by using (12).

$$P_{ij} = \frac{X_{ij}}{X_i} \tag{12}$$

So, their relation and ratio function are captured with the help of (13) and (14).

$$\frac{P_{ik}}{P_{jk}} = \frac{X_{ik}/X_i}{X_{jk}/X_j} = \frac{X_{ik}X_j}{X_{jk}X_i} \tag{13}$$

$$F\left(w_i, w_i, \hat{w}_k\right) = \frac{P_{ik}}{P_{jk}} \tag{14}$$

By applying logarithmic function helps to linearize the relationship between the counts and word vectors with the help of (15).

$$w_i^T w_i + b_i + b_j = \log\left(X_{ij}\right) \tag{15}$$

To minimize the difference between word pairs, basic loss function for single word pair (i, j) is applied with the help of (16).

$$L_{ij} = \left(w_i^T w_i + b_i + b_j = \log\left(X_{ij}\right)\right)^2 \tag{16}$$

We also compute less frequent co-occurrence words as in (17); weighting function is used that helps to prevent them dominating the loss function.

$$f(x) = \begin{cases} \left(\frac{X_{ij}}{X_{max}}\right)^{\alpha} & \text{if } X_{ij} < X_{max} \\ 1, & \text{if } X_{ij} \geq X_{max} \end{cases} \tag{17}$$

Combining the loss function with the weighting function, the final loss function for the GloVe over all word pair is calculated as in (18).

$$J = \prod_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T w_i + b_i + b_j = \log\left(X_{ij}\right)\right)^2 \tag{18}$$

Moreover, by combining Sentence Transformer, an effective version skilled to generate sentence embeddings, allowing contextual data and semantics similarities at the sentence stage. Through systematic assessment and evaluation. We utilized embeddings with SBERT bert-base-nli-means-tokens library using Siamese network. Two input Sentences are passed to BERT and pooling layer generate their embeddings. Each sentence is converted to a sequence of tokens through (19).

$$S = [w_1, w_2, \ldots.., w_N] \tag{19}$$

Each token w_i is converted into vector through embeddings. To assign weights to layer representation of sentence embeddings calculated through (20).

$$R = \sum_{i=1}^{L} w_i . E_i \tag{20}$$

After assigning weights to the sentence, we computed the similarity score through (21) between pairs of sentences by using cosine similarity and then normalized the weights as in (22).

$$Sim(A, B) = \frac{R_A . R_B}{\|R_A\| . \|R_B\|} = \frac{\sum_{i=1}^{n} R_A . R_B}{\left(\sqrt{\sum_{i=1}^{n} R_A^2}\right) . \left(\sqrt{\sum_{i=1}^{n} R_B^2}\right)} \tag{21}$$

$$w_i = \frac{\exp(w_i)}{\sum_{i=1}^{L} \exp(w_i)} \tag{22}$$

Then objective function involves minimizing a loss function, as in (23).

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left[y_i . d^2 + (1 - y_i) . \max\left(0, m - d^2\right) \right] \tag{23}$$

Hence, we propose to decide the effective embeddings technique for reinforcing the classification overall performance of the proposed model. This complete method guarantees leveraging a new word embeddings approach to complement the semantic information of textual information and optimized type outcome in this research.

D. APPLIED ALGORITHMS

We propose a complete approach integrating diverse ML and DL based models for personality traits predictions. Analysis of how different NLP, ML and DL methods, including supervised, unsupervised, and hybrid approaches, can effectively categorize the personality traits based on text classification by considering language as concerned elements [41]. Initially, preprocess the textual descriptions of personality types to make sure consistency and cleanliness. For training, ensemble learning strategies are utilized that consisting of GB, RF, XGBoost, and Ada-boost, leveraging their capability to detect complicated patterns in high dimensional statistics. Additionally, we employ cutting-edge learning transformation model BERT which excels in information and generating natural language textual content. Subsequently, also applied DL model LSTM and Bi-LSTM to predict personality traits. For comparison, use of a set of shallow ML models to discover the different procedures for prediction.

By systematically comparing these models on the MBTI dataset, goal to pick out the efficient set of rules for as it should be predicting the traits based on textual description, hence contributing to deeper expertise of persona evaluation using machine and natural language processing technique. Here is a certain discussion of each model and how they can be applied to predict personality traits using the data set.

Here is a certain discussion of each model and how they can be applied to predict personality traits using the data set.

1) CONVENTIONAL MACHINE LEARNING MODELS

Conventional machine learning models that have a small number of layers in their architecture including input, hidden and output layer. These models are often used for simpler classification and regression tasks and are commonly easier to interpret and understand.

- **K-Nearest Neighbors:** K-NN is a simple but powerful algorithm for classification and regression tasks. It works by means of finding the closest factor point in the features to the given dataset. Once the K most similar

TABLE 3. Parameters of defined equations.

Symbols	Description
$P_{i,j}$	Represents the j-th part of the i-th post
n_i	Number of paths in the i-th post
P'_i	Number of i-th posts in the dataset
$tf_{i,j}$	Measure of how frequently a term i appears in a document j
$POS(T)$	Function of pos tagging that takes text T as input
w_t	Target word
w_{t+j}	Context words
C	Context window size
T	Number of texts in the corpus.
v_{wo}	Output vector of the context word
v_{wI}	Input vector of targeted word
V	Vocabulary
X_{ij}	Co-occurrence count of words i and j
X_i	Total number of times word i appears in the corpus
w_i and w_j	Vector words for i and j
$\hat{w}_k$	Vector word for ratio function
b_i and b_j	Bias terms
X_{max}	Threshold for co-occurrence count
α	Parameter is set to control the scaling
R	Weighted representation
L	Total number of layers
E_i	Representation of from the i-th layer
w_i	Representation of assign weights to the i-th layer
d	Distance between embeddings
M	Margin parameter

data points have been observed, the algorithm assigns the new data point the majority class label of those points. Prediction is primarily based on majority class. K-NN used to predict persona traits primarily based at the similarity of textual description in dataset. It calculates the similarity among the input textual content and present data points and predicts the maximum frequent personality types among many of the nearest neighbors.

- **Support Vector Machine:** For classification, is a supervised model, utilized by uncovering the premier hyperplane that separates different traits inside the function area described through the dataset. By mapping textual description to characteristic vectors, SVM classifies new instances into suitable personality categories. It works by locating the optimal boundary, known as the hyperplane, that best splits the two classes of data points in space. The hyperplane is decided by exploiting the edge, which is the distance between the hyperplane and the nearest points from each class.
- **Naïve Bayes:** NB is a probabilistic classifier primarily based on Bayes theorem and the theory of independence between functions. It computes the probability of every class given the enter features and chooses the class with the best possibility as the prediction. It predicts the persona trait based totally on the incidence of phrases or phrases in textual descriptions furnished within the dataset.
- **Logistics Regression:** LR is a linear model suitable for binary classification. It estimates the possibility that a given point relates to a specific point using logistic features. By learning the connection among input, LR makes predictions based on textual description.
- **Decision Trees:** DT used to predict traits by iteratively partitioning the features relying on data to expand the information gain with prediction that is based on the classified structure of trees. To maximize class in each partition, it chooses each node of the tree. In a decision tree, each leaf node shows a decision, and each parent node displays the result. The attribute of the tree denotes the results of each decision. The tree is initiated with a root node, which shows the primary decision. Then, for each result, a new branch is linked to the tree. This process is followed until all results have been measured for.

2) ENSEMBLE LEARNING MODELS

Ensemble Learning is a ML method that combines many models to boost the performance and predictive accuracy of the models. The essential idea at the back of the ensemble studying is to leverage the collective intelligence of more than one model instead of relying on single model. Here we are using these ensemble models.

- **AdaBoost (Adaptive boosting)** is an algorithm for classification, operates by successively training a sequence of weak learners (such as decision trees) using weighted data that is trained on input. In each iteration, AdaBoost allocates higher weights to misclassified classes, forcing subsequent weak learners to focus more on the hard to classify instances. The final prediction is made by integrating the prediction of all weak learners, commonly weighted with the aid of their performance for the duration of training.
- **Gradient Boosting (GB):** This algorithm is used for regression and classification tasks. Leveraging the strength of multiple weak learners like DT, GB constructs a strong predictive model by iteratively minimizing the errors made by ensemble through gradient descent optimization. It is a sequential approach to building models to ensure high accuracy and the ability to acquire complex relationships in the data. Despite its computational demands and the need for hyperparameter tuning, GB stands as a robust and versatile method for predicting models.
- **XGBoost:** Extreme Gradient Boosting is a scalable and competent implementation of machines for boosting purposes. It is widely used in regression and challenging problems of classifications. It works by means of constructing an ensemble of susceptible selection trees sequentially, with each tree to correct the mistakes made by the preceding nodes. It makes use of a regularized objective function to prevent overfitting and comprise features like tree pruning, column subsampling and coping with missing values to improve model overall performance and robustness.

- **Random Forest:** RF is an ensemble method that structures the multiple decision trees to forecast personality trends for training based on textual data. By pooling individual tree predictions, RF gives solid results for each personality type.

Ensemble strategies like XGBoost and AdaBoost can efficiently handle excessive-dimensional statistics like the dataset and capture complex relationships between functions and traits. By combining multiple weak novices, those can detect patterns and make correct predictions.

3) CONCURRENT DEEP LEARNING ALGORITHMS

The training of deep learning models using parallel processing techniques to accelerate training time and improve accuracy refers to the concurrent deep learning models, enables large scale learning models on massive datasets, offering rapid experimentation and model optimization [42]. Models include RNN, CNN, LSTM, and Bi-LSTM. Here is an analysis of existing studies with this model used for personality traits detection.

- **Long Short-Term Memory (LSTM):** This algorithm excels in modeling sequential data by introducing specialized memory cells that can retail information over extended time intervals. This can be learned from an input data as choice of words and linguistic patterns over time to infer personality traits. LSTM enables selectively to store the information, allowing them to predict complex temporal patterns and relationships.
- **Bi-LSTM: Bidirectional:** Long Short-Term Memory (Bi-LSTM) enhance the capabilities by handling sequences in both past and future directions concurrently, enabling the model to access backward and forward context for better understanding of the input sequences. For personality trait prediction this model hold promises for the advance understanding of human behavior through contextual data.

4) TRANSFER LEARNING MODELS

Transformer based model has drastically boosted the state-of-the-art in NLP tasks inclusive of text type, sentiment evaluation, machine translation, and query answering. Their capacity to capture long-term dependencies and contextual statistics has made them indispensable in NLP studies.

- **Bidirectional Encoder Representation from Transformer (BERT):** BERT, unlike traditional models that rely on both position direction like from left to right or right to left context, BERT captures bidirectional context by using both previous and succeeding phrases. It is pre- trained in big textual content corpora using masked language modelling and next sentence prediction responsibilities, enabling it to understand the contextual dating inside text information.

These models are state-of-the-art language models capable of understanding and generating natural language textual content. By fine-tuning these models on selected dataset, they could learn how to encode textual description of personality developments and expect persona based on text inputs. By applying these models to the dataset, we discover numerous strategies to predict persona traits primarily based on textual description. Each model offers unique advantages and can be carried out depending on the traits of the dataset and complexity of the prediction undertaking.

E. EXPERIMENTAL SETUP

Here, in the experimental setup, we discussed the data set to be used for the empirical analysis and mentioned the standard performance-related aviation measure which will be used for the comparison of the method with the existing studies.

1) MBTI DATASET

The MBTI dataset includes data gathered based on the Myers-Briggs Type Indicator,[1] a tool to classify individuals into 16 types with 4 indicators. This dataset has about 8673 rows of data. This dataset usually consists of survey responses and textual data, where respondents' language use and behavioral cues are linked with one of the MBTI types by mapping personality trait extroversion onto the MBTI dataset IE. Each entry in the dataset is often tagged with the relevant MBTI type, like INFP, ESTJ, etc., based on the individual's replies or textual analysis. In research, this dataset attends as a useful material for studying personality traits, enabling the development, and testing of models that predict MBTI types from distinct types of data, especially in studies where text analysis and NLP approaches are used to infer personality types from written content.

2) FRIENDS PERSONA DATASET

This dataset[2] contains dialogs from the TV show Friends based on short conversations from the first four shows session based on personality traits. The dataset consists of total 711 number of entries with six columns showing text posts against each trait.

3) PERFORMANCE EVALUATION MEASURE

Personality traits prediction is measured using a standard performance evaluation measure, including accuracy, precision, recall, and F1 score. The efficacy of datasets, and prediction models, is evaluated using these parameters. In evaluating a model's performance on an MBTI dataset, key metrics are calculated by following equations in table 4.

These metrics offer a multifaceted view of the model's ability to classify personality types accurately from the dataset.

Where TP, TN FP, and FN stand for True Positive, True Negative, False Positive and False Negative, respectively.

[1] https://www.kaggle.com/datasets/datasnaek/mbti-type / last accessed on 09 Feb, 2024

[2] https://github.com/emorynlp/personality-detection / last accessed on 26 July 2024

TABLE 4. Equations of evaluation measures.

Metrics	Equation
Accuracy	$\frac{TP+TN}{TF+FN+FP+TP}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-score	$\frac{2(Precision*Recall)}{Precision+Recall}$
ROC - True Positive Rate	$\frac{TP}{TP+FN}$
ROC - False Positive Rate	$\frac{FP}{FP+TN}$
AUC	$\sum_{i=1}^{n-1} \frac{(FP_{i+1} - FP_i).(TP_{i+1} - TP_i)}{2}$

IV. RESULTS AND DISCUSSION

In this section, unveil the investigations of proposed methodology through experiment, evaluations, and empirical analysis of predictive results on personality traits from textual data. Through the examination of findings with performance matrices, and comparative analyses shedding light on the strengths, limitations and implications of each approach that further contribute to the advancement of this field. To predict the results of personality traits with the given dataset label, trait extroversion maps to Introversion/Extroversion (I/E) axes. This dimension reveals where individuals focus their concentration and get their energy. Introverts tend to focus inwardly and solely, while introverts focus outwardly and expand themselves as a socially. We apply selected text features TF-IDF and POS tagging applied with models SVM, Gradient Boosting, DT, NB, RF, XGBoost, AdaBoost, KNN and LR are applied on the selected feature set to train the test dataset. Additionally, deep models with word embeddings (Word2Vec, GloVe and Sentence Embeddings) also employ concurrent models such as LSTM, and Bi-LSTM and transformer-based model BERT with encoder for comparison. A systematic approach was used to preprocessing including tokenization, and lemmatization to standardize the meaningful features. The text was transformed into numerical representations using textual and word embeddings as feature engineering. After this model training was employed to learn the patterns in textual data that distinguish between introvert, or introvert, collectively, aiming to maximize the margin between classes while minimizing the classification errors. The resulting model was then evaluated using performance metrics of these techniques.

A. EXPLORATORY DATA ANALYSIS OF MBTI DATASET

Let us discuss the data visualization and interpreting the outcomes using MBTI dataset. In this research, features are extracted against number of posts from selected datasets. These features are analyzed by data visualization process as depicted in Fig 2. The dataset is clearly unbalanced

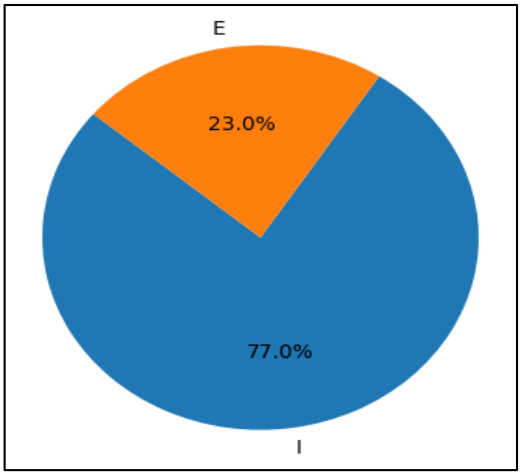

FIGURE 2. Analysis of MBTI dataset posts visualization.

throughout the classes. For overall 80000, the number of posts available that contain a content-based features whose values lies traits between I and E. Users who comment on social media frequently are more introvert on emotional based. Word cloud is generated by the most frequently occurring words among the text, as display in Fig 3.

B. EXPLORATORY DATA ANALYSIS OF FRIENDSPERSONA DATASET

Integrating the visualizations and statistics for labeled text that how individuals differ in their online behavior. Fig 4 showing the distribution of text lengths in a dataset refer to the number of words in the dialogues spoken by characters. Right-skewed indicating that most dialogues are relatively short, with majority falling between 0 and 1000 words. This shows that characters tend to have brief interactions for a TV show where quick changes are common. Fig 5 shows the distribution of extroversion trait as the main target of this research to explore extroversion personality. The

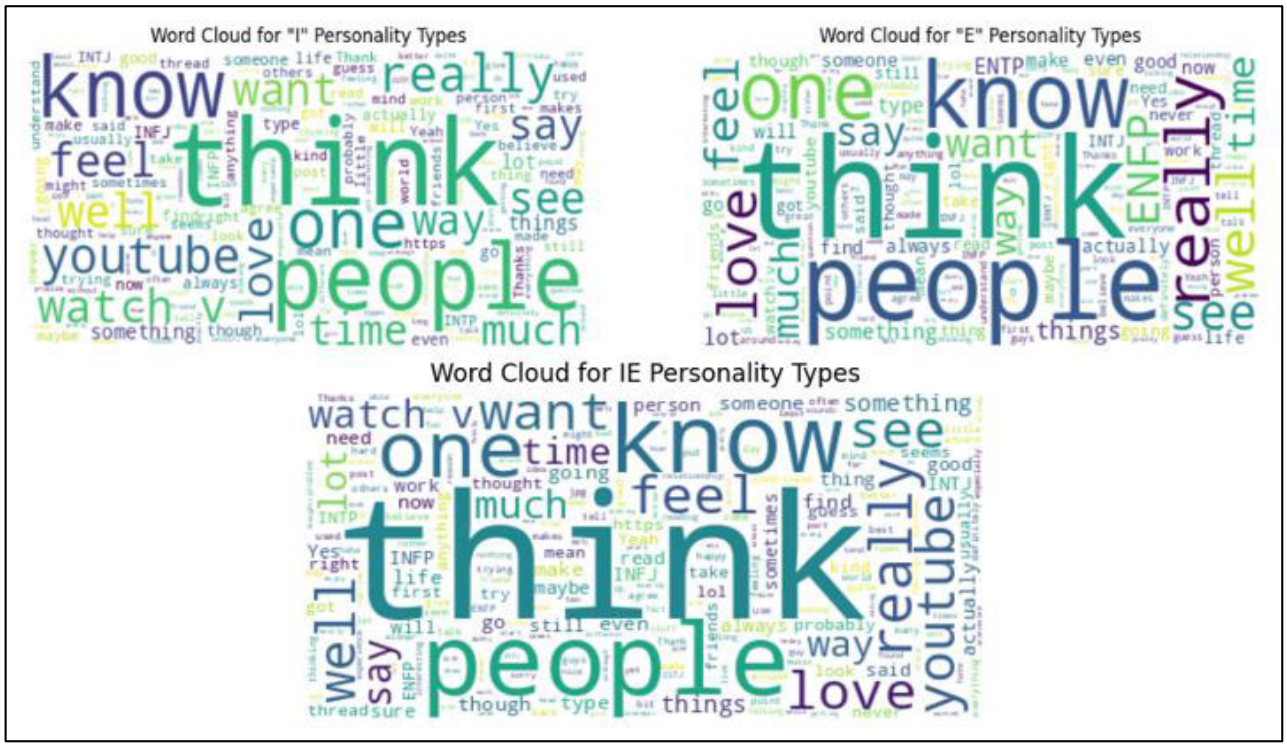

FIGURE 3. Word cloud visualization of a) I for Introvert content only, b) E for Extrovert content only and c) for both IE.

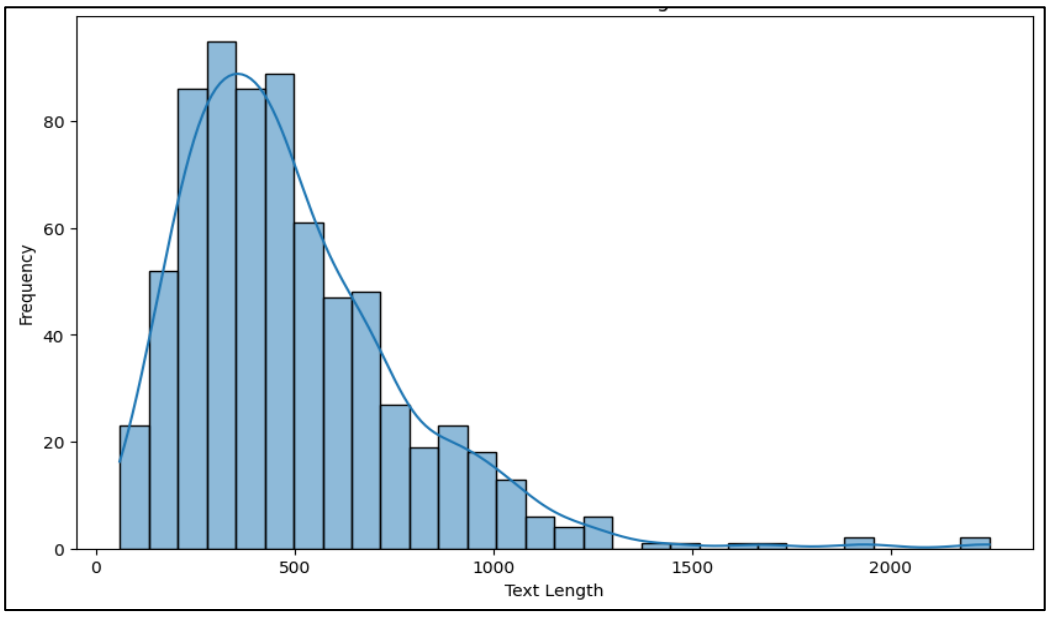

FIGURE 4. Histogram showing the distribution of text lengths in dataset.

distribution is bimodal, with significant peaks at 0 and 1, indicating a clear divide between introverted and extroverted characters. This showed that dataset includes characters with distinct personality types, either very introverted or extroverted, with fewer characters falling in the middle range. This reflect the diverse range of personalities shows that from highly extroverted Joey character to the more introverted Ross character, showing their personality in the TV shows session.

Fig 6 displaying most frequent words used in the dialogue of extroverted characters. Prominent words like 'look', 'how', 'go', 'well', 'okay' and character names like 'geller' (Ross

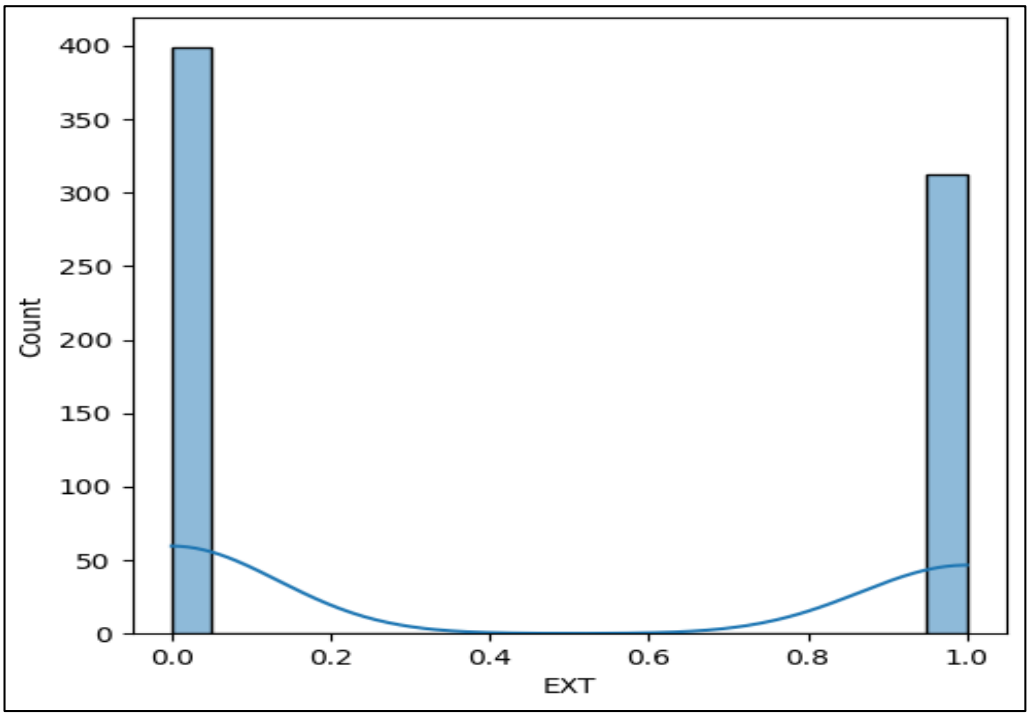

FIGURE 5. Histogram showing the distribution of extroversion trait score in dataset.

FIGURE 6. Most frequent words of extroverted trait.

and Monica Geller) and 'tribbiani' (Joey Tribbiani) appear frequently. This word cloud gives and impression of common themes and interactions in the show. Words like 'know', 'look', and 'well' suggest casual, conversational dialogues which is typical of the extroverted and social nature of the characters.

C. RESULTS WITH MACHINE LEARNING MODELS

By applying ML models to classify personality traits on MBTI dataset, using various algorithms shows promising performance across multiple metrics. Notably, SVM, GB, and XGB achieved the highest accuracy of 86.05%, indicating effectiveness in distinguishing between introverted and

TABLE 5. Performance of ML classifiers with feature selection (results in% age).

MBTI DATASET					
Features	**Models**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
TF-IDF	**SVM**	**86.05**	**85**	**86**	**85**
	NB	78.0	61	78	68
	LR	84.0	84	84	84
	DT	77.47	77	77	77
	KNN	79.71	77	80	78
	GB	86.05	85	86	85
	RF	79.6	81	80	79
	XGB	86.05	85	86	85
	AdaBoost	84.09	83	84	83
POS	SVM	76	68	76	70
	NB	70	64	68	69
	LR	78	61	78	68
	DT	64	67	64	65
	KNN	72	65	72	68
	GB	**78**	**64**	**78**	**68**
	RF	78	67	78	68
	XGB	76	68	76	70
	AdaBoost	78	66	78	69
FRIENDSPERSONA DATASET					
Features	**Models**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
TF-IDF	SVM	**82**	**80**	**79**	**77**
	NB	75	71	75	65
	LR	64	61	61	59
	DT	74	73	73	73
	KNN	55	55	55	55
	GB	81	79	79	78
	RF	72	71	70	68
	XGB	73	73	73	73
	AdaBoost	70	70	70	70
POS	SVM	65	63	62	60
	NB	60	54	58	59
	LR	52	52	52	52
	DT	56	55	55	55
	KNN	55	52	52	50
	GB	70	69	69	68
	RF	64	62	62	61
	XGB	75	75	75	75
	AdaBoost	56	56	56	56

extroverted individuals. These models also exhibit high precision, recall and f-score highlighting a balanced performance in correctly identifying both introverted and extroverted personalities.

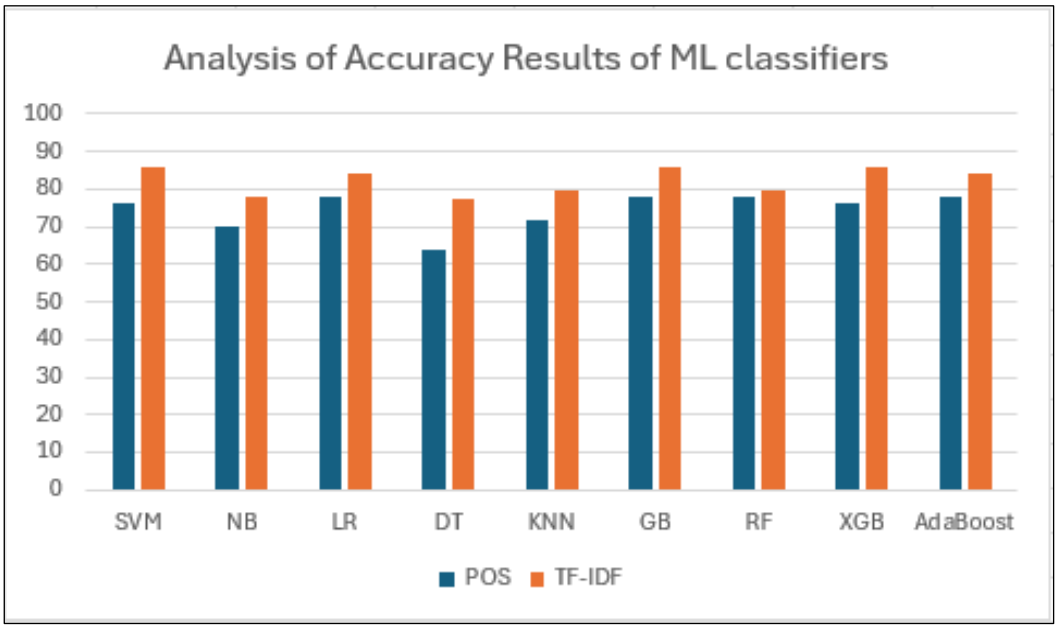

FIGURE 7. Combined accuracy analysis of proposed classifiers with selected features using MBTI dataset.

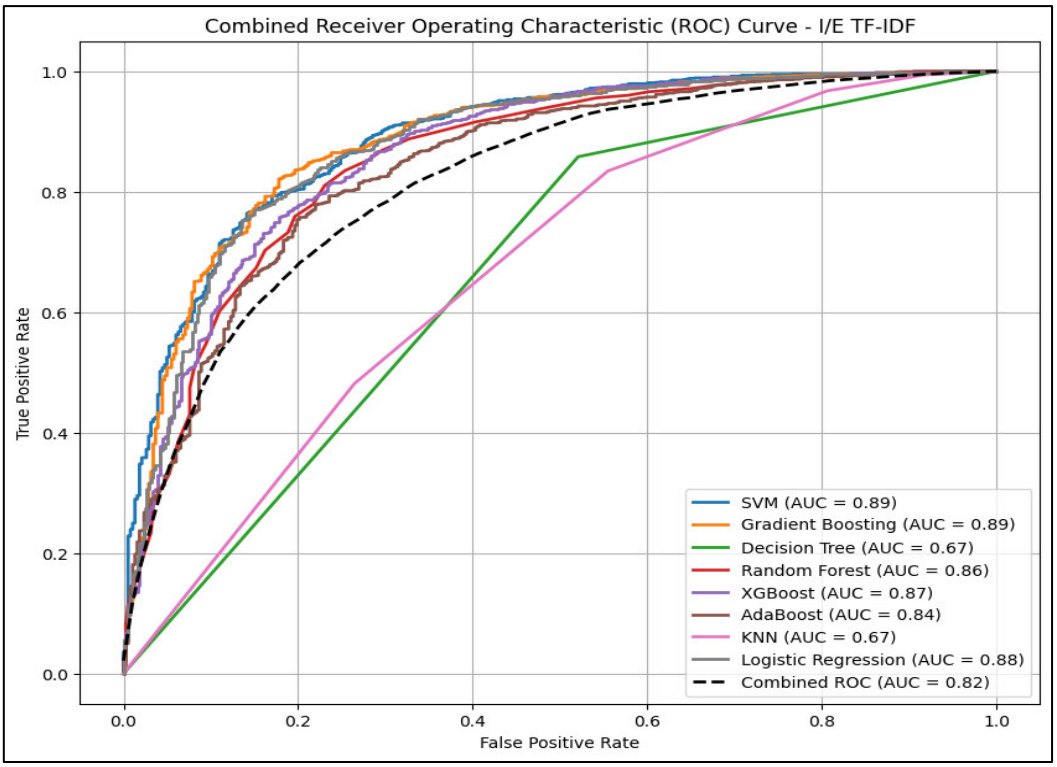

FIGURE 8. Comparative analysis of ROC curve with ML-TFIDF Feature using MBTI dataset.

These models are known for their capability to handle noisy and complex data. Similarly, GB and XGB iteratively improve model performance by focusing on instances that are difficult to correctly classify, thus enhancing the model's predictive accuracy. Moreover, NB or DT achieving accuracy of 78% and 77% respectively, as they employ techniques such as regularization and ensemble learning to mitigate overfitting tendencies. LR and Adaboost also performed well with accuracies of 84% and 84.09%, respectively. KNN and RF with an accuracy of 79.71% and 79.6% respectively,

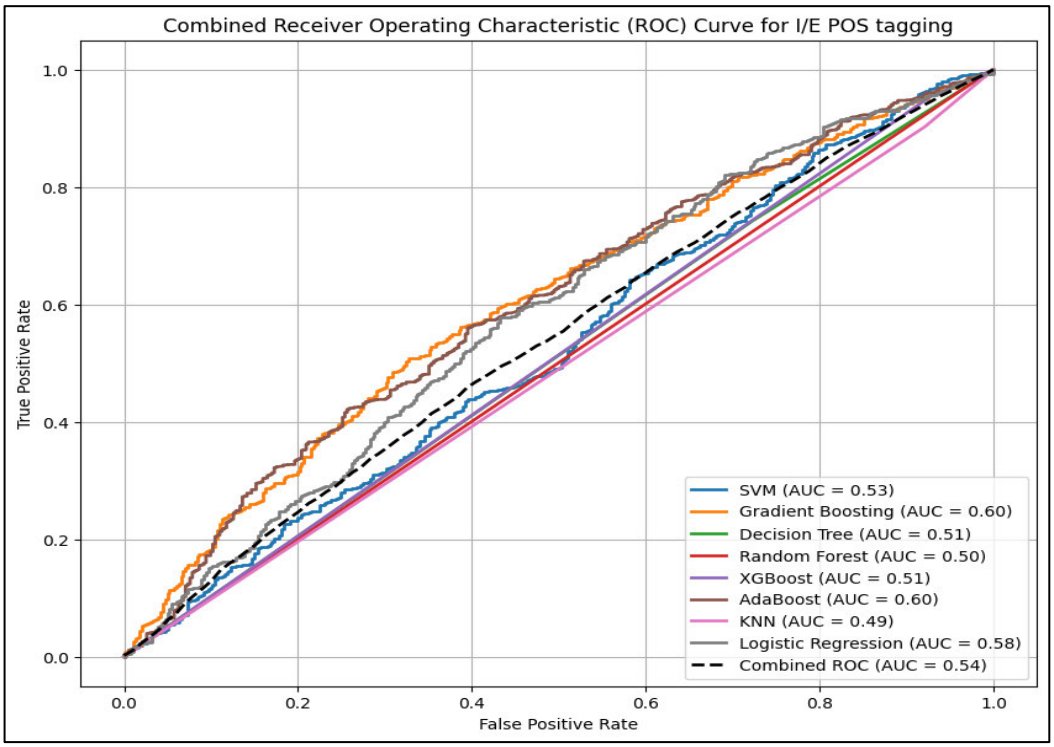

FIGURE 9. Comparative analysis of ROC curve with ML-POS tagging Feature using MBTI dataset.

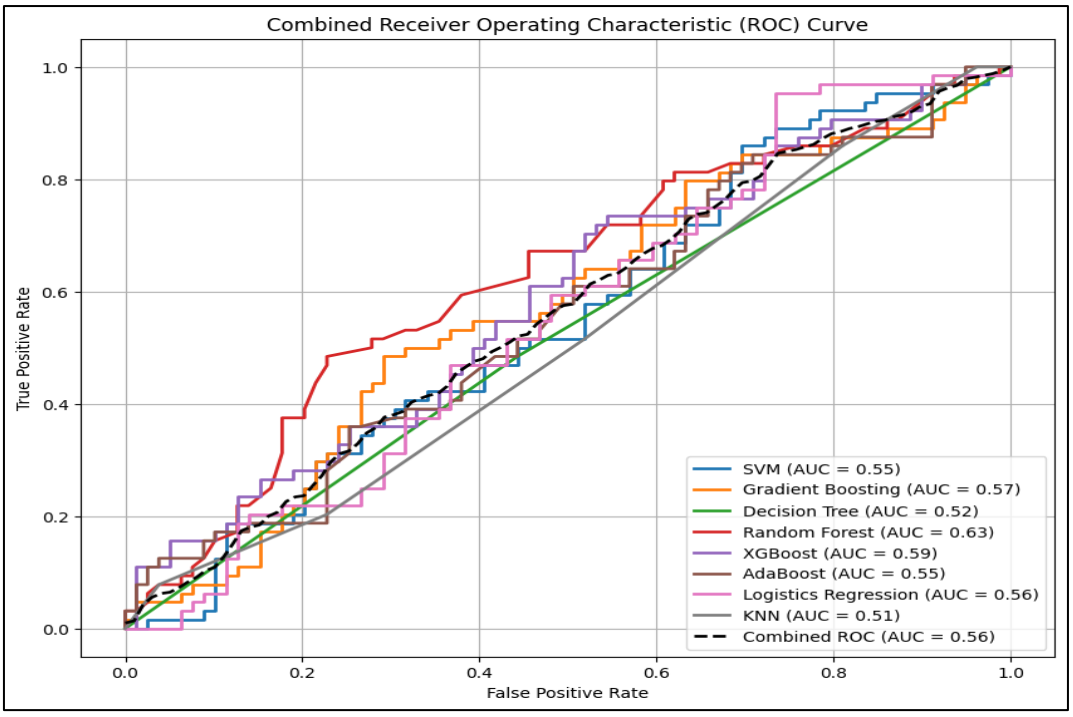

FIGURE 10. Comparative analysis of ROC curve with ML- TFIDF Feature using FriendsPersona dataset.

slightly below of top-performing models as KNN model relies on similarity principle where instances are classified based on the majority class among their nearest neighbors in feature space. However, The use of feature TF-IDF contributed to enhancing the performance of these models by prioritizing informative features and reducing noise in the

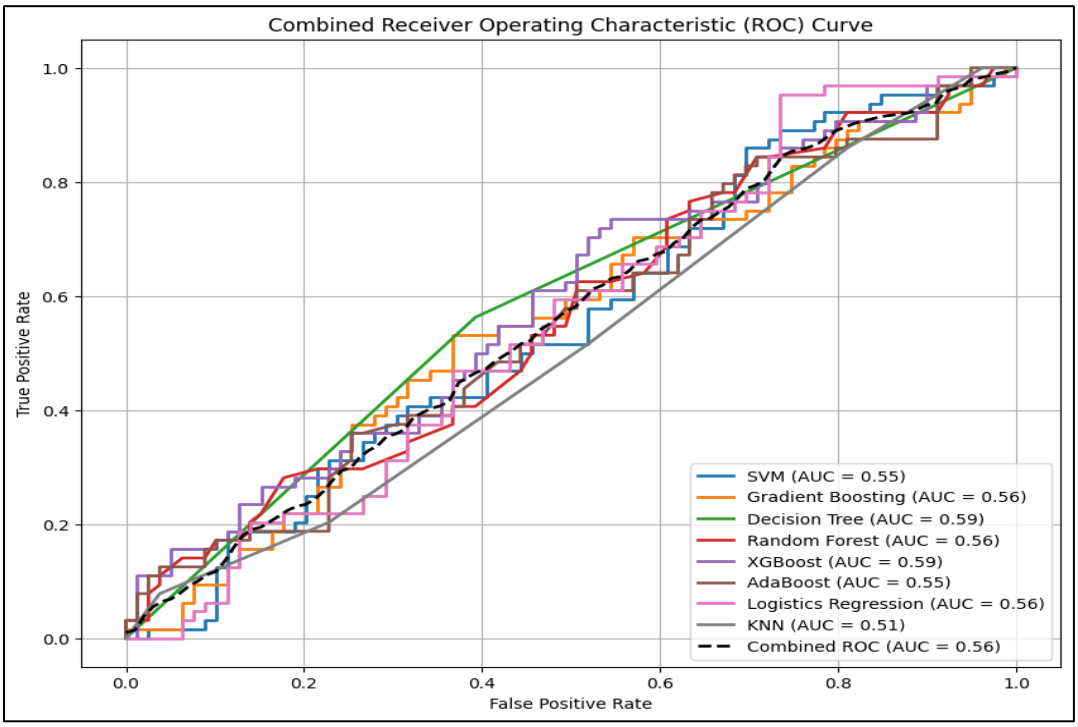

FIGURE 11. Comparative analysis of ROC curve with ML-POS tagging Feature using FriendsPersona dataset.

dataset. Results with shallow machine learning are illustrated in table 5. Afterwards, categorized textual features with different classifiers are combined for more investigation to assess the performance of selected ML models. For all combinations of classifiers. Comparing of results obtained from feature selection using TF-IDF and POS tagging reveals notable performance. When employing TF-IDF with SVM, GB and XGB achieved high accuracy and f-score as well. This indicates that TF-IDF effectively captures information for distinguishing between introverts or extroverts' traits. POS tagged exhibit lower accuracies, with SVM achieving the highest accuracy of 78%, but overall low as compared to TF-IDF. The overall performance of these proposed classifiers with textual TF-IDF and POS feature for personality trait detection are helpful for classifiers to detect traits more accurately. Evident from the result of average and best accuracy displayed in Fig 7. For the comparative analysis of model performance to identify the superior model results ROC was also employed, as shown in Fig 8 and 9.

Results on FriendsPersona dataset obtained as shown in table 5. The performance of various ML models shows TF-IDF text feature with SVM model, emerged as best model with an accuracy of 82% while GB and NB also performed well with accuracies of 81% and 75% respectively. In contrast, POS tagging results were generally lower, with XGB achieving the highest accuracy of 75%. Overall, TF-IDF feature extraction method for trait prediction as it consistently yielded better performance across the models compared to POS tagging.

Fig 10 and 11, provided ROC curves compare different ML models in predicting the extroversion personality trait using two different textual features. The AUC-ROC (Area Under the Curve – Receiver Operating Characteristics) score indicates the performance of these models. In both graphs, RF and XGB generally outperform using ensemble models showing the highest AUC (0.63 with TF-IDF and 0.56 with POS tagging). The combined ROC for TF-IDF has an AUC of 0.57, slightly better than the combined ROC for POS tagging (AUC of 0.56) across shallow ML models. Overall, while both features set show predictive ability, TF-IDF appears to provide a marginally better basis for model performance, especially with RF model.

D. RESULTS WITH DEEP LEARNING ALGORITHMS

By applying word embeddings with models' settings LSTM, and Bi-LSTM are applied on the selected feature set to train the test dataset. The text was transformed into lower-dimensional as a feature engineering. After this model training was employed to learn the patterns in textual data that distinguish between introvert, introvert, agreeableness, and consciousness collectively, aiming to maximize the margin between classes while minimizing the classification errors. Using LSTM and Bi-LSTM models and pre-trained word embeddings like word2vec, glove, and sentence transformer, highly effective in predicting personality traits from the text data. LSTM is used to retain the information over long sequences and capture dependencies over time. Bi-LSTM enhances the architecture of LSTM, processing in both

backward and forward direction. The following table 6 shows the parameters that are used to define the model architecture.

TABLE 6. Parameter settings of LSTM and BI-LSTM.

Parameters	Values
Input vector size	2000
Vocabulary size	1000
Embedding dimension	128
Unit size	100
Number of hidden layers	4
Activation function	Sigmoid
Learning Rate	0.001
Random State	42
Optimizer	Adam
Number of epochs	35
Batch size	64

With these embeddings to better understand the traits from text, results of models along both datasets are illustrated in table 7. Using MBTI dataset, LSTM with Word2vec embeddings demonstrate steady improvement over epochs reaches to 89.67%. Glove embeddings peak at accuracy of 87.56%, that exhibits positive trends. Sentence transformers outperform both word2vec and glove and reach 91.50% accuracy while Bi-LSTM with Word2vec embeddings demonstrate steady improvement over epochs reaches 88.63%. Glove embeddings peak at accuracy of 89.45%, that exhibits positive trends. Sentence transformer outperforms both word2vec and glove and reaches 92.52% accuracy as combined results are shown in Fig 12. The superior results of Bi-LSTM are attributed to the ability of sentence embeddings to capture semantic information from textual data. Additionally, the nature of Bi-LSTM model, Overall, these results highlight the model's potential for predicting personality traits effectively, as shown in Fig 13 and 14. Further complemented the utilization of sentence embeddings by capturing both past and future context, allowing for a more comprehending understanding of input text data. While Glove and word2vec embeddings also yielded notable accuracies, as their reliance on static word representations limited their ability to capture the dynamics and context nature, thus resulting in slightly lower performance compared to sentence embeddings. Overall, the combination of Bi-LSTM with sentence embeddings proved the most effective in accuracy predicting personality traits.

The FriendsPersona dataset was also evaluated using deep models to predict the trait. The traditional embeddings Word2Vec and Glove achieved the highest performance, with accuracy, precision, recall and f1-score of 89%. The LSTM model with Glove also performed strongly, with an accuracy of 87%, enhancing the model's ability to capture semantic relationships and context. Both models showed improved results with Glove embeddings compared to word2vec and advanced sentence embeddings. Advance word embeddings also applied which consider the entire sentence context rather than individual words, also demonstrated robust performance to understand the language patterns and relationships for personality prediction. Fig 15 depicts the combined analysis of these results along with deep features.

Transformer based models are also used to estimate the performance of model on personality traits by generating coherent and contextually relevant text using both datasets, use BERT in this research study, which is an attention mechanism that provides high quality representation of text. For BERT encoding with following parameters as shown in table 8. In comparison to other deep models used for predicting personality traits, the BERT encoder highlighted competitive performance on MBTI dataset, with an accuracy of 82.80% with 73% f-measure results. While using Friends Persona dataset, BERT achieves accuracy of 72.5%, comparatively lower than MBTI dataset to predict the trait, as shown in table 9. BERT's ability to capture personality traits is attributed to its deep bidirectional nature, allowing it to understand context from both directions in a text. This capability helps in capturing linguistic patterns for prediction. BERT demonstrated its capability to classify traits based on encoded text representation.

The comparative analysis of both dataset MBTI and Friends Persona reveals notable differences in the performance of various models and features to predict the extroversion personality trait using textual data. For ML model using textual feature TF-IDF, MBTI dataset consistently achieved higher accuracies, with SVM, GB and XGB reaching 86% whereas the Friends Persona dataset had its performance with SVM at 82%. POS features resulted in lower performance overall across both datasets, with the MBTI dataset's SVM at 76% and Friends Persona XGB at 75% to analyze the linguistic patterns for prediction. Similarly, DL models using deep features show highest performance compared to conventional ML approaches to predict the extrovert patterns from datasets. On the MBTI dataset, Bi-LSTM with sentence embeddings achieved highest accuracy of 92% significantly higher than any model on the Friends Persona dataset, which shows its best capability to trace the text pattern of introvert or extrovert achieving an accuracy of 89% with Glove model. The BERT transformer-based model showed consistent performance across both datasets with an accuracy of 82% on MBTI dataset and 72% on Friends Persona dataset. This indicates that while BERT is effective, the choice of embeddings and dataset characteristics significantly influence the results. Overall, the MBTI dataset achieved better accuracy across most models and techniques, emphasizing the importance of dataset quality and the suitability of advanced methods for extrovert personality detection shows the pattern of text either a person in extrovert or introvert in their behavior.

TABLE 7. Performance of deep results with word embeddings (results in %).

MBTI DATASET					
Model	**Embeddings**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
LSTM	Word2Vec	89.67	89.13	90.30	89.69
	GloVe	87.56	86.74	88.47	87.54
	Sentence Embeddings	91.50	90.85	92.39	91.54
Bi-LSTM	Word2Vec	88.63	87.84	89.47	88.59
	GloVe	89.45	88.68	90.32	89.44
	Sentence Embeddings	**92.52**	**91.69**	**93.44**	**92.49**
FRIENDSPERSONA DATASET					
Model	**Embeddings**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
LSTM	Word2Vec	81	80	78	78
	GloVe	87	87	87	87
	Sentence Embeddings	79	78	78	78
Bi-LSTM	Word2Vec	83	82	81	80
	GloVe	**89**	**88**	**88**	**88**
	Sentence Embeddings	83	82	82	82

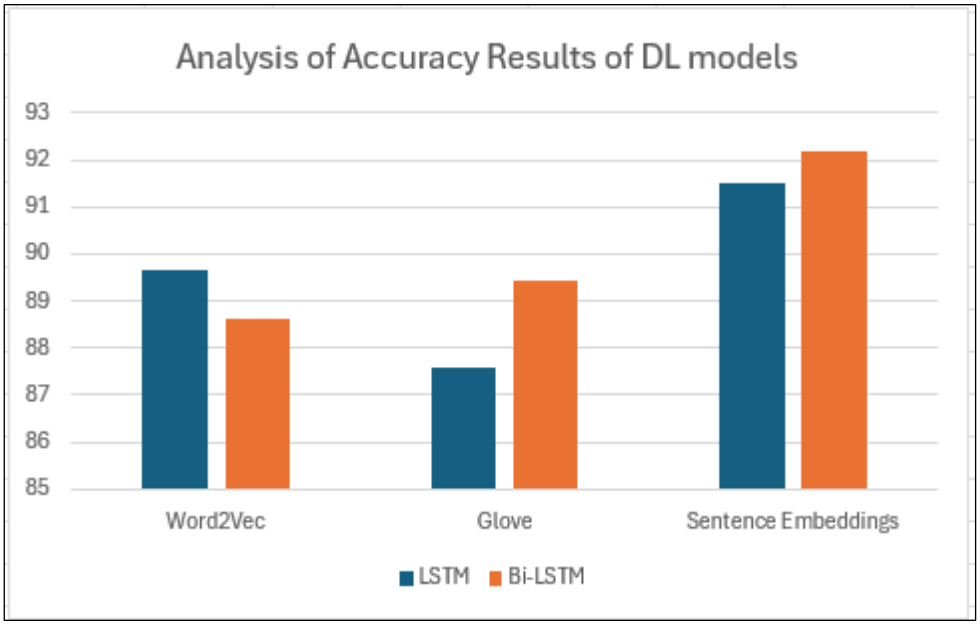

FIGURE 12. Combined accuracy analysis of proposed classifiers with word embeddings features using MBTI dataset.

E. RESULTS ON OTHER PERSONALITY TRAITS

By applying other traits also using MBTI dataset, Here are the results of mapped traits that are acquired with the help of MBTI dataset. Through analysis with the help of table 10, the SVM outperform accuracy of 89% with trait openness mapped on intuitive (future-oriented) and sensing (fact-based) labels for prediction, shows the person's willingness to embrace new experience and ideas scoring slightly low accuracy with LR achieved with trait extroversion and agreeableness with 84% respectively, with same

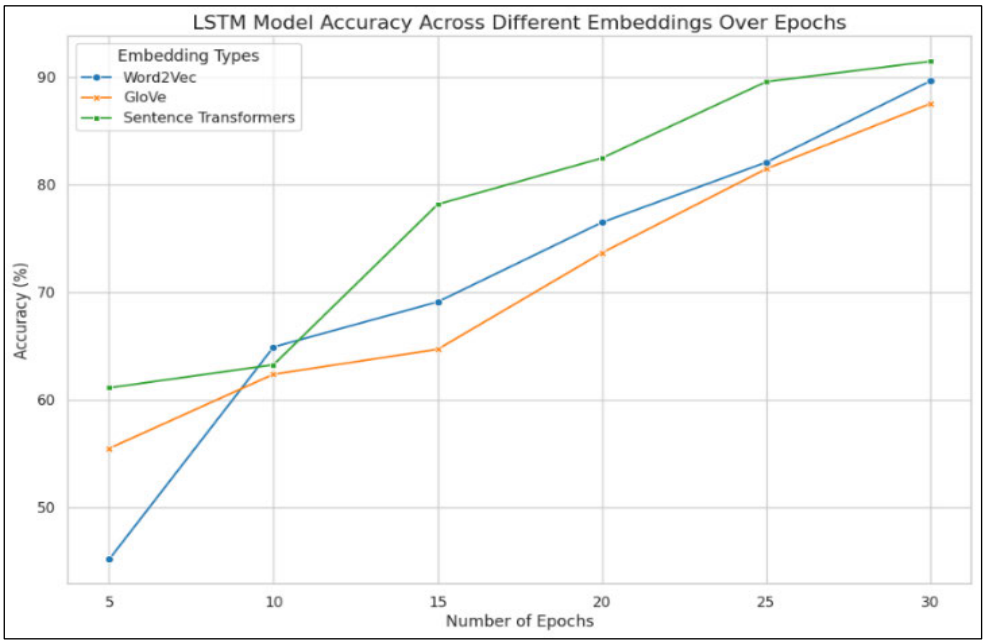

FIGURE 13. LSTM line accuracy with all applied embeddings feature.

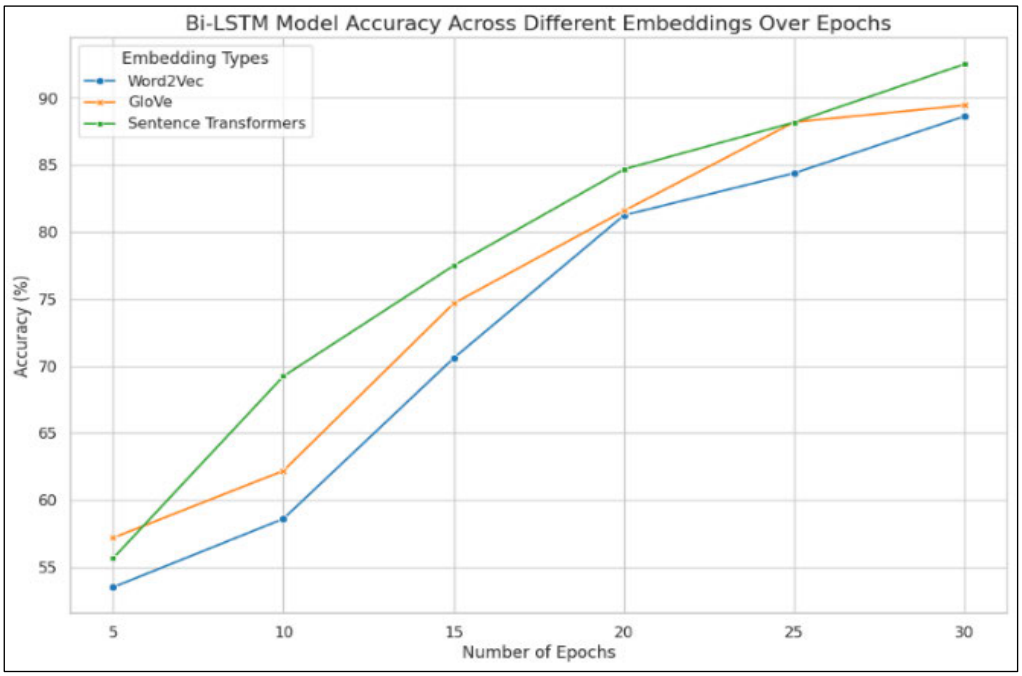

FIGURE 14. Bi-LSTM line accuracy with all applied embeddings feature.

classifier LR achieves 87% accuracy to predict openness personality features. Moreover, comprising performance with ensemble boosting model GB achieve the overall highest accuracy of 90% among openness. Overall extroversion achieves the highest results at 86% with SVM, conscientiousness mapped on judging (systematic) and perceiving (flexible) dimensions shows the trait to be organized and disciplined in task acquired the performance of 81% with GB, agreeableness trait mapped on the dichotomous of feeling (emotional) and thinking (logical) shows the behavior of person who has agreeable behavioral trait in their personality that reflect the individual's tendence to be compassionate, cooperative and friendly achieved 84% accuracy with SVM and openness accuracy is 90% with equally acquire by GB

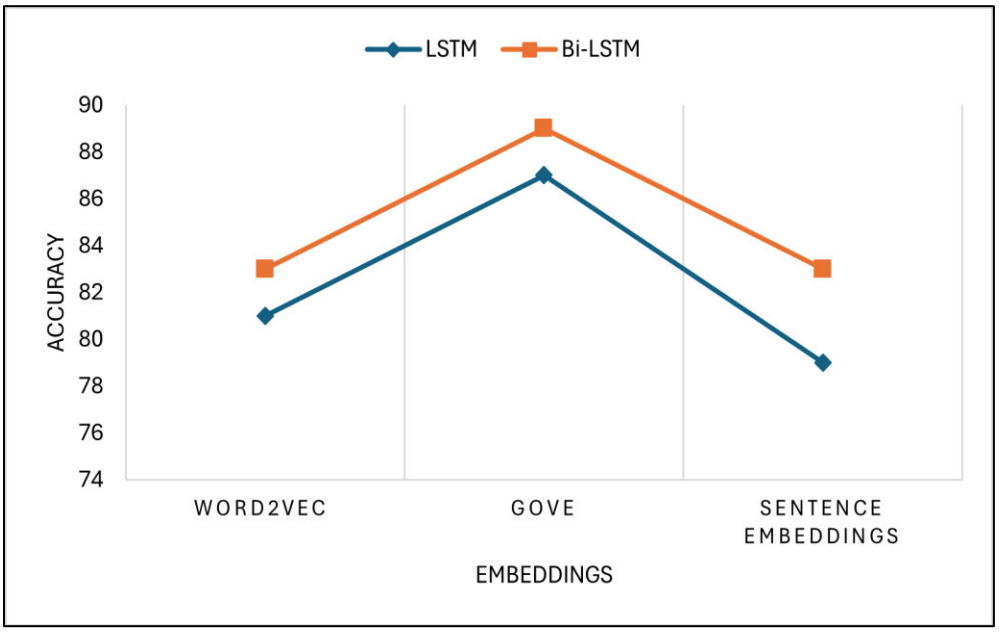

FIGURE 15. Combined comparative analysis of deep models using FriendsPersona dataset with deep embeddings.

TABLE 8. Details of bert hyperparameters.

Hyperparameter	Value	Description
vocab_size	30522	Size of embedding vocabulary
hidden_size	768	Size of hidden layers
num_hidden_layers	12	Number of transformer blocks
num_attention_heads	12	Number of attention heads in each transformer block
hidden_dropout_probs	0.1	Dropout probability for the hidden layers
attention_probs_dropout_probs	0.1	Dropout probability for the attention probabilities
intermediate_size	3072	Size of the intermediate layer, called feed-forward
hidden_act	GELU	Activation function used in the hidden layers
initializer_range	0.02	Range for weight initialization
max_position_embeddings	512	Maximum sequence length the model can accept
type_vocab_size	2	Size of the token type embeddings
num_labels	-	Number of output labels for classification tasks
layer_norm_eps	1e-12	Epsilon value for layer normalization

TABLE 9. Performance of bert model (results in %age).

Feature	Model	Dataset	Accuracy	Precision	Recall	F1-Score
Bert Encoder	Bert	**MBTI**	**82.8**	**80**	**77**	**73**
		FriendsPersona	72.5	72	70	70

and XGB with textual feature TF-IDF. On the other hand, POS tagging, another textual feature, utilizes model SVM with accuracy of 86% coupled with openness trait, which is equally comprises with GB,LR and RF, respectively. Overall extroversion achieves the highest results at 86% with GB, RF and Adaboost, conscientiousness 61% with SVM and GB.

With the help of DL algorithms, mapping of traits prediction with the help of MBTI dataset labels that direct and oppose the personality of individual to predict the characteristics of users by using social media content. In this regard deep model LSTM with word embeddings features Word2Vec achieves highest accuracy of 89.67% with trait extroversion. Another feature, GloVe with 89.39% accuracy prediction of agreeableness trait, among others. The sentence embeddings, achieving highest accuracy outperform with 91.50% to predict the trait extroversion more likely. On the other hand, Bi-LSTM with word embeddings features Word2Vec achieves highest accuracy of 88.63% with trait extroversion. Another feature, GloVe, has the highest 91.34% accuracy prediction of agreeableness trait, among others. The sentence embeddings, achieving highest accuracy outperform with 92.52% to predict the trait extroversion, as detailed results are shown in table 11.

TABLE 10. Performance of ML model on other personality traits (results in %age).

Features	Models	Extroversion	Conscientiousness	Agreeableness	Openness
TF-IDF	SVM	86	80	84	89
	NB	78	63	75	86
	LR	84	70	84	87
	DT	77	70	84	84
	KNN	79	70	70	86
	GB	86	81	82	90
	RF	79	74	79	86
	XGB	86	80	83	90
	AdaBoost	84	79	81	89
POS	SVM	76	61	63	86
	NB	70	60	58	70
	LR	78	62	62	86
	DT	64	53	54	75
	KNN	72	56	57	85
	GB	78	61	62	86
	RF	78	60	61	86
	XGB	76	56	58	85
	AdaBoost	78	60	61	78

TABLE 11. Performance of DL model on other personality traits (results in %age).

Model	Embeddings	Extroversion	Conscientiousness	Agreeableness	Openness
LSTM	Word2Vec	89.67	87.88	86.48	86.41
	GloVe	87.56	85.85	89.39	84.57
	Sentence Embeddings	91.50	89.56	90.17	88.28
Bi-LSTM	Word2Vec	88.63	86.69	85.35	85.50
	GloVe	89.45	87.38	91.34	86.27
	Sentence Embeddings	92.52	90.50	91.17	89.23
BERT	Bert Encoder	82.80	77.90	82.3	88.24

In comparison to other deep models used for predicting personality traits, the BERT encoder highlighted competitive performance, with an accuracy of 88.24% to predict the highest rate of trait measure openness. BERT demonstrated its capability to classify traits based on encoded text representation while achieving slightly lower performance compared to other models like Bi-LSTM with sentence embeddings.

F. COMPARISON OF PROPOSED WITH EXISTING STUDIES

In comparison to existing studies, the results obtained from the proposed algorithms in this research demonstrate a notable improvement in accuracy rates for predicting personality traits from social media content using MBTI dataset. Previous studies have explored various ML and DL approaches, but overall, results have achieved the high accuracy rate attained in this research, as shown in table 12. The utilization of comprehensive set of models like, SVM, DT, RF, LR, KNN, NB, GB, XGB and Adaboost as well as DL models like LSTM and Bi-LSTM and transformer-based model BERT sets this study apart from existing works. The incorporation of diverse textual features such as TF-IDF, POS tagging, and word embeddings including Word2Vec, GloVe, and sentence embeddings further enhances the predictive power of the models. Moreover, sentence embedding achieving the highest accuracy rate of 92.52% with the Bi-LSTM model underscores the efficacy of leveraging advanced deep learning architectures for personality trait prediction. In addition, Transformer-based model, which has gained significant attention in NLP task due to its contextual understanding of language, BERT is employed to capture the linguistic patterns

TABLE 12. Comparisons of all models with proposed models (results in% age).

Ref	Year	Model	Features	Results (%)
[20]	2020	SVM	TF-IDF	84
[22]	2020	XGB	TF-IDF	79
[21]	2021	SVM, NB, LR	count vectorization	84
[25]	2021	CNN-LSTM	BoW	85
[29]	2021	BERT	Sentence Embedding	72
[23]	2023	LR, SVM, NB, RF	POS tagging	67
Proposed	**2024**	**Bi-LSTM**	**Sentence Embeddings**	**92.52**

from social media for personality traits prediction with the highest accuracy rate of 82% as compared to existing studies. Overall, this significant improvement in accuracy rates demonstrates the superiority of the proposed approaches over existing methodologies as shown in table 12 and highlights its potential for advancing research in the field of personality trait prediction from social media content.

V. CONCLUSION AND FUTURE DIRECTIONS

In this research study, a detailed study is carried out on the MBTI and FriendsPersona dataset to assess the classification capability of various classifiers on traits I/E. New proposed features are proposed based on social media posts to predict personality traits. These features are used along with UGC based on posts, comments and emojis features. The main purpose of proposed features is to enhance the classification accuracy of classifiers and detect traits efficiently. The features are extracted from the selected data of personality traits named MBTI. This dataset is the latest and explored by fewer research based on social content. To evaluate the performance of ML classifiers (SVM, DT, LR, KNN), Ensemble models (GB, XGB, Adaboost, RF), transformer-based model (BERT) and DL models (LSTM and Bi-LSTM) on proposed features to find the most promising of each trait. Performance of all classifiers is measured by the evaluation metrics.

The main findings in this study compare the performance of various models for predicting personality traits. The results across various personality traits and modeling approaches exhibit distinct patterns in predictive performance. TF-IDF features consistently yield strong results across traits, highlighting their effectiveness in capturing relevant textual information for personality prediction tasks. Among the shallow machine learning models, SVM and ensemble model such as XGB frequently emerges as a top performer, particularly when integrated with TF-IDF features. Notably, SVM achieves high accuracies for I/E labels maps as extroversion, with accuracies reaching 86.05%. Additionally, the performance of Bi-LSTM model underscores the importance of exploring deep learning architecture for capturing intricate linguistic context in personality traits. However, DL models surpassed these performances, with Bi-LSTM achieving an impressive accuracy of 92% by leveraging sentence embeddings. Subsequently, Transformer based model, BERT with its encoder architecture, yielded slightly lower accuracy at 82%. This highlights the effectiveness of bidirectional architectures in capturing contextual information from text data, particularly when combined with advanced embeddings techniques like sentence transformer. While shallow machine learning models and TF-IDF often provide a strong baseline for personality predictions tasks, their inclusion in the comparison with deep models highlights their potential for further exploration in personality trait prediction tasks. Hence this study deals with too noisy and complex data in all cases of text as well as word embeddings to figure out the results accuracy reaches the highest prediction of personality traits. The consequences of this research, shows how machine and deep learning has vast promise in the collaboration of psychology domain to predict the personality traits with user generated content and cover the technique for further advancement in this exciting and vital field. These findings underscore the effectiveness of DL approaches particularly Bi-LSTM in capturing intricate patterns within the textual data, offering valuable insights for future research and practical applications in personality assessment and behavior analysis.

In this, our focus has been on exploring predictive models for personality trait. Moving forward encompasses a broader range of traits, allowing for a more comprehensive understanding of human behavior and personality dynamics. Enlarging personality prediction to languages such as Urdu and Pashto presents difficulties due to data availability, linguistic complexities, and cultural differences, how these variables affect personality prediction across languages. Majorly, English based language datasets are used to predict as high resource, but low resource data is still unavailable. Creating strong models that can be generic across several languages and cultural conditions is a significant task. Leveraging advanced model of DL models, LLMs such as XLNET, and GPT-3 an autoregressive language model presents opportunities for a more enhanced analysis of personality traits prediction in various applications such as customer personalization, recruitment, and healthcare sectors. However, challenges include ensuring domain relevance and adapting the model to the intricacies of professional language. By expanding this scope to include additional traits in the domain of personality assessment and contribute to a deeper understanding of individual differences and behavioral patterns.

REFERENCES

[1] A. Koutsoumpis, S. Ghassemi, J. K. Oostrom, D. Holtrop, W. van Breda, T. Zhang, and R. E. de Vries, "Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning," *Comput. Hum. Behav.*, vol. 154, May 2024, Art. no. 108128, doi: 10.1016/j.chb.2023.108128.

[2] S. Iqbal, R. Khan, H. U. Khan, F. K. Alarfaj, A. M. Alomair, and M. Ahmed, "Association rule analysis-based identification of influential users in the social media," *Comput., Mater. Continua*, vol. 73, no. 3, pp. 6479–6493, 2022, doi: 10.32604/cmc.2022.030881.

[3] H. U. Khan, S. Nasir, K. Nasim, D. Shabbir, and A. Mahmood, "Twitter trends: A ranking algorithm analysis on real time data," *Expert Syst. Appl.*, vol. 164, Feb. 2021, Art. no. 113990, doi: 10.1016/j.eswa.2020.113990.

[4] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," *Kuwait J. Sci.*, vol. 43, no. 4, pp. 1–19, 2016.

[5] S. M. M. R. Naqvi, S. Batool, M. Ahmed, H. U. Khan, and M. A. Shahid, "A novel approach for building domain-specific chatbots by exploring sentence transformers-based encoding," in *Proc. Int. Conf. IT Ind. Technol. (ICIT)*, vol. 10, Oct. 2023, pp. 1–7, doi: 10.1109/icit59216.2023.10335884.

[6] K. Oyibo and J. Vassileva, "The relationship between personality traits and susceptibility to social influence," *Comput. Hum. Behav.*, vol. 98, pp. 174–188, Sep. 2019, doi: 10.1016/j.chb.2019.01.032.

[7] H. U. Khan, *Modelling to Identify Influential Bloggers in the Blogosphere: A Survey*. Amsterdam, The Netherlands: Elsevier, Mar. 2017, doi: 10.1016/j.chb.2016.11.012.

[8] A. Mahmood, H. U. Khan, and M. Ramzan, "On modelling for bias-aware sentiment analysis and its impact in Twitter," *J. Web Eng.*, vol. 19, pp. 1–28, Mar. 2020, doi: 10.13052/jwe1540-9589.1911.

[9] M. K. I. Zim, M. A. Hanif, and H. Kaur, "Prediction of personality for mental health detection using hybrid deep learning model," in *Proc. IEEE Int. Conf. Interdiscipl. Approaches Technol. Manage. Social Innov. (IATMSI)*, Mar. 2024, pp. 1–6, doi: 10.1109/iatmsi60426.2024.10503423.

[10] A. S. Khan, H. Ahmad, M. Z. Asghar, F. K. Saddozai, A. Arif, and H. A. Khalid, "Personality classification from online text using machine learning approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, pp. 460–476, 2020.

[11] K. Yang, R. Y. K. Lau, and A. Abbasi, "Getting personal: A deep learning artifact for text-based measurement of personality," *Inf. Syst. Res.*, vol. 34, no. 1, pp. 194–222, Mar. 2023, doi: 10.1287/isre.2022.1111.

[12] S. I. S. Iqbal, F. K. S. Iqbal, H. U. K. F. Khan, T. I. H. U. Khan, and J. H. S. T. Iqbal, "Sentiment analysis of social media content in pashto language using deep learning algorithms," *J. Internet Technol.*, vol. 23, no. 7, pp. 1669–1677, Dec. 2022, doi: 10.53106/160792642022122307021.

[13] T. Bowden-Green, J. Hinds, and A. Joinson, "How is extraversion related to social media use? A literature review," *Personality Individual Differences*, vol. 164, Oct. 2020, Art. no. 110040, doi: 10.1016/j.paid.2020.110040.

[14] A. Bashkirova, A. Compagner, D. M. Henningsen, and J. Treur, "An adaptive modelling approach to employee burnout in the context of the big five personality traits," *Cognit. Syst. Res.*, vol. 79, pp. 109–125, Jun. 2023, doi: 10.1016/j.cogsys.2022.12.010.

[15] F. Giannini, M. Marelli, F. Stella, D. Monzani, and L. Pancani, "Surfing the OCEAN: The machine learning psycholexical approach 2.0 to detect personality traits in texts," *J. Pers.*, 2024, doi: 10.1111/jopy.12915.

[16] U. Ishfaq, H. U. Khan, and S. Iqbal, "Identifying the influential nodes in complex social networks using centrality-based approach," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 10, pp. 9376–9392, Nov. 2022, doi: 10.1016/j.jksuci.2022.09.016.

[17] W. Ahmad, H. U. Khan, T. Iqbal, M. A. Khan, U. Tariq, and J.-H. Cha, "Hybrid multichannel-based deep models using deep features for feature-oriented sentiment analysis," *Sustainability*, vol. 15, no. 9, p. 7213, Apr. 2023, doi: 10.3390/su15097213.

[18] P. Sánchez-Fernández, L. G. B. Ruiz, and M. D. C. P. Jiménez, "Application of classical and advanced machine learning models to predict personality on social media," *Expert Syst. Appl.*, vol. 216, Apr. 2023, Art. no. 119498, doi: 10.1016/j.eswa.2022.119498.

[19] H. Samota, S. Sharma, H. Khan, M. Malathy, G. Singh, and R. Rambabu, "A novel approach to predicting personality behaviour from social media data using deep learning," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 15s, pp. 539–547, 2024. [Online]. Available: https://www.ijisae.org

[20] C. So, "Are you an introvert or extrovert? Accurate classification with only ten predictors," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Feb. 2020, pp. 693–696, doi: 10.1109/ICAIIC48513.2020.9065069.

[21] B. Fieri, J. La'la, and D. Suhartono, "Introversion-extraversion prediction using machine learning," *Int. J. Inform. Vis.*, vol. 7, no. 4, pp. 2154–2160, 2023.

[22] P. and S. A. and B. A. Bhardwaj Harshit and Tomar, "Classification of extraversion and introversion personality trait using electroencephalogram signals," in *Artificial Intelligence and Sustainable Computing for Smart City*. Cham, Switzerland: Springer, 2021, pp. 31–39.

[23] M. N. Sahono, F. U. Sidiastahta, G. F. Shidik, A. Z. Fanani, Muljono, S. Nuraisha, and E. Lutfina, "Extrovert and introvert classification based on Myers-Briggs type Indicator(MBTI) using support vector machine (SVM)," in *Proc. Int. Seminar Appl. Technol. Inf. Commun. (iSemantic)*, Sep. 2020, pp. 572–577, doi: 10.1109/iSemantic50169.2020.9234288.

[24] K. Orynbekova, A. Talasbek, A. Omar, A. Bogdanchikov, and S. Kadyrov, "MBTI personality classification using apache spark," in *Proc. 16th Int. Conf. Electron. Comput. Comput. (ICECCO)*, Nov. 2021, pp. 1–4, doi: 10.1109/ICECCO53203.2021.9663858.

[25] N. Cerkez and V. Vareškic, "Machine learning approaches to personality classification on imbalanced MBTI datasets," in *Proc. 44th Int. Conv. Inf., Commun. Electron. Technol. (MIPRO)*, Sep. 2021, pp. 1259–1264, doi: 10.23919/MIPRO52101.2021.9596742.

[26] P. Kumar R, B. Mohan G, and G. D. Sai, "Ensemble machine learning models in predicting personality traits and insights using myers-briggs dataset," in *Proc. Int. Conf. Adv. Comput., Commun. Appl. Informat. (ACCAI)*, vol. 528, May 2023, pp. 1–7, doi: 10.1109/accai58221.2023.10199294.

[27] J. Serrano-Guerrero, B. Alshouha, M. Bani-Doumi, F. Chiclana, F. P. Romero, and J. A. Olivas, "Combining machine learning algorithms for personality trait prediction," *Egyptian Informat. J.*, vol. 25, Mar. 2024, Art. no. 100439, doi: 10.1016/j.eij.2024.100439.

[28] M. K. Hayat, A. Daud, A. A. Alshdadi, A. Banjar, R. A. Abbasi, Y. Bao, and H. Dawood, "Towards deep learning prospects: Insights for social media analytics," *IEEE Access*, vol. 7, pp. 36958–36979, 2019, doi: 10.1109/ACCESS.2019.2905101.

[29] M. A. Teli and M. A. Chachoo, "Pre-trained word embeddings in deep multi-label personality classification of Youtube transliterations," in *Proc. Int. Conf. Intell. Syst., Adv. Comput. Commun. (ISACC)*, Feb. 2023, pp. 1–6, doi: 10.1109/ISACC56298.2023.10084047.

[30] H. Ahmad, M. U. Asghar, M. Z. Asghar, A. Khan, and A. H. Mosavi, "A hybrid deep learning technique for personality trait classification from text," *IEEE Access*, vol. 9, pp. 146214–146232, 2021, doi: 10.1109/ACCESS.2021.3121791.

[31] H. Bhardwaj, P. Tomar, A. Sakalle, D. Acharya, T. Badal, and A. Bhardwaj, "A DeepLSTM model for personality traits classification using EEG signals," *IETE J. Res.*, vol. 69, no. 10, pp. 7272–7280, Oct. 2023, doi: 10.1080/03772063.2021.2012278.

[32] J. Zhao, J. Lin, S. Liang, and M. Wang, "Sentimental prediction model of personality based on CNN-LSTM in a social media environment," *J. Intell. Fuzzy Syst.*, vol. 40, no. 2, pp. 3097–3106, Feb. 2021, doi: 10.3233/jifs-189348.

[33] J. J. Sirasapalli and R. M. Malla, "A deep learning approach to text-based personality prediction using multiple data sources mapping," *Neural Comput. Appl.*, vol. 35, no. 28, pp. 20619–20630, Oct. 2023, doi: 10.1007/s00521-023-08846-w.

[34] Z. Ren, Q. Shen, X. Diao, and H. Xu, "A sentiment-aware deep learning approach for personality detection from text," *Inf. Process. Manage.*, vol. 58, no. 3, May 2021, Art. no. 102532, doi: 10.1016/j.ipm.2021.102532.

[35] F. Elourajini and E. Aïmeur, "AWS-EP: A multi-task prediction approach for MBTI/Big5 personality tests," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2022, pp. 1–8, doi: 10.1109/ICDMW58026.2022.00049.

[36] F. Yang, X. Quan, Y. Yang, and J. Yu, "Multi-document transformer for personality detection," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 16, pp. 14221–14229. [Online]. Available: www.aaai.org

[37] K. El-Demerdash, R. A. El-Khoribi, M. A. I. Shoman, and S. Abdou, "Deep learning based fusion strategies for personality prediction," *Egyptian Informat. J.*, vol. 23, no. 1, pp. 47–53, Mar. 2022, doi: 10.1016/j.eij.2021.05.004.

[38] M. A. Akber, T. Ferdousi, R. Ahmed, R. Asfara, and R. Rab, "Personality prediction based on contextual feature embedding SBERT," in *Proc. IEEE Region Symp. (TENSYMP)*, 2023, pp. 1–5, doi: 10.1109/TENSYMP55890.2023.10223609.

[39] T. Wang, P. Ye, H. Lv, W. Gong, H. Lu, and F.-Y. Wang, "Modeling digital personality: A fuzzy-logic-based myers–briggs type indicator for fine-grained analytics of digital human," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 1, pp. 1096–1107, 2023, doi: 10.1109/TCSS.2023.3245127.

[40] W. Khan, A. Daud, K. Khan, J. A. Nasir, M. Basheri, N. Aljohani, and F. S. Alotaibi, "Part of speech tagging in urdu: Comparison of machine and deep learning approaches," *IEEE Access*, vol. 7, pp. 38918–38936, 2019, doi: 10.1109/ACCESS.2019.2897327.

[41] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Natural Lang. Process. J.*, vol. 4, Sep. 2023, Art. no. 100026, doi: 10.1016/j.nlp.2023.100026.

[42] S. Kazi, S. Khoja, and A. Daud, "A survey of deep learning techniques for machine reading comprehension," *Artif. Intell. Rev.*, vol. 56, no. S2, pp. 2509–2569, Nov. 2023, doi: 10.1007/s10462-023-10583-4.

ANAM NAZ received the B.S.C.S. degree from the University of Sargodha, Pakistan, where she is currently pursuing the M.S. degree in computer science. Previously, she has conducted research on the semantic web, specifically focusing on DBpedia, to advance data integration and retrieval capabilities. Her research interests include machine learning, deep learning, and natural language processing.

HIKMAT ULLAH KHAN received the master's degree in computer science and the Ph.D. degree in computer science from International Islamic University, Islamabad. He has been an Active Researcher for the last ten years. He is currently a Professor and the Chairperson of the Department of Information Technology, University of Sargodha, Pakistan. He has authored more than 70 research articles in top peer-reviewed journals and international conferences. His research interests include social web mining, semantic web, data science, information retrieval, and scientometrics. He is an editorial board member of a number of prestigious impact factor journals.

SAMI ALESAWI received the Bachelor of Science degree in computer engineering and the Master of Science degree in computer science from King Abdulaziz University, Jeddah, Saudi Arabia, and the Doctor of Philosophy degree from The University of Texas at Arlington, Arlington, TX, USA. He is a highly accomplished individual within the realm of computer engineering and computer science. He holds the esteemed position of an Assistant Professor with the Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University. In this capacity, he actively partakes in a myriad of academic pursuits and is actively involved in his ongoing educational and research interests include games, computer graphics, computer vision, human–computer interaction, and AI.

OMAR IBRAHIM ABOUOLA received the Bachelor of Science degree in computer science from KAU, in 2001, the master's degree in information science from the University of Indiana, Bloomington, USA, in 2009, and the Ph.D. degree in information systems and technology from Claremont Graduate University (CGU), USA, in 2018. He is currently an Assistant Professor with the Information Systems and Technology Department, College of Computer Science and Engineering (CCSE), University of Jeddah. His master's thesis was related to the technology of banking. His Ph.D. dissertation aimed to design an innovative assistive technology to help retail companies to predict optimum locations for their businesses.

ALI DAUD received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in July 2010. Currently, he is a Full Professor with the Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates. He has 13 years' post-Ph.D. experience of teaching, supervision, and research at B.S., M.S., and Ph.D. level. He has published more than 100 research papers in reputed international impact factor journals and conferences. He has taken part in many research projects as well and have written and acquired many research funding's. He has proven and extensive experience in data mining, artificial intelligence (machine learning/deep learning) applications to social networks, data science, natural language processing, and the Internet of Things.

MUHAMMAD RAMZAN received the Ph.D. degree in CS from the University of Management & Technology, Lahore, Pakistan. He is currently a Lecturer with the CS and IT Department, University of Sargodha. Previously, he was a Lecturer with the Virtual University of Pakistan. Also, he was with the Higher Education Department as a Database/Network Administrator. Before this, he was a Lecturer with Minhaj University Lahore (Sharia College). He has authored more than 40 research articles published in reputed peer-reviewed journals. His research interests include medical imaging, deep learning, machine learning, video surveillance, activity recognition, and computer vision.