IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 4 March 2025, accepted 3 April 2025, date of publication 8 April 2025, date of current version 21 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3558714

RESEARCH ARTICLE

Convolutional Bi-LSTM for Automatic Personality Recognition From Social Media Texts

NOUF M. ALJUHANI[1], ABDULLAH AL-MALAISE AL-GHAMDI[1,2], HANAN S. ALGHAMDI[1], AND FARRUKH SALEEM[3]

[1]Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Information Systems Department, School of Engineering, Computing and Design, Dar Al-Hekma University, Jeddah 22246, Saudi Arabia
[3]School of Built Environment, Engineering, and Computing, Leeds Beckett University, LS6 3QR Leeds, U.K.

Corresponding authors: Nouf M. Aljuhani (nhamdanaljuhani@stu.kau.edu.sa) and Farrukh Saleem (f.saleem@leedsbeckett.ac.uk)

ABSTRACT Automatic Personality Recognition (APR) has received much attention in recent years due to its wide range of important applications across various fields. The growing use of online social networks provides valuable opportunities for APR, as a strong correlation has been found between what users post on these platforms and their personality traits. Consequently, various APR models have been developed to infer the Big Five personality traits from social media user-generated texts. However, most of these models heavily relied on hand-crafted features, which are unable to capture deep contextual information and learn complex patterns from texts. More importantly, the performance of text-based APR is still unsatisfactory, especially at the level of each personality dimension. To tackle this issue, we propose a new model, called APR_ConvLSTM, that aims to improve text-based APR performance by integrating two robust deep learning architectures: CNN and Bi-LSTM. Unlike existing APR models, the APR_ConvLSTM is a unified end-to-end model where all personality traits are predicted simultaneously and effectively without a need for laborious feature engineering. We also developed a new labeled Big Five personality dataset, called X-Big5, which has been in need for a long time in the APR field. Extensive experiments on the X-Big5 and a publicly available benchmark dataset (PAN-2015 Author Profiling) demonstrate the promising performance of our model over its contenders. Overall, the proposed model achieved the highest Accuracy and F-1 score of 79.51% and 86.54% on the PAN-2015 dataset and 87.95% and 81.35%, respectively, on the X-Big5 dataset. Moreover, it shows promising performance over its competitors, with the highest average Accuracy and F-1 score of 79.01% and 80.56%, respectively, on the combined dataset. The model reached competitive results in predicting Openness, Extraversion, Agreeableness, and Neuroticism traits with the highest F1 scores of 88.60%, 77.35%, 76.16%, and 74.52%, respectively, on the combined dataset. The proposed model can positively impact the analysis of social media text generated by different users and help identify their personality traits.

INDEX TERMS Automatic personality recognition, big five personality traits, convolutional neural networks, long-short term memory, social media texts.

I. INTRODUCTION

Personality is one of the fundamental concepts in the psychology field, profoundly affecting human behavior and distinguishing individuals from each other. Throughout the long history of personality psychology, which is traced back to 1937, different theories and models have emerged to understand and conceptualize the structure of personality [1]. Some of the common examples of such models include Allport's Trait model [2], Myers-Briggs Type Indicator (MBTI) [3], HEXACO model [4], Psychoticism, Extraversion, and Neuroticism (PEN) model [5], and Big Five model [6]. However, the Big Five model, also known as the Five-Factor model, is considered the most comprehensive and widely

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa M. Fouda.

TABLE 1. Big five personality traits.

Trait	High	Low
Extraversion (EXT)	Extraverts: outgoing, active, talkative, assertive, enthusiastic, and gregarious.	Introverted: solitary, quiet, serious, prefers to be alone or stay within close circles.
Conscientiousness (CON)	Efficient, organized, planful, responsible, productive, dependable, careful, self-disciplined, ambitious, and persistent.	Unconscientious: sloppy, careless, spontaneous, easy-going, inefficient, and disorganized.
Openness (OPN)	Imaginative, adventurous, creative, curious, inventive, tend to change, have an excellent aesthetic sense, and have unusual ideas and broad interests.	Conventional, dogmatic, cautious, unimaginative, and conservative.
Agreeableness (AGR)	Agreeable: trustworthy, generous, modest, forgiving, respectful, cooperative, tend to maintain positive relationships with others and avoid conflicts.	Disagreeable: unreliable, complicated, antagonistic towards other people, tends to find faults, and is less likely to compromise.
Neuroticism (NEU)	Neurotics: anxious, impulsive, nervous, sensitive, depressed, and tend to guilt, sadness, and self-pitying.	Emotional Stability: calmer, confident, comfortable, and self-controlled.

accepted model in all psychology. This model assumes that an individual's personality can be measured by considering only five main traits, also called personality dimensions, which are Extraversion, Conscientiousness, Openness, Agreeableness, and Neuroticism [see Table 1].

Indeed, the role of personality is not only limited to explaining one's behavior but also extends to affect dozens of human life outcomes such as interpersonal skills and social relations [7], [8], job performance [9], [10], job satisfaction [11], choices [12], preferences [13], physical health [14], mental and psychological disorders [15], [16]. Additionally, being able to effectively detect a user's Big Five personality traits has been found to significantly benefit a broad range of academic and industrial applications including: marketing strategies [17], job screening [18], social networks analysis [19], recommender systems [20], sentiment analysis [21], specialized healthcare and guiding [22], political forecasting [23], to name a few.

Thus, Automatic Personality Recognition (APR), one of the basic tasks in Personality Computing [24], has attracted wide attention in recent years, especially with the growing use of social network platforms. In particular, the proliferation of User-Generated Content (UGC) and self-disclosed information on these platforms provide valuable opportunities for the APR [25], [26], [27], [28], [29]. Among various types of valuable content, textual data has been repeatedly found to significantly correlate and indicate a user's personality traits. For example, Tadesse et al. showed that users with high extraversion tend to use more positive emotional words and 2nd and 3rd person-singular pronouns, while those with high neuroticism frequently use negative emotion words and singulars first-person pronouns (e.g., I, mine, me) [30]. Yuan et al. found that conscientiousness has a negative relationship with the words corresponding with the cognitive process, such as insight or self-reflection words (e.g., know and realize) and certainly words (e.g., absolute and always) [31].

Consequently, APR based on social media users' texts has received great research interest in the last decade [30], [31], [32], [33], [34], [35], [36]. The main concern of this task is to effectively infer the underlying personality dimensions of users from their generated texts. According to recent surveys [25], [37], existing text-based APR methods can mainly be classified under two categories: machine learning-based methods and deep learning-based methods. Most prior works [e.g., [30], [32], [33], [36], [38] are based on machine-learning methods where hand-crafted features such as bag-of-words, lexical features, stylistic features, and other statistical features are used along with classical machine-learning classifiers such as Support Vector Machines (SVM), Naïve Bayes (NB), Logistic Regression (LR). However, the performance of this approach heavily relied on finding the optimal synthesis of hand-crafted features that can explain and predict personality, which is labor- and time-intensive. Additionally, shallow text-based APR models are unable, to some extent, to capture deep contextual information and complex patterns from texts, making them unable to handle common textual- issues (e.g. polysemy).

Recently, deep learning-based models have achieved remarkable performance improvement in different fields, such as health Informatics [39], computer vision [40], and Natural Language Processing (NLP) [41]. Convolutional Neural Networks (CNNs) [42] and Recurrent Neural Networks (RNNs) [43] are the most prevailing types of deep learning networks in the NLP context. CNNs, which are mainly used in computer vision, achieved a good performance in extracting meaningful local features by moving filters (i.e., kernels) over different text regions. On the other hand, RNNs have been proposed to handle time-sequential data, such as texts, by using a recurrent hidden state that captures and passes useful contextual information between inputs. Although RNNs are suitable for different NLP applications, they suffer from vanishing and exploding gradients when processing long sequences, making them unable to capture long-term dependencies between inputs [44]. Alternatively, Long-Short Term Memory (LSTM) [45], one of the best variants of RNNs, is designed to eliminate the risk of gradient disappearance by adding a series of memory cells and gate units.

Despite the impressive success of deep learning methods with different NLP tasks, utilizing them to improve the performance of text-based APR has gained little research attention. The existing efforts are often based on a single deep-learning architecture [31], [46], [47], [48], [49], [50], which alone cannot fully benefit from rich textual information posted on social media platforms. For example, Yu et al. employed CNN with multiple convolutional filters to extract local patterns from Facebook status [50]. However, this CNN architecture cannot learn sequential correlation and contextual dependencies between input data. Additionally, and more importantly, the performance of APR from social media texts is still unsatisfactory, especially at the level of each personality dimension.

To tackle the above issues, a new text-based APR model, called APR_ConvLSTM, has been developed, aiming to improve APR performance by integrating the individual merits of both CNN and LSTM networks. The choice of CNN and Bi-LSTM architecture was motivated by their complementary strengths and individual success achieved with different NLP tasks. Specifically, we used the CNN layer to extract informative patterns (i.e., n-gram[1] features) and reduce the dimensionality of feature maps. Then, feature sequences obtained from the CNN will be fed into a bidirectional-LSTM (Bi-LSTM) layer to learn semantic textual features and long-term dependencies between inputs. Here, Bi-LSTM is used instead of the standard unidirectional LSTM due to its ability to access both the preceding and succeeding context information of input sequences [51]. Notably, most APR approaches treat personality dimensions separately by training a model for each dimension. Unlike such approaches, the APR_ConvLSTM is designed to be a unified end-to-end APR model by decomposing the output layer into five independent parallel layers in which all personality dimensions are predicted simultaneously. Moreover, the APR_ConvLSTM model has a worthwhile advantage as it is free from the feature engineering process, which is a labor- and time-intensive task.

In addition, APR from social media texts lacks the availability of well-annotated benchmark datasets. To date, only two textual datasets have been publicly available for academic use: the MBTI dataset [52], which is annotated according to the MBTI personality theory's dimensions, and the PAN-2015 Author Profiling dataset [53], which is annotated with the Big Five personality dimensions. Therefore, in this work, we collected a new real-world personality dataset from the X platform, one of the most popular social media channels worldwide, namely the X-Big5 dataset. The dataset comprises a set of users' tweets labeled with their ground-truth Big Five personality dimensions.

Extensive experiments have been conducted on the PAN-2015, X-Big5 dataset, and a combined set of them. The results showed that the proposed model APR_ConvLSTM can effectively improve APR performance at each personality dimension and significantly outperform other baseline methods.

[1]N-grams are contiguous sequences of n words from a given text.

A. RESEARCH OBJECTIVES

The main objective of this work is to propose a new APR model, named the APR_ConvLSTM, aiming at improving the performance of the APR from social media user-generated texts and reducing the dependence on feature engineering. To achieve this main objective, a set of sub-objectives will be addressed in this work as follows:

1) Building a unified end-to-end deep learning-based APR model to effectively and simultaneously infer the Big Five personality traits from social media texts.
2) Collecting a new real-world dataset for the text-based APR field.
3) Evaluating the performance of our model in predicting each personality trait (i.e., openness, conscientiousness, extraversion, agreeableness, and neuroticism) in terms of four evaluation metrics: accuracy, precision, recall, and F1-score.
4) Comparing the performance of the proposed model against traditional machine-learning methods.
5) Comparing the performance of the proposed model with state-of-the-art deep learning-based APR models and evaluating if the improvements achieved are statistically significant using a statistical t-test analysis.

B. RESEARCH QUESTIONS

Based on the above discussion, this research mainly focuses on answering the following research questions:

RQ1: Does the proposed APR_ConvLSTM outperform traditional machine-learning methods in predicting the Big Five personality traits from social media text?

RQ2: Can the proposed APR_ConvLSTM model improve the performance of the APR from social media texts in terms of each personality dimension, compared to the state-of-the-art deep-learning APR models?

C. RESEARCH CONTRIBUTIONS

This research makes the following contributions to this field of study:

1) We devise a new APR deep network model APR_ConvLSTM, which integrates CNN and Bi-LSTM architectures, to effectively predict the Big Five personality traits from social media users' texts. The APR_ConvLSTM has two main advantages over existing APR models: a) It is a unified end-to-end model where all the personality dimensions are predicted simultaneously and independently from each other instead of training multiple APR models, one for each dimension. b) It learns contextual features and complex linguistic patterns from texts without a need for a laborious feature engineering process.
2) Introducing a real-world personality dataset for the text-based APR task composed of a set of social media

users' posts annotated with their Big Five personality dimensions.

3) We conducted extensive experiments on two real-world datasets and a combined version of them, which clearly show the competitive performance of the APR_ConvLSTM in predicting each of the Big Five personality dimensions.
4) We provide a comparative performance analysis between the APR_ConvLSTM and state-of-the-art APR baselines, demonstrating the outperformance of our model against other APR models.

To summarize, this work aims to address two main research gaps in the APR field:

1) The poor performance of the APR from social media texts, especially in predicting each personality dimension.
2) The paucity of real-world social media datasets that are annotated with the Big Five personality traits.

The remainder of this paper is organized as follows: Section II presents an overview of previous works in the domain of APR from social media content. Section III presents methods for collecting, labeling, and preparing the X-Big5 dataset. Section IV explains the proposed APR_ConvLSTM model in detail. In Section V, we present a series of four experimental analyses of our model on two real-world datasets and a combined set of them. In addition, the final discussion presented in Section VI highlights the outcomes of this study based on the research questions. Finally, we conclude this study in Section VII.

II. RELATED WORK

In the last decade, there has been a growing increase in APR works that infer users' personality traits from their generated texts on online social networks. Existing APR models are mainly classified into two categories: machine learning-based and deep learning-based methods. However, according to our review, we noticed that there is a small set of studies that attempted to combine more than one deep-learning architecture, such as CNN, LSTM, or GRU, for the APR task. Therefore, in this section, we classified prior ARP works into three categories: machine learning-based APR (Section II-A), deep learning-based APR (Section II-B), and hybrid deep learning-based APR (Section II-C). Since interest in APR from social media texts has increased in the last five years [48], [49], and to ensure that recent advances in text-based APR are covered, we limited our review to studies published after 2014.

A. MACHINE-LEARNING BASED APR

Most existing APR studies depend on machine learning methods such as SVM, Random Forest, Logistic Regression, and Naïve Bayes (NB) to predict users' personality traits from their social media behavior [26], [30], [32], [36], [38], [54], [55], [56], [57]. The performance of machine-learning models mainly depends on the quality of the handcrafted features extracted from texts. Farnadi et al. provided comparative analysis work to find the optimal combination of features that works well with the APR regardless of dataset sources [26]. Unfortunately, they concluded that it was not possible to find a general set of features suitable for all varieties of social networks' datasets. Tadesse et al. combined lexical linguistics features with Social Network Analysis features (e.g., network size, betweenness, density), and fed them into a Gradient-Boosted Decision Tree (XGBoost) [30]. They found that the XGBoost outperformed all other baseline methods (i.e., SVM, Logistic Regression, and Gradient Boosting) for all Big Five traits with an average accuracy score of 74.2%. Considering only linguistics indicators, Carducci et al. used the SVM to classify the Big Five personality traits of 18,473 tweets collected from 24 Twitter users [36]. Their approach was mainly based on the transfer learning mechanism in which the predictive model was trained on 9913 Facebook status updates and tested on the Twitter dataset. Their proposed approach predicts Twitter users' Big Five traits with a lower Mean Squared Error (MSE) than the one suggested by [58].

Although most pertaining studies focus on predicting the Big Five personality traits, KN et al. put effort into building an APR model capable of estimating both Big Five and MBTI personality dimensions [38]. First, the authors developed a novel input representation method composed of linguistic features (e.g., word and phrase Frequency, Co-occurrence of words, Contextual meaning of a word) and stylistic features (e.g., the count of hashtags, the average word length of a post). Then, tweet representations are inputted to an ensemble machine learning method comprising Linear SVM and XGBoost. The most exciting finding is that their proposed input representation mechanism achieved a superior performance against other prevalent representation methods, such as one-hot encoding and count-based vectorization.

B. DEEP LEARNING-BASED APR

The tremendous achievements of deep learning methods in different NLP areas, such as sentiment analysis [59], sarcasm detection [60], and emotion analysis [61], urged personality researchers to begin utilizing these methods for APR, replacing traditional machine learning approaches. Some existing deep-learning-based studies adopted the concept of transfer learning in which pre-trained language models, such as BERT, RoBERTa, XLNet, and FastText, have been adopted to learn semantic representations of a user's texts. For example, Ren et al. proposed an APR model that uses BERT to dynamically generate sentence-level embeddings of two users' text datasets (MBTI dataset and stream-of-consciousness essays), which are then combined with the sentiment polarity of texts [49]. Among three neural networks (i.e., CNN, GRU, and LSTM), the CNN architecture achieved better performance results with average accuracy improvements of 6.91% and 6.04% on MBTI and Big Five personality datasets, respectively. In the same fashion, Christian et al. developed their model based on the three pre-trained models, BERT, RoBERT, and XLNet [47]. The semantic representations are then passed to the self-attention mechanism, allowing the model to focus on important words contributing to the APR. Kosan et al. conducted a study to perform a structural analysis of 1,769,202 tweets collected from 5081 users, which are embedded with the FastText and then fed into a Bi-LSTM layer [48]. The main aim of their study was to investigate the effect of different text pre-processing methods on the performance of the APR task.

Instead of relying on pre-trained language models, Xue et al. developed a new semantic-enhanced APR model based on bidirectional GRU (Bi-GRU) to deeply learn semantic features from essays and YouTube transcriptions datasets [46]. Yuan used a CNN layer composed of three convolutional filters to extract n-gram features (1,2,3), which are then combined with LIWC features and sent to fully connected layers [30]. However, compared with other classifiers, the model only outperforms in terms of the Openness personality trait. The details of selected models and their performance is depicted in Table 2.

C. HYBRID DEEP LEARNING-BASED APR

Recently, there have been few attempts in personality literature to combine two or more deep learning architectures to improve the APR performance. For example, Xue et al. designed a novel hierarchical deep neural network model based on RNN, attention mechanism, and CNN, referred to as AttRCNN-CNNs, to extract semantic and syntactic features from each user's text posts [62]. Although incorporating the semantic features extracted from their proposed model into different regression algorithms achieved lower MAE values, the study did not conclude with a complete end-to-end APR solution. Sun et al. introduced a model that concatenates Bi-LSTMs and CNN layers to learn Latent Sentence Group features, i.e., a set of successive sentences strongly connected in logic and semantic structure, from text structure [63]. Their results demonstrated that combining LSTM and CNN can effectively improve the performance of APR rather than adopting them separately. In the opposite way, Ahmad et al. designed a model based on MBTI personality theory in which CNN layers come at the first component of the model, followed by a unidirectional LSTM layer [64]. Recently, Zhao et al. proposed an attention-based LSTM APR model that combines thematic (i.e., topic preferences), sentiment, and contextual features extracted from Facebook users' status updates [65].

Although some progress has been made, the performance of APR from social media texts is still unsatisfactory, and the task is still in its early stages, especially at the level of each personality dimension. Most ARP models are based on hand-crafted features, which have limited representation ability and cannot capture complex dependencies between input data. In this work, we aim to improve the performance of the text-based APR in terms of each personality dimension and reduce the burden associated with feature engineering. To this end, we proposed a deep learning based-APR model APR_ConvLSTM, that combines the advantages of two robust deep neural networks, CNN, and Bi-LSTM. CNN is adopted due to its ability to extract informative local features, while the Bi-LSTM to learn preceding and succeeding long-term dependencies between inputs. In contrast to previous models, our APR_ConvLSTM is a unified trainable end-to-end APR model in which all personality dimensions are inferred simultaneously without a need for cumbersome feature engineering.

III. METHOD

The lack of publicly available real-world personality datasets represents a significant obstacle in the APR domain. Therefore, we introduced a new personality dataset based on the Big Five personality taxonomy. This section details the methods used to collect, label, and prepare the dataset to use with the model experimentation and validation.

A. DATA COLLECTION

In this work, we developed a new real-world personality dataset, namely X-Big5, composed of a set of users' tweets annotated with their ground-truth Big Five personality traits. We selected the X platform, formerly known as Twitter, because it is one of the most popular and widely used social media channels worldwide, with more than 368 million active users registered in 2022 [66]. According to [34] and [67], X is an effective platform for studying personality, as users can share their views and opinions and express themselves through tweets without hesitation or shame.

There are two main approaches adopted in literature to collect a personality corpus. The first one, the classical approach, requires participants to fill up at least one of the personality questionnaires, such as the Big Five Inventory (BFI, 44 items) and International Personality Item Poll (IPIP, 300 items), to calculate their personality scores, which are then used for labeling their texts [32], [36], [68]. In this approach, participants are volunteered to answer personality tests and share their social media posts for research purposes. The second approach is based on the anonymous collection of social media users' posts, which are then labeled by a group of domain experts [47], [57], [69].

In this work, we faced difficulty hiring enough psychological experts since manual data annotation is costly and time-consuming. Thus, we followed the classical approach by developing an online questionnaire based on the Big Five Inventory-2 (BFI-2) [70], a major revision of the original Big Five Inventory (BFI) [71]. In this inventory, each personality trait is measured using a balanced number of true-keyed and false-keyed items to mitigate the effect of response biases such as acquiescence. To avoid participant fatigue and careless responses [72], we allowed participants to either fill up the complete set of the BFI-2 (60 items) or the short form of the BFI-2 (BFI-2-S) [73], which consists

TABLE 2. Comparison of models for personality prediction.

Ref	Features	Models	Dataset	Model's Performance	Limitations
[48]	Word embeddings using the FastText model	Bi-LSTM	1,769,202 tweets from 5,081 users.	RMSE (**EXT**: 0.143, **CON**: 0.211, **OPN**: 0.137, **AGR**: 0.190, **NEU**: 0.159)	• Not enough comparative analysis (only compared with FastText model). • Single evaluation metric.
[31]	- LIWC features - N-grams	CNN & Two FC layers	A subset of myPersonality dataset: 9,900 status updates of 250 users.	Accuracy (**EXT**: 57%, **CON**: 58%, **OPN**: 76%, **AGR**: 57%, **NEU**: 60%)	• Outperformed other methods only for Openness. • Multiple classifiers (handling traits separately). • Relatively small dataset.
[47]	- TF-IGM - Sentiment - Lexical features	BERT, RoBERTa, XLNet & Attention Mechanism	myPersonality dataset & Twitter dataset	MyPersonality dataset: F-1 score (**EXT**: 74 .8 %, **CON**: 65.2%, **OPN**: 91.2%, **AGR**: 69.0%, **NEU**: 70.9%) Twitter dataset: F-1 Score (**EXT**: 88.2%, **CON**: 73.6%, **OPN**: 74.0%, **AGR**: 73.4%, **NEU**: 69.4%)	• Underperformed in Agreeableness. • Multiple classifiers.
[50]	Word embeddings & Social Network Analysis	CNN, RNN, FC Networks	A subset of myPersonality dataset: 9,917 status updates	CNN with pooling achieved 60.0 ± 6.5% F1-score	• Relatively small dataset. • Performance is not well-satisfying.
[49]	- Sentiment polarity - BERT embeddings	BERT, CNN, GRU, LSTM	MBTI dataset: 50 posts by 8,675 volunteers, Stream-of-consciousness dataset	Accuracy of Big Five Traits (**EXT**: 79.94%, **CON**: 80.23%, **OPN**: 80.35%, **AGR**: 80.30%, **NEU**: 80.14%)	• Small dataset. • Single evaluation metric.
[46]	Semantic features using Bi-GRU	Bi-GRU, FC Layer	Stream-of-consciousness dataset: 2,467 essays, YouTube dataset: 404 vlogs	Accuracy on YouTube dataset (**EXT**: 70.73%, **CON**: 65.85%, **OPN**: 65.85%, **AGR**: 73.17%, **NEU**: 78.05%)	• Small social media dataset. • Performance is not well-satisfying.

OPN = Openness, CON = Conscientiousness, EXT = Extraversion, AGR = Agreeableness, NEU = Neuroticism
LIWC = Linguistic Inquiry and Word Count, TF-IGM = Term Frequency-Inverse Gravity Moment

of 30 items. In both inventories, items are short, easy-to-understand descriptive phrases that respondents will rate on a 5-point scale ranging from "disagree strongly" to "agree strongly." Once the participants complete the personality test, they are presented with their personality scores in a range of 0-100, with feedback on each personality dimension. In this setting, we motivate them to complete the questionnaire and respond honestly, eliminating the random response problem.

The dataset collection strategy specifically aimed to capture a large volume of tweets per user to enable the analysis of user-level linguistic and behavioral patterns. This approach aligns with our main research objectives, which focus on analyzing users' personalities based on their social media posts. Collecting a higher average tweet count per user allows for a more detailed analysis to understand individual behavior deeply. Compared to broader datasets like PAN-2015, increasing the tweet count per user may reduce noise and enhance the robustness of the findings at the individual level. Moreover, the results generated in this paper align with trends observed in other datasets, further supporting the general applicability of the proposed model, and providing detailed linguistic and behavioral analysis.

The online questionnaire link has been distributed over public social network platforms (i.e., Twitter, Facebook, Reddit, Telegram) to reach participants who meet the following criteria: 1) have an active Twitter account with at least 25 tweets, and 2) they use English as the primary language. The selection of 25 as the minimum number of tweets is motivated by other studies' findings [49], [55], which found that having 25 tweets from each user can effectively increase the APR performance. All participants provided informed consent to participate in the study and voluntarily agreed to share their tweets and personality scores for research purposes[2]

A total of 47 valid responses were received over three months. Following informed consent, 18 of them provided us with their Twitter handles (@usernames) and voluntarily

[2]To keep privacy, the dataset is available upon request for research purposes only, and is not released as an open-access resource.

TABLE 3. X-Big5 Dataset's Participants' demographic characteristics.

Characteristic	Percent of Respondents (%)
Gender	
Male	39%
Female	61%
Age	
15 to 20 years old	33%
21 to 30 years old	44%
31 to 40 years old	5%
41 to 50 years old	17%
Over 50 years	1%
Education Level	
Have not completed high school	19%
High school graduate	17%
Some college	8%
Bachelor's degree	39%
Master's degree	11%
Doctorate	6%
Marital Status	
Married	83%
Single	17%

accepted to use their tweets and personality scores for our research. We then checked the validation of the provided Twitter usernames to ensure the authenticity of the participants. Given a user's handle, and through the official Twitter Application Programming Interface (API), we retrieved all non-retweet tweets for each user that were posted from 30-12-2019 until 31- 12-2022. Notably, we specified the start date of tweets collection to the date of COVID-19 beginning, where numerous studies indicated an unprecedented change in human Big Five personality dimensions as a consequence of this pandemic [74], [75], [76]. To maintain privacy and confidentiality, all volunteers' Twitter handles (@usernames) have been permanently removed from the dataset; thus, the dataset is used anonymously and never linked to participants. Moreover, only publicly available tweets are retrieved, with no inclusion of private or restricted content.

The access to Twitter API was by Tweepy [77], a widely used open-source library composed of methods and classes that facilitate the connection with different Twitter API versions. As a result, a total of 38,488 tweets from 18 Twitter users were obtained, labeled with the Big Five personality traits scores ranging from 0 to 100. Table 3 presents demographic information about the participants.

B. DATA LABELING

Most existing personality studies tackled the APR either as a regression task (personality traits labeled with real-valued scores) or a classification task (personality traits labeled with high, middle, and low values). However, in the real-world environment, service/product providers are more interested in extreme values of a user's traits (i.e., high or low) and pay less, or almost no, attention to the middle scores [32], [54]. Taking game recommendation systems as an example, it is easier to persuade users high in Extraversion to install or buy socializer games, while it is less likely for those low in Extraversion (i.e., introverted) to accept such recommendations. Thereby, this work focuses on predicting the extreme values for each Big Five personality dimension and ignoring the middle trait level. Specifically, we transformed the continuous personality scores retrieved from the personality questionnaire (0-100) into discrete personality labels (0,1), in which 0 refers to the low personality score of a dimension and 1 to the high personality score. The transformation was applied by using the mean value for each dimension's scores as a splitting point so that any value less than the mean is transformed into 0, and anyone greater than or equal to the mean is converted into one.

C. DATA PRE-PROCESSING

Text preprocessing is a critical step in NLP, including different practices to clean texts from noise and less informative data, which may mislead the model and affect its accuracy. In our context, Kosan et al. experimented with various preprocessing steps to determine the best synthesis that could improve the success rate of their APR model [48]. Specifically, they used 11 preprocessing steps as standard steps (involved in all experiments) and 13 steps used for comparative analysis. The authors revealed different performance results when applying different combinations of these steps. For example, they found that removing stop words from the corpus effectively increased the model's success rate, while applying stemming resulted in worse results. Inspired by their findings, as well as other correlation results between personality traits and different language habits retrieved from other state-of-the-art studies [30], [36], [38], we applied the following pre-processing methods to our dataset:

1) LOWERCASING AND FILTERING TWEETS

Lowercasing is one of the essential preprocessing techniques in which all tweets will be converted into a constant format by transforming them into lowercase form. We used the lower, built-in Python method to lowercase all uppercase letters in the dataset. Then, we filtered tweets by excluding all short tweets (a tweet with a number of words < 3).

2) REPLACING SLANG WORDS AND ABBREVIATIONS

Social media texts generally contain informal language use, including a lot of abbreviations, acronyms, and slang words. Slang words are informal words, phrases, or expressions typically restricted to a particular context or network of people, such as F2F" refers to Face to Face, "ty" to "thank you," "TBH," to be honest. Additionally, abbreviations and acronyms, shortened forms of a word or phrase, such as omg "oh my god," ASAP " as soon as possible," covid-19" commonly appeared in social media texts. Due to the variety of their structures, classification models faced difficulty distinguishing and interpreting them. This issue can be solved by constructing an index of slang and abbreviations associated with their standard replacements. Thus, we scraped the NoSlang [78], a common Internet slang dictionary website, to create a dictionary consisting of 5,817 common English slang with their formal replacements. Using this dictionary, we converted all slang into its formal form.

3) DELETING PUNCTUATIONS

KN et al. revealed that the appearance of punctuation (e.g., question marks, ellipses, exclamations) does not provide considerable predictive value for personality recognition [38]. Therefore, we removed them from tweets by using the punctuation constant, a predefined constant in the Python string module consisting of a set of commonly used punctuation characters.

4) EXPANDING CONTRACTIONS

Expanding contractions, such as I'm, she's, hasn't, and let's, into their complete forms is an effective pre-processing technique, especially if it is undertaken before the tokenization step. This is because tokenizers split contractions in a nonsense manner; for example, the contraction haven't will tokenizing into two words: haven and t, and I'm into I and m. In this step, we used the contractions Python library to expand all contractions in the dataset.

5) REMOVING NUMBERS

All numbers are removed from the dataset using a regular expression, as they do not reflect any aspect of personality. This step is carried out after replacing all slang and abbreviations because some of them include numbers in their syntax, such as F2F and 2good for " too good."

6) REMOVING STOP-WORDS

Stop words are function words or the most frequent words in any language, such as conjunctions, pronouns, articles, and linking verbs. Although most APR studies removed these words as a critical practice in NLP, we believe it is more beneficial to keep some of them, which have been proven to have strong associations with personality traits. For example, Tadesse et al. revealed that people scoring high in openness strongly correlated with high-frequency words, including articles (i.e., a, an, the), prepositions (e.g., with, to), third-person singular (e.g., she, he,) and third-person plural (e, g., they, their) [30]. Additionally, extraverts were found to strongly correlate with 2nd and 3rd person singular pronouns and with agreement words such as "OK" and "yes," while neurotics frequently used 1st person plural (e.g., we, ours) [30]. Therefore, we utilized the stop words list provided by the NLTK package [79] to remove unrelated stop words for the APR task and keep relevant ones.

7) TOKENIZATION

As a last pre-processing step, we broke down tweets into separate words called tokens using a Tokenizer class of the TensorFlow Keras module. Tokenization is an essential step before passing text samples into the model. Stemming is not applied to our dataset to preserve sentence tense, which is found helpful in the APR task [30], [80].

Although the above pre-processing steps are critical to improve data quality and remove noise, excluding certain texts, such as shorter tweets, may affect the model's generalizability. For example, while removing shorter tweets, punctuations, and replacing slang and abbreviations reduce noise in the dataset, they may make the model less adaptable to diverse data. Future comparative analysis could be conducted by training models with and without certain text types to evaluate their impact on prediction performance and generalizability.

IV. PROPOSED MODEL

Text-based APR is still a challenging problem in affective computing. Despite existing efforts, the performance is far from being optimal, especially at the level of each personality dimension. Hence, we aim to solve this issue by proposing an efficient APR model, APR_ConvLSTM, that combines the impressive capabilities of both CNN and LSTM deep neural networks. The APR_ConvLSTM is a unified end-to-end model where all personality dimensions are predicted simultaneously. As depicted in Fig. 1, the CNN comprises parallel convolution layers with multiple convolution kernels of different sizes to extract local features from the input representation. Then, a max-overtime pooling operation is performed to capture the most important features, which are then passed into a Bi-LSTM. The Bi-LSTM, in turn, learns both past and future long-term semantic dependencies, which are concatenated and passed into two successive fully connected layers. Finally, a user's Big Five personality traits classification results are generated at the output layer, which is composed of five binary layers, each for a specific personality trait. The upcoming sections explain each component in detail.

A. WORD EMBEDDING LAYER

Traditional word representation methods, like one-hot encoding, suffer from different drawbacks, including data sparsity, losing word order, and dimensionality-oversize [81]. Alternatively, word embedding, a technique used to convert a word into a dense, low-dimensional, real-valued vector, has been applied to various NLP tasks and achieved remarkable performance compared to traditional methods. Thus, we used the (GloVe) embedding method, which is developed through unsupervised learning trained on 60 billion words of Wikipedia 2014 and Gigaword2 corpus, to generate initial word representations [82].

Let's assume a text has m words; the embedding vector t of $w_{(n)}$, $n \in [1, m]$ will be constructed through a Glove-embedding matrix $W_g \in \mathbb{R}^{|v|\times d}$, (where v is the vocabulary size and d the embedding dimension), by the following equation:

$$t_{(n)} = W_g w_n, \quad n \in [1, m] \tag{1}$$

We will specify the initial length of a word embedding vector as 300 dimensions, as most NLP studies [83], [84] have demonstrated that it is the optimal length for word embedding

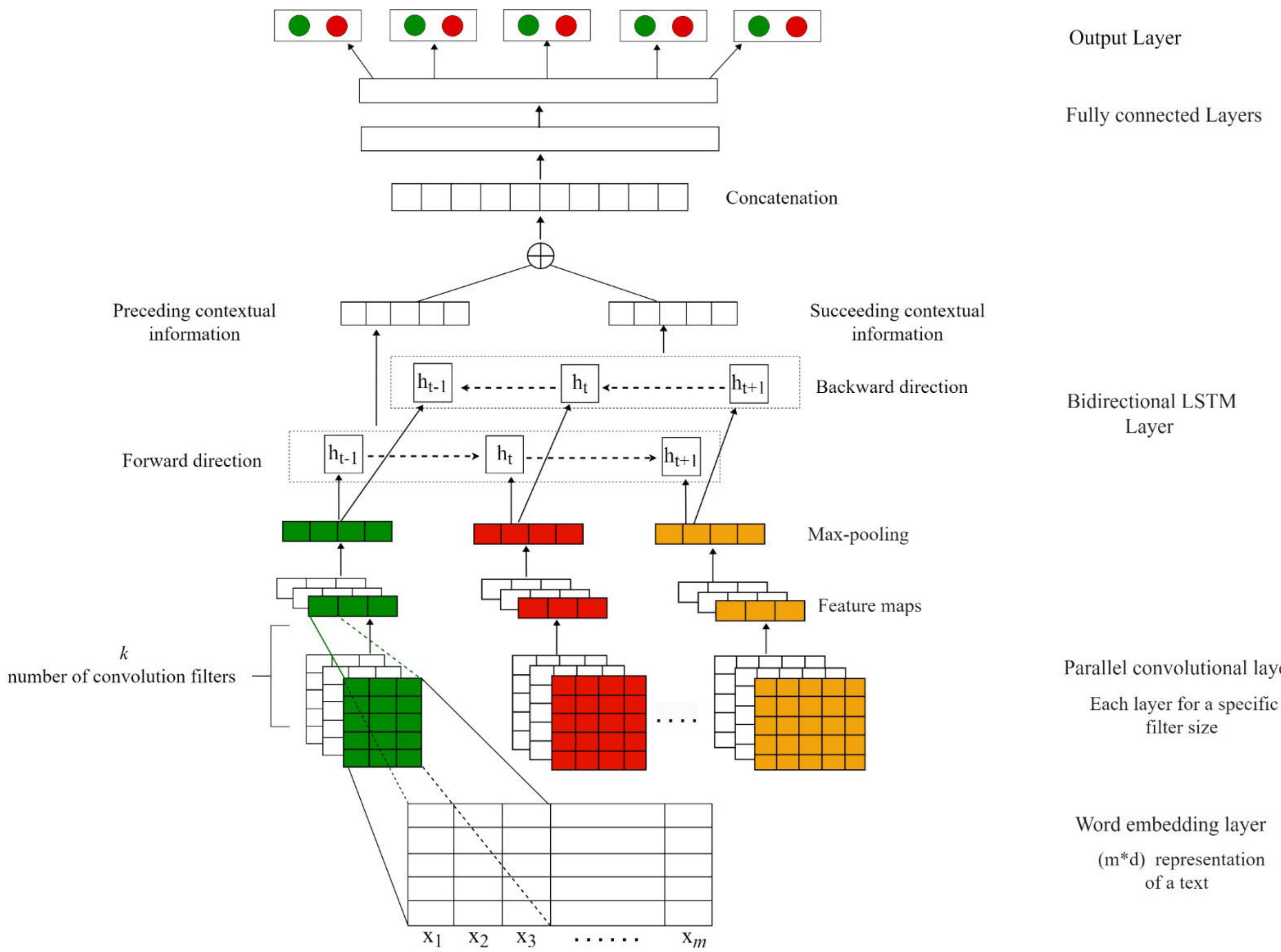


FIGURE 1. The network architecture of the APR_ConvLSTM model. It comprises five main layers: the word embedding layer, CNN layers, Bi-LSTM layer, fully connected layers, and the output layer.

vectors. Therefore, the input data of our model will be a two-dimensional matrix $T \in \mathbb{R}^{m\times d}$, in which m is the maximum number of words in a sample (note that we apply padding when necessary), and d is the embedding dimension of each token (which contains 300 numerical values).

B. CNN LAYERS

CNN was invented, for the first time, to be used with computer vision tasks such as image recognition, image classification, and object detection. Afterward, this network architecture was adopted with various NLP tasks and showed impressive performance in extracting robust and abstract features. Here, the embedding matrix retrieved from the previous layer is fed into a parallel convolutional layer and max-overtime pooling to capture meaningful local features (i.e., various n-grams features) and, most importantly, reduce the input data's dimensionality.

1) CONVOLUTION LAYER

In this layer, k convolution filters with various sizes are used to perform a convolutional operation over a window of h word vectors successively. More precisely, let x refer to a tweet sample composed of words x_1 to x_m; a new feature c_i will be generated from a window of words $x_{(i:i+h-1)}$ by:

$$c_i = f(w \cdot x_{(i:i+h-1)} + b) \tag{2}$$

where $c_i \in \mathbb{R}^{(m-h+1)}$, f refers to the Exponential Linear Unit (ELU) [85], a nonlinear activation function, $w \in \mathbb{R}^{h\times d}$ is a randomly initialized weight, and b is a bias vector. Here, we selected the ELU function to provide nonlinearity for the convolution operation since it can accelerate the learning process and achieve higher performance results with CNN, compared with other activation functions [85], [86], [87]. By repeatedly applying this operation over all possible h words of a given tweet, a new feature map will be produced: $c \in [c_1, c_2, \ldots, c_{(m-h+1)}]$. Therefore, each convolution filter will produce a feature map whose dimension is $1\times(m-h+1)$.

If q parallel convolution layers of various filter sizes (i.e., each layer is dedicated to a specific h value) are used, and the number of filters in each layer is k, then a total of $(k \times q)$ feature maps will be obtained from the convolution operation.

2) POOLING LAYER

After the convolution operation, a max-overtime pooling operation is applied over each $1 \times (n-h+1)$ feature map to extract the maximum value $p = \max\{c\}$, which is used as the feature corresponding to a given feature map. The idea here is to preserve the most important feature (i.e., the one with the highest value) from each feature map. All maximum features corresponding to the same window size h, which are produced by k convolution filters, are concatenated to obtain a one-dimensional feature vector:

$$v = [p_{(h,1)}, p_{(h,2)}, \ldots, p_{(h,k)}] \tag{3}$$

As the APR_ConvLSTM model is composed of q parallel convolution layers, each corresponding to a specific window size h, q one-dimensional feature vectors are outputted from the max-pooling operation, where $V = [v^{(1)}, v^{(2)}, \ldots, v^{(q)}]$.

C. BI-LSTM LAYER

LSTM is a powerful variant of RNNs mainly proposed to alleviate the vanishing and exploding gradients problem faced by traditional RNNs, which prevent capturing long-term dependencies between sequential data. Along with the previous-time step hidden state (a short-term memory) equipped with vanilla RNNs, the LSTM structure contains an additional cell state called long-term memory to enhance the remembering capacity of the model. However, the standard version of LSTMs processes the input data in a forward direction, allowing them access only to preceding contextual information. Therefore, to completely understand a text context, we applied Bi-LSTM, which can capture both past and future contextual features by combining a forward hidden layer (represented as $\overrightarrow{\text{LSTM}}$) and a backward hidden layer (represented as $\overleftarrow{\text{LSTM}}$).

Given the feature sequences from the CNN layer $[v^{(1)}, v^{(2)}, \ldots, v^{(q)}]$, the $\overrightarrow{\text{LSTM}}$ layer reads the feature sequences from $v^{(1)}$ to $v^{(q)}$, respectively, to learn the forward hidden state $\overrightarrow{h_f}$, while the $\overleftarrow{\text{LSTM}}$ layer reads the feature sequences from $v^{(q)}$ to $v^{(1)}$, respectively, to learn the backward hidden state $\overleftarrow{h_b}$. Then, both forward and backward hidden states are concatenated at each time step (i) to obtain the final hidden representation $h_i = [\overrightarrow{h_i}\,;\,\overleftarrow{h_i}]$, where:

$$\overrightarrow{h_l} = \overrightarrow{\text{LSTM}}(v^{(i)}),\ i \in [1, q] \tag{4}$$

$$\overleftarrow{h_l} = \overleftarrow{\text{LSTM}}(v^{(i)}),\ i \in [q, 1] \tag{5}$$

As depicted in Fig. 2, each $\overrightarrow{\text{LSTM}}$ or $\overleftarrow{\text{LSTM}}$ unit (i.e., neuron) is composed of three main inputs: the input of the current time step (v_t), previous hidden state (h_{t-1}), and previous cell state (c_{t-1}), which are used to generate two main outputs: the current time step hidden state (h_t) and updated cell state (c_t). To regulate information flow into and out of the memory cell, three main gates are utilized: an input gate i_t, a forget gate f_t, and an output gate o_t.

The forget gate determines the amount of information required to omit from the previous state of long-term memory

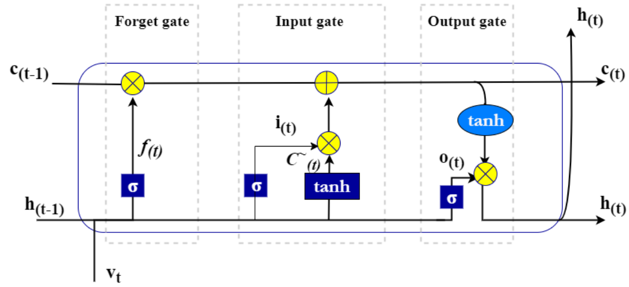


FIGURE 2. Structure of the LSTM unit consists of three main gates: forget gate f_t, input gate i_t, and output gate o_t. Given three input vectors v_t, h_{t-1}, and c_{t-1}, the LSTM unit produces two main output vectors: the current time step hidden state h_t and the updated cell state c_t.

(c_{t-1}) by applying the following equation:

$$f_t = \sigma\left(W_f v_t + W_f h_{t-1} + b_f\right) \tag{6}$$

As we know, the sigmoid activation function outputs a number in [0, 1], where 0 means removing all previous information held by c_{t-1}, and 1 indicates keeping the information.

The input gate, in turn, determines what new information should be added to the cell state by applying the following equations:

$$i_t = \sigma\left(W_i v_t + W_i h_{t-1} + b_i\right) \tag{7}$$

$$\tilde{c}_t = \tanh\left(W_{\tilde{c}} v_t + W_{\tilde{c}} h_{t-1} + b_{\tilde{c}}\right) \tag{8}$$

As shown in Fig. 2, the two vectors retrieved from the above two equations will then undergo point-wise multiplication. As the tanh activation function produces a value in [−1, 1], the multiplication result vector will either have a positive or negative sign, which is then added to or subtracted from the cell state (i.e., updating the cell state).

Finally, the output gate comes into play to decide what information the current time step hidden state h_t should contain by considering the previously hidden state h_{t-1}, the current input data v_t, as well as the recently updated cell state through the following equations:

$$o_t = \sigma\left(W_o v_t + W_o h_{t-1} + b_o\right) \tag{9}$$

$$h_t = o_t \odot \tanh(c_t) \tag{10}$$

In more detail, the previously hidden state h_{t-1}, the current input data v_t, will pass into a sigmoid function to squash its value between [0, 1]. Then, the recently updated cell state goes through a tanh function to squash its value between [−1, 1]. Lastly, both values will undergo point-wise multiplication to decide what information the current time step hidden state should carry for the next time step input (i.e., v_{t+1}).

D. FULLY CONNECTED LAYERS

Then, the hidden representation h, received from the Bi-LSTM layer, is passed into two successive fully connected dense layers, which are activated with the Rectified Linear

Unit (ReLU) function f, as follows:

$$Z^1 = f(W^1.h + b^1) \tag{11}$$
$$Z^2 = f(W^2.Z^1 + b^2) \tag{12}$$

where W^1, b^1 and W^2, b^2 are weights and bias matrices for the first and second dense layer, respectively. To avoid overfitting, a dropout technique [88] is applied between these dense layers with a rate of 0.2. Dropout is a common regularization technique that randomly sets a fraction of input units to 0 at each update of the training process, which in turn mitigates overfitting and improves the generalization of the model.

E. OUTPUT LAYER

Finally, features obtained from the last dense layer are connected to the final output layer. As our APR_ConvLSTM aims to predict all Big Five personality traits simultaneously, we decompose the output layer into five parallel binary layers; each layer is assigned to a specific personality trait t, where $t \in$ {EXT, CON, OPN, AGR, NEU}. Here, SoftMax is used at each layer for binary classification, whose output is the probability distribution over the trait's classes k (i.e., 0 or 1). The output of the i-th neuron of each trait layer is formalized as follows:

$$\hat{y}_t = (w_t Z^2 + b_t) \tag{13}$$
$$\text{SoftMax}(\hat{y}_t)_i = \frac{e^{\hat{y}_{t_i}}}{\sum_{j=1}^{k} e^{\hat{y}_{t_j}}} \tag{14}$$

The cross-entropy loss between a trait's predicted probabilities and the ground-truth labels is calculated at each trait's layer, as follows:

$$L(y, \hat{y})_t = -\sum_{d=1}^{D}\sum_{c=1}^{C} y_d^c \log \hat{y}_d^c \tag{15}$$

where $y, \hat{y}$ are ground-truth labels and predicted probabilities of a text sample d, respectively, D is the number of samples, and C is the number of classes of each trait (binary classes). Finally, the training process aims to minimize the overall loss of APR_ConvLSTM, which is composed of the individual loss of each trait layer as follows:

$$L_{\text{total}} = L(y, \hat{y})_{\text{EXT}} + L(y, \hat{y})_{\text{CON}} + L(y, \hat{y})_{\text{OPN}} + L(y, \hat{y})_{\text{AGR}} + L(y, \hat{y})_{\text{NEU}} \tag{16}$$

V. EXPERIMENTS AND RESULTS ANALYSIS

In this section, four comparative experiments have been conducted to evaluate the performance of the APR_ConvLSTM model on the text-based APR task. We also provide a detailed analysis of the overall performance of the proposed model and verify whether the improvements achieved are statistically significant compared to the state-of-the-art baselines. First, Section V-A displays the experiment setup, including datasets, evaluation metrics, training procedure, and an overview of the four experiment groups. Then, the results of each experiment set are presented and analyzed in Section V-B. We further show the overall performance of the proposed model and the statistical significance test results in Section V-C.

TABLE 4. Statistics of each personality dataset.

Dataset	Number of Instances	Average Length	Max Length	Vocabulary Size
PAN-2015	25,816	15	46	40,229
X-Big5	31,417	11	51	30,511

A. EXPERIMENTS SETUP

1) DATASETS

In this work, two real-world data sets are used to evaluate the APR_ConvLSTM model. The first dataset is the PAN-2015[3] Author Profiling dataset [53], which was released as a part of the 3rd Author Profiling Shared Task at PAN 2015. The dataset consists of a set of tweets collected in four languages: English, Spanish, Italian, and Dutch, and annotated with a user's age, gender, and Big Five personality dimensions. The Big Five personality scores were assessed with the BFI-10 personality test and reported in the range of $[-0.5, 0.5]$. In our experiments, we considered only the English corpora, which consists of 27,344 tweets collected from 294 users. Since our APR_ConvLSTM is a multi-label personality classification model concerned with predicting extreme values (i.e., high, and low) of each personality trait, we need to convert personality scores into two classes (i.e., 1 if a score is high and 0 if it is low). Following a study [89], we used the median point (i.e., 0.1) as a threshold such that the trait would be high if its score equals or is above the median; otherwise, it would be low. All preprocessing methods explained in Section III-C are also applied to this dataset.

The second dataset is the X-Big5 dataset, which is previously described in detail in Section III. To recap, the dataset comprises a total of 31,417 (non-retweet) tweets, annotated with their Big Five personality scores in the range of [0 to 100]. The scores are also discretized into two classes for a trait binary classification. Table 4 shows some statistics of both datasets.

2) EVALUATION METRICS

We used Accuracy, Precision, Recall, and F1-score to evaluate the performance of the APR_ConvLSTM on the APR task. The calculation formulas of these metrics are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$
$$\text{Precision} = \frac{TP}{TP + FP} \tag{18}$$

[3]PAN is not an acronym but rather a standalone name for a series of shared tasks on digital text forensics. See: https://github.com/pan-webis-de/pan-webis-de.github.io/blob/master/FAQ.md

TABLE 5. Parameters values of APR_ConvLSTM model.

Parameters	Value
Filter sizes	1, 2, 3
Number of filters	128
Bi-LSTM hidden units	256
Dropout rate	0.2
Epochs	30

$$\text{Recall} = \frac{TP}{TP + FN} \tag{19}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{20}$$

where TP (i.e., True Positive) represents the number of positive examples that are classified correctly, FN (i.e., False Negative) represents the number of negative examples that are classified incorrectly, FP (i.e., False Positive) indicates the number of positive examples that are classified incorrectly, and TN (i.e., True Negative) represents the number of negative examples that are classified correctly.

3) TRAINING

To properly validate our model and prevent overfitting, we adopted a training-testing schema on both datasets as follows. First, we used the official split to randomly select 80% of the data for training and 20% for testing. Then, 10-fold cross-validation is applied to the training set (i.e., 80% of the data) so that at each k-training iteration, the training set will be divided into ten mutually exclusive subsets; one is used as a validation set, and the remaining nine are used for training. The stratified sampling is applied over the 10-fold to ensure that each fold has nearly the same class distribution of each trait as the complete dataset. This technique provides a robust evaluation of the model by training and testing it on multiple data subsets.

Moreover, a grid search method was used to find the optimal configuration of three model hyperparameters: dropout rate, number of epochs, and LSTM hidden units. The search ranges of these hyperparameters are set as follows: dropout rate = [0.2, 0.5], number of epochs = [20, 50], and LSTM hidden units = [64, 128, 256]. We also conducted extensive experiments (Experiment 3) to investigate the effect of two structure-related hyperparameters (i.e., filter sizes and filter numbers) and various word embedding methods on the model performance. The best-suited model hyperparameters are presented in Table 5. With the optimized hyperparameters, all models are trained through the Adam optimizer on shuffled mini-batches of size 32 and a fixed learning rate of 0.001.

4) PERFORMANCE COMPARISON

The main question we seek to answer in this work is whether the proposed APR_ConvLSTM can improve the performance of text-based APR from social media users' content. For this purpose, four comparative experiments were set up in this paper.

The first experiment set compared our APR_ConvLSTM with classical machine-learning methods such as SVM, Decision Tree, and Random Forest. The second set is designed to investigate the effect of the large dataset on the APR_ConvLSTM performance. As we mentioned, the text APR from social media content lacks the availability of benchmark datasets, representing a significant challenge in the domain. Thus, a large personality dataset has been created in the second experiment by combining both PAN-2015 and X-Big5 datasets. Then, we compare the model performance on the large dataset against the performance of using each single dataset separately.

The third experiment compares the model performance under different hyperparameters (e.g., filter sizes, filter numbers) and with various word embedding methods (e.g., Glove, Word2Vec, FastText). The fourth experiment is conducted to compare the APR_ConvLSTM performance with state-of-the-art APR models. All these experiments were conducted on an NVIDIA T4 GPU with 16 GB high-bandwidth memory (GDDR6).

B. RESULTS AND ANALYSIS

1) EXPERIMENT 1

The first set of experiments aimed at comparing the APR_ConvLSTM performance against traditional machine learning classification methods, including SVM, Decision Tree, K-Nearest Neighbors (KNN), and Naive Bayes. Experimental results for both datasets, PAN-2015 and X-Big5, are represented in Table 6 and Table 7, respectively.

As can be seen from Table 6, the proposed APR_ConvLSTM achieved higher performance results on the PAN-2015 dataset with all personality traits in the four metrics (i.e., Accuracy, Precision, Recall, and F1-score) compared to SVM, Decision Tree, and Naive Bayes methods. Although the K-Nearest Neighbors method achieved almost similar results in terms of four traits (Openness, Conscientiousness, Extraversion, and Agreeableness), our model was better at predicting Neuroticism with an increase of 5.67%, 5.87%, and 2.72% in Accuracy, Recall, and F1-score, respectively.

Interestingly, except for the Naive Bayes method, we notice that the Openness trait dominated the highest performance with our APR_ConvLSTM and other machine learning methods. This suggested the high potential of detecting this trait from social media users' generated texts, compared with other traits. Additionally, our model again achieved superior performance on the X-Big5 dataset in all metrics for each personality dimension, compared to other machine-learning methods, as demonstrated in Table 7.

2) EXPERIMENT 2

Datasets play a key role in advancing the APR task, not just as a source for training models but also as a means for evaluating and comparing performance. However, as we mentioned previously, APR from social media users' texts lacks the availability of well-annotated Big Five personality

TABLE 6. Comparison with machine-learning methods on PAN-2015 dataset.

Model	Trait	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	OPN	97.62	97.69	99.92	98.79
	CON	48.76	78.01	48.81	60.07
	EXT	55.65	80.68	58.40	67.76
	AGR	34.68	70.72	21.69	33.19
	NEU	48.24	65.87	45.00	53.47
	Average	56.99	78.59	54.76	62.66
KNN	OPN	**97.81**	97.85	**99.96**	**98.89**
	CON	74.71	79.75	**91.10**	85.05
	EXT	**76.49**	79.84	**94.37**	**86.50**
	AGR	70.55	76.00	**88.61**	81.83
	NEU	60.79	67.46	**78.55**	72.59
	Average	76.07	80.18	**90.52**	84.97
Decision Tree	OPN	96.15	98.08	97.98	98.03
	CON	69.79	81.64	79.64	80.63
	EXT	70.62	81.25	82.14	81.69
	AGR	67.02	78.09	77.74	77.91
	NEU	60.61	70.10	70.47	70.28
	Average	72.84	81.83	81.59	81.71
Naïve Bayes	OPN	61.62	97.40	62.38	76.05
	CON	47.08	81.08	42.99	56.19
	EXT	61.44	80.70	67.91	73.76
	AGR	64.31	76.35	75.78	76.06
	NEU	56.08	68.52	62.06	65.13
	Average	58.11	80.81	62.22	69.44
APR_ConvLSTM	OPN	97.73	**98.20**	99.50	98.85
	CON	**75.94**	**82.67**	87.98	**85.24**
	EXT	75.89	**84.15**	85.97	85.05
	AGR	**72.54**	**79.64**	85.04	**82.25**
	NEU	**66.46**	**73.33**	77.41	**75.31**
	Average	**77.71**	**83.60**	87.18	**85.34**

OPN= Openness, CON= Conscientiousness, EXT= Extraversion, AGR= Agreeableness, NEU= Neuroticism

TABLE 7. Comparison with machine-learning methods on X-Big5 Dataset.

Model	Trait	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	OPN	64.47	45.82	09.24	15.38
	CON	52.64	67.27	58.41	62.53
	EXT	61.09	30.46	46.60	36.84
	AGR	65.17	28.79	27.55	28.16
	NEU	62.91	38.09	15.17	21.70
	Average	61.26	42.09	31.39	32.92
KNN	OPN	64.50	48.85	33.83	39.98
	CON	64.59	70.28	82.59	75.94
	EXT	74.12	43.02	19.35	26.69
	AGR	74.09	45.21	21.52	29.16
	NEU	64.99	47.47	31.28	37.71
	Average	68.46	50.97	37.71	41.90
Decision Tree	OPN	70.29	57.39	58.20	57.79
	CON	70.35	78.32	77.68	77.99
	EXT	73.93	46.49	46.73	46.61
	AGR	75.19	49.94	49.39	49.66
	NEU	70.81	56.93	56.93	56.93
	Average	72.11	57.81	57.79	57.80
Naïve Bayes	OPN	63.42	47.08	37.75	41.90
	CON	64.70	72.05	78.15	74.97
	EXT	71.15	39.71	35.69	37.59
	AGR	71.24	41.27	37.96	39.55
	NEU	63.96	46.01	36.82	40.91
	Average	66.89	49.22	45.27	46.98
APR_ ConvLSTM	OPN	**82.77**	**78.12**	**70.40**	**74.06**
	CON	**82.16**	**85.58**	**88.54**	**87.04**
	EXT	**85.31**	**72.46**	**63.99**	**67.96**
	AGR	**86.33**	**76.24**	**65.13**	**70.25**
	NEU	**82.70**	**77.16**	**69.52**	**73.14**
	Average	**83.85**	**77.91**	**71.52**	**74.49**

OPN= Openness, CON= Conscientiousness, EXT= Extraversion, AGR= Agreeableness, NEU= Neuroticism

datasets. To date, only the PAN-2015 dataset, which is detailed in Section V-A1, is a publicly available benchmark. Although some studies tried to collect their own personality datasets from different social network platforms [36], [47], their datasets are relatively small or no longer available for sharing. This can be mainly attributed to the cost of data annotations and user's privacy risks.

Therefore, we designed this experiment to verify the effectiveness of the APR_ConvLSTM when running on a large-scale personality dataset. Following previous work [90], we constructed the large dataset by combining the aforementioned experimental datasets, i.e., PAN-2015 and X-Big5, which resulted in a dataset composed of 57,233 tweets from 267 users. In this experiment, we used the same training protocol described in Section V-A3.

Table 8 reports the performance of our APR_ConvLSTM on the large-scale dataset against its performance on each single dataset. Not surprisingly, we observed that the highest averaged performance in terms of all metrics was achieved when training on small datasets. Specifically, the best average precision, recall, and F1-score were on the PAN-2015 dataset at 83.60%, 87.18%, and 85.34%, respectively, while the best average Accuracy was on our dataset with a value of 83.85%. This can be explained by the lack of diversity of personality traits in the small-scale training datasets. However, by looking at the performance at each personality trait, we can notice that training on the large-scale dataset achieved excellent results (i.e., above 70%) for all traits across all metrics. For example, it achieved an important increase in the Accuracy for the Neuroticism trait from 66.46% to 72.52% compared to the PAN-2015, and an F1-score increase from 67.96 % to 74.89 % in the Extraversion trait compared to the X-Big5 dataset.

TABLE 8. Performance on a large dataset against small-scale datasets.

Dataset	Trait	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
PAN-2015	OPN	**97.73**	**98.20**	**99.50**	**98.85**
	CON	75.94	82.67	87.98	85.24
	EXT	75.89	**84.15**	**85.97**	**85.05**
	AGR	72.54	**79.64**	**85.04**	**82.25**
	NEU	66.46	73.33	**77.41**	**75.31**
	Average	77.71	**83.60**	**87.18**	**85.34**
X-Big5	OPN	82.77	78.12	70.40	74.06
	CON	**82.16**	**85.58**	**88.54**	**87.04**
	EXT	**85.31**	72.46	63.99	67.96
	AGR	**86.33**	76.24	65.13	70.25
	NEU	**82.70**	**77.16**	69.52	73.14
	Average	**83.85**	77.91	71.52	74.49
Combined	OPN	82.96	86.83	86.30	86.57
	CON	78.00	83.52	86.93	85.19
	EXT	75.41	74.98	74.79	74.89
	AGR	75.55	74.66	73.44	74.04
	NEU	72.52	72.79	70.09	71.42
	Average	76.89	78.56	78.31	78.42

OPN= Openness, CON= Conscientiousness, EXT= Extraversion, AGR= Agreeableness, NEU= Neuroticism

TABLE 9. Comparison with machine-learning methods on the combined dataset.

Model	Trait	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	OPN	43.18	59.04	34.79	43.79
	CON	54.15	73.34	58.15	64.87
	EXT	45.61	41.07	22.42	29.01
	AGR	50.35	48.49	73.25	58.35
	NEU	52.83	52.08	46.48	49.12
	Average	49.22	54.80	47.02	49.03
KNN	OPN	62.90	68.71	76.50	72.40
	CON	68.44	73.94	**87.48**	80.14
	EXT	56.15	55.81	55.30	55.55
	AGR	56.77	54.81	51.20	52.94
	NEU	55.77	55.17	51.74	53.40
	Average	60.01	61.69	64.44	62.89
Decision Tree	OPN	70.53	76.59	77.29	76.94
	CON	68.38	78.75	77.48	78.11
	EXT	63.59	63.30	63.14	63.22
	AGR	64.86	62.97	63.13	63.05
	NEU	62.42	61.70	61.35	61.53
	Average	65.96	68.66	68.48	68.57
Naïve Bayes	OPN	61.61	69.13	71.63	70.36
	CON	68.15	75.22	83.87	79.31
	EXT	57.33	55.74	67.46	61.04
	AGR	56.60	54.01	57.98	55.93
	NEU	55.67	55.72	46.33	50.59
	Average	59.87	61.96	65.45	63.45
APR_ConvLSTM	OPN	**82.96**	**86.83**	**86.30**	**86.57**
	CON	**78.00**	**83.52**	86.93	**85.19**
	EXT	**75.41**	**74.98**	**74.79**	**74.89**
	AGR	**75.55**	**74.66**	**73.44**	**74.04**
	NEU	**72.52**	**72.79**	**70.09**	**71.42**
	Average	**76.89**	**78.56**	**78.31**	**78.42**

OPN= Openness, CON= Conscientiousness, EXT= Extraversion, AGR= Agreeableness, NEU= Neuroticism

Additionally, it boosts the performance for predicting Extraversion and Agreeableness in terms of the recall metric, with an increase of 10.8 % and 8.31 %, respectively. These observations demonstrate the essential role of the availability of large-scale datasets in improving the performance of deep neural networks. Further, we compared the performance on the large dataset with previously mentioned machine learning methods. The results are summarized in Table 9. Again, our APR_ConvLSTM outperforms other methods in all personality traits across various metrics. This clearly demonstrates the superiority of our model over machine learning methods regardless of the dataset size.

3) EXPERIMENT 3

There are numerous factors that affect the performance of deep learning models for text-based APR. Thus, we dedicated this section to investigating the effect of model hyperparameters and input data factors on the APR_ConvLSTM performance. The average of the four metrics: Accuracy, Precision, Recall, and F1- measure are used to evaluate the performance of the three datasets.

a: IMPACT OF FILTER SIZE

The size of convolution filters is one of the main factors that affect CNN performance. Intuitively, the larger the filter size, the more long-term contextual dependencies can be captured. However, when the filter size is too large, the

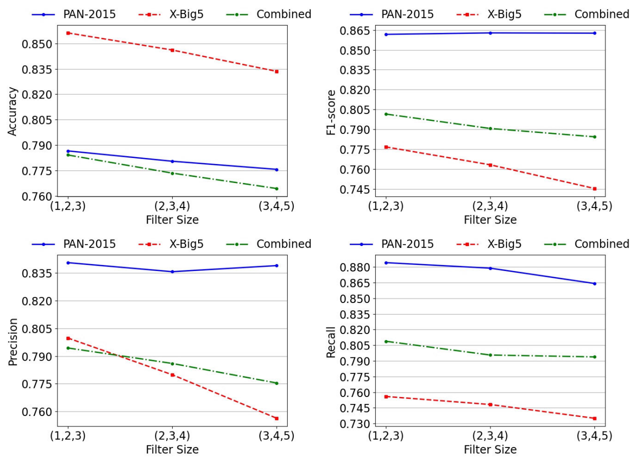


FIGURE 3. Impact of convolution filter size on the APR_ConvLSTM performance.

effect of the noise cannot be ignored, which may lead to significant performance reduction. To investigate the effect of filter size on the APR task, we conducted experiments by changing filter size combinations as follows (1, 2, 3-gram, 2, 3, 4-gram, 3, 4, 5-gram). The results are represented in Fig. 3. As expected, the best APR_ConvLSTM performance on all datasets and over all metrics was achieved by using a combination of the smallest filter sizes (i.e., uni-gram, bi-gram, and tri-gram combination). In general, the performance metrics started to decrease as the filter sizes increased due to the noise introduced with long sequences. This demonstrates the effectiveness of CNN at extracting local patterns, but at the same time, it fails to capture long-term contextual information.

b: IMPACT OF NUMBER OF FILTERS

The number of convolution filters (i.e., kernels) is another factor that affects the performance of deep learning models. In this section, we study the impact of changing the number of convolution filters on the APR_ConvLSTM performance. As shown in Fig. 4, except for the recall metric on the X-Big5 dataset, increasing the number of filters leads to an increase in the average performance of all metrics on the three datasets. This observation is not difficult to explain since the more number of filters means the more local patterns the convolution operation can be captured.

c: IMPACT OF WORD EMBEDDING METHODS

The selection of word embedding methods plays a vital role in the performance of text-based deep learning models. Therefore, this experiment is conducted with the aim of examining the impact of the commonly used word embedding techniques, Glove, Word2Vec: Skip-gram, Word2Vec: Continuous Bag-of-Words (CBOW), and FastText, on the APR_ConvLSTM performance.

The comparative results are presented in Table 10 and Fig. 5. As can be seen, the Glove and Word2Vec:

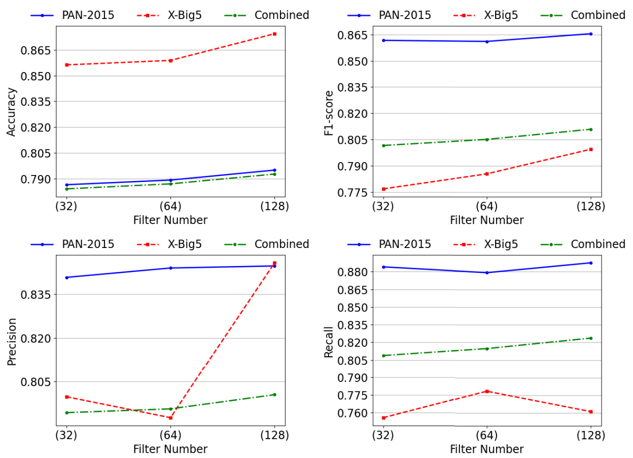


FIGURE 4. Impact of convolution filter number on the APR_ConvLSTM performance.

Skip-gram methods provide the highest performance for the APR_ConvLSTM in terms of all four metrics, even though their results vary across datasets. Specifically, while the Glove achieved better accuracy, F1-score, recall on the PAN-2015, and better precision on X-Big5, the Skip-gram outperforms it in all other cases. Notably, FastText performs the worst among all methods in all metrics, which indicates the poor ability of this method to provide a reliable representation for the APR task.

4) EXPERIMENT 4

In this section, we compared the performance of our proposed APR_ConvLSTM with state-of-the-art deep-learning APR baselines, including:

1) SEPRNN [46]: A semantic-enhanced personality recognition model composed of two main components: word-level representation and personality trait recognition. In the first component, the Bi-GRU is used to extract the left and right contextual vectors of each word, which are then combined with the word embeddings to obtain a more precise semantic representation. The output of the word-level representation process is passed to a fully connected layer with a sigmoid function to predict the binary values (i.e., yes/no) for each Big Five personality trait.
2) LMs + Attention mechanism [47]: A multi-model deep learning architecture built on top of three pre-trained language models (LMs): BERT, RoBERTa, and XLNet. Each input embedding generated from the pre-trained language models is fed into a self-attention mechanism to associate each word in the text with other words. Then, the output of the self-attention mechanism is combined with a variety of statistical features and passed to a feed-forward neural network consisting of three connected layers, each with a ReLU activation function. Finally, a model-averaging function calculates the unweighted average of SoftMax probabilities produced from the three classification models to make the final personality trait prediction (i.e., a label with the highest average probability).
3) Attention-based Uni-LSTM [65]: This model combines user topic preferences and sentiment features and feeds them into an attention-based LSTM model. First, the word embedding vectors are generated using the Word2Vec method and combined with sentiment features into a two-dimensional matrix. Second, the input matrix is fed into a unidirectional LSTM (Uni-LSTM) model with 2 hidden layers, each with 128 neurons. Then, topic information extracted by the LDA model is combined with LSTM output and passed to an Attention layer. Finally, a fully connected layer with 1 neuron and SoftMax function outputs the prediction of a user's personality category.
4) Bi-LSTM [48]: In this model, a user's tweets go through extensive preprocessing steps and then embed using the FastText embedding method. The FastText embeddings are then fed to a Bi-LSTM layer with 256 hidden units and a Swish activation function. Then, a dropout layer is used, followed by a dense layer composed of five neurons, each corresponding to one of the personality traits. The model is optimized using the Adam optimizer, and the Mean Squared Error (MSE) is used as a loss function.
5) CNN+Mairesse [91]: This model is one of the well-known baselines in the APR domain. It begins with the input layer, where texts are embedded using the Word2Vec embeddings method and sent to a convolutional layer. The convolutional layer extracts unigram, bigram, and trigram features and passes them to a max pooling to reduce dimensionality and generate sentence vectors. Further, a 1-max pooling is applied to aggregate the sentence vectors into one document vector, which is then concatenated with 84 Mairesse features. For final classification, document vectors are fed to a fully connected layer with a Sigmoid activation function, followed by a SoftMax layer with two neurons, representing the two classes (i.e., yes or no) for each personality trait.

Since all the above baselines, except the model of [91], do not have publicly available implementations, we re-implemented them using the same structure and parameters provided in their papers. For the CNN + Mairesse model [91], we used the open-source implementation available at `http://github.com/senticnet/personality-detection`. Notably, there are some parameters not explicitly specified by the authors, such as the number of epochs used to train the model proposed in [47] and the attention-based Uni-LSTM model [65]. Therefore, to ensure a fair comparison, we assigned to these parameters the same parameter values used with our proposed model.

In this experiment, we used the large-scale dataset (i.e., combined dataset) to compare the performance of our model against the baselines, as it contains more representative and

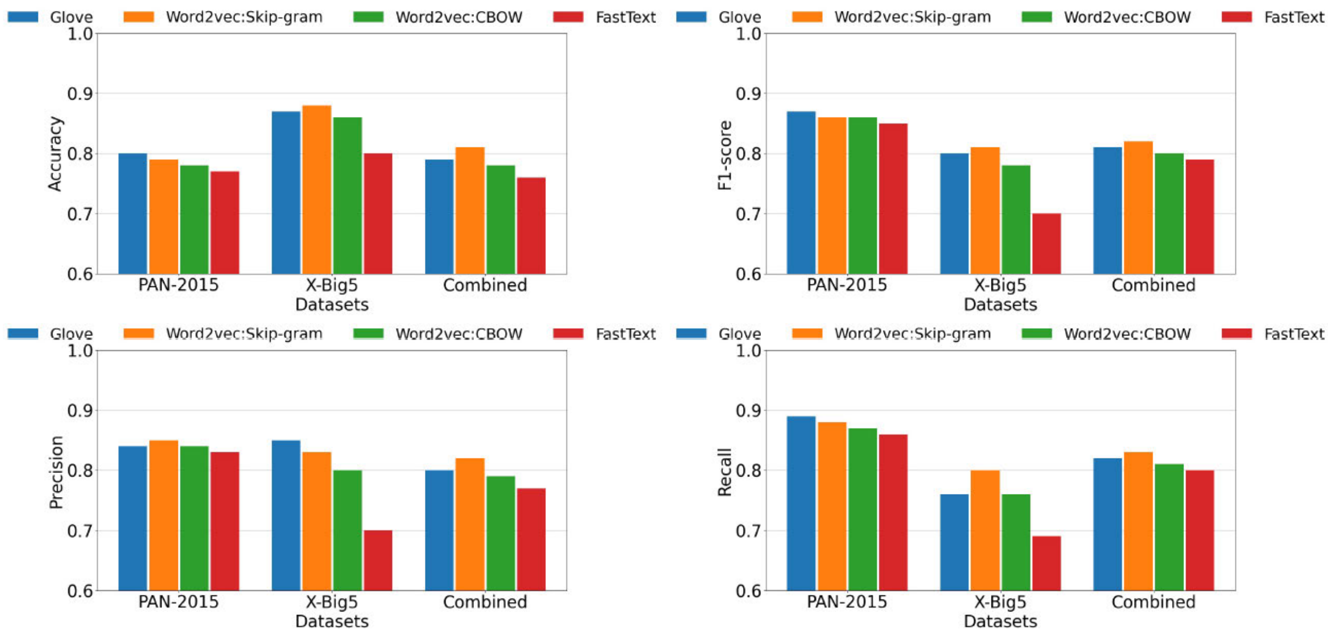


FIGURE 5. Impact of word embedding methods on the APR_ConvLSTM performance.

TABLE 10. Comparison Results of the proposed model with different word embedding methods, results are given in terms of average Accuracy, Precision, Recall, and F-1 Score.

Dataset	Word Embedding Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
PAN-2015	Glove	**79.51**	84.46	**88.76**	**86.54**
	Word2Vec: Skip-gram	78.98	**84.66**	87.67	86.13
	Word2Vec: CBOW	78.18	83.95	87.36	85.61
	FastText	76.70	83.17	86.07	84.59
X-Big5	Glove	87.43	**84.58**	76.13	79.95
	Word2Vec: Skip-gram	**87.95**	83.30	**79.56**	**81.35**
	Word2Vec: CBOW	85.61	79.58	75.91	77.66
	FastText	80.15	70.41	69.17	69.78
Combined	Glove	79.28	80.05	82.17	80.56
	Word2Vec: Skip-gram	**80.75**	**81.83**	**82.66**	**82.23**
	Word2Vec: CBOW	78.40	79.16	81.48	80.30
	FastText	76.47	77.13	80.17	78.62

diverse personality instances. Moreover, a t-test is performed on 10 independent runs of each model to check if the performance differences are statistically significant. The comparative results in terms of the average Accuracy and F-1 score are reported in Table 11 and Table 12, respectively.

As can be seen from Table 11, our model consistently and significantly outperforms all competitors in predicting all personality traits in terms of the Accuracy metric. Specifically, it achieved the highest Accuracy of 85.37%, 79.59%, 77.34%, 77.39%, and 75.38% for the Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, respectively. Moreover, the t-test results indicated that the performance difference between our proposed model and the best competitor of each personality trait is statistically significant, with the results of p-values < 0.01.

From Table 12, we notice that the APR_ConvLSTM again achieved the highest performance result in predicting each personality trait compared with other competitors in terms of the F-1 score. Interestingly, it reached an important F1-score increase from 72.33% to 77.35% in predicting the Extraversion compared to the model of [47] and from 71.56% to 76.16% in predicting the Agreeableness compared to the model of [48]. Moreover, we achieved competitive results in predicting Openness and Neuroticism traits with the highest F1-score of 88.60% and 74.52%, respectively. These improvements are statistically significant, except for the Conscientiousness trait, with the results of p-values < 0.01 for Openness, Extraversion, and Agreeableness and a p-value of 0.0259 (i.e., < 0.05) for the Neuroticism trait. Although the SEPRNN model [46] obtained a similar F-1 score for the Conscientiousness trait, our model was better in terms of Accuracy, with an increase of 2.07%.

C. FURTHER ANALYSIS

In this section, we present the overall performance of our proposed model in terms of the average Accuracy and F-1 score across all personality traits. Then, a statistical significance test is conducted to check whether the performance difference between our proposed model and each baseline is statistically significant.

1) OVERALL PERFORMANCE OF THE APR_CONVLSTM VS. BASELINES

Table 13 reports the overall performance of the proposed model on the combined dataset compared to the above-mentioned baselines. As one can see, our model shows

TABLE 11. Comparison of our proposed model APR_ConvLSTM with state-of-the-art APR models on the combined dataset in terms of accuracy (Mean ± STD).

Model & t-test	OPN	CON	EXT	AGR	NEU
SEPRNN [46]	0.7940 ± 0.0030	0.7752 ± 0.0022	0.7195 ± 0.0036	0.7333 ± 0.0107	0.7170 ± 0.0049
LMs + Attention mechanism [47]	0.8038 ± 0.0048	0.7373 ± 0.0143	0.7386 ± 0.0036	0.7352 ± 0.0067	0.7257 ± 0.0031
Attention-based Uni-LSTM [65]	0.7692 ± 0.0412	0.7501 ± 0.0215	0.7008 ± 0.0494	0.7101 ± 0.0518	0.6867 ± 0.0437
Bi-LSTM [48]	0.7546 ± 0.0059	0.7670 ± 0.0020	0.7176 ± 0.0050	0.7428 ± 0.0045	0.7162 ± 0.0077
CNN + Mairesse [91]	0.7489 ± 0.0020	0.7357 ± 0.0026	0.7053 ± 0.0021	0.7093 ± 0.0050	0.6860 ± 0.0042
APR_ConvLSTM (Ours)	$\mathbf{0.8537} \pm 0.0017$	$\mathbf{0.7959} \pm 0.0039$	$\mathbf{0.7734} \pm 0.0029$	$\mathbf{0.7739} \pm 0.0041$	$\mathbf{0.7538} \pm 0.0022$
P-value	< 0.01**	< 0.01**	< 0.01**	< 0.01**	< 0.01**
95% CI	[0.8513, 0.8560]	[0.7904, 0.8014]	[0.7694, 0.7775]	[0.7683, 0.7796]	[0.7508, 0.7569]
Sig.	✓	✓	✓	✓	✓

OPN = Openness, CON = Conscientiousness, EXT = Extraversion, AGR = Agreeableness, NEU = Neuroticism. Asterisks indicate significance level where * corresponds to a p-value < 0.05 and ** corresponds to a p-value < 0.01.

TABLE 12. Comparison of our proposed model APR_ConvLSTM with state-of-the-art APR models on the combined dataset in terms of F1-Score (Mean ± STD).

Model & t-test	OPN	CON	EXT	AGR	NEU
SEPRNN [46]	0.8441 ± 0.0071	0.8613 ± 0.0021	0.7215 ± 0.0184	0.6989 ± 0.0251	0.7175 ± 0.0092
LMs + Attention mechanism [47]	0.8506 ± 0.0031	0.8452 ± 0.0064	0.7233 ± 0.0108	0.7031 ± 0.0147	0.7037 ± 0.0105
Attention-based Uni-LSTM [65]	0.8349 ± 0.0249	0.8488 ± 0.0081	0.7197 ± 0.0393	0.7112 ± 0.0471	0.7071 ± 0.0348
Bi-LSTM [48]	0.8113 ± 0.0110	0.8543 ± 0.0019	0.7026 ± 0.0082	0.7156 ± 0.0093	0.6962 ± 0.0092
CNN + Mairesse [91]	0.8056 ± 0.0032	0.8425 ± 0.0018	0.6783 ± 0.0073	0.6765 ± 0.0132	0.6792 ± 0.0047
APR_ConvLSTM (Ours)	$\mathbf{0.8860} \pm 0.0012$	$\mathbf{0.8619} \pm 0.0039$	$\mathbf{0.7735} \pm 0.0038$	$\mathbf{0.7616} \pm 0.0133$	$\mathbf{0.7452} \pm 0.0092$
P-value	< 0.01**	0.8127	< 0.01**	< 0.01**	< 0.05*
95% CI	[0.8843, 0.8876]	[0.8564, 0.8673]	[0.7683, 0.7787]	[0.7432, 0.7800]	[0.7324, 0.7579]
Sig.	✓	×	✓	✓	✓

OPN = Openness, CON = Conscientiousness, EXT = Extraversion, AGR = Agreeableness, NEU = Neuroticism. Asterisks indicate significance level where * corresponds to a p-value < 0.05 and ** corresponds to a p-value < 0.01.

promising performance over its competitors, with the highest average Accuracy and F1-score of 79.01% and 80.56%, respectively. Training and prediction time are also important metrics that should be considered when comparing deep-learning-based models. From the results, we notice that our model demonstrated significant computational efficiency compared with most baselines, requiring only 247.24 seconds for training.

Although the model proposed in [47] was faster in training (i.e., taking 170.15 seconds), our model remains competitive by consuming only 3.5353 seconds in prediction, which is a fraction of the time compared with other competitors. These results highlight the efficiency of our proposed model in balancing performance and computational time, making it a practical choice for real-time and large-scale applications. Moreover, we notice that the APR_ConvLSTM model has an additional advantage as it is free from feature engineering, which is a labor- and time-intensive task.

2) STATISTICAL SIGNIFICANCE TEST

In this section, a statistical t-test [92] is conducted to examine whether the performance difference between the proposed model and each baseline is statistically significant. Basically, the t-test is run based on two hypotheses:

Null Hypothesis (H_0): There is no significant difference between the overall performance of our proposed model and each baseline.

Alternative Hypothesis (H_a): There is a significant difference between the overall performance of our proposed model and each baseline.

Table 14 represents the t-test results on both Accuracy and F-1 score, in addition to the 95% Confidence Interval

TABLE 13. Comparison of our proposed model and other baselines in terms of accuracy and F1-Score w.r.t. Training Time, Prediction Time, and Handcrafted features used.

Architecture	Accuracy	F1-score	Handcrafted Features Used	Training Time (Sec.)	Prediction Time (Sec.)
SEPRNN [46]	0.7478 ± 0.0049	0.7687 ± 0.0124	×	1945.42	18.1854
LMs + Attention mechanism [47]	0.7481 ± 0.0065	0.7652 ± 0.0091	✓	170.15	10.4628
Attention-based Uni-LSTM [65]	0.7271 ± 0.0031	0.7466 ± 0.0078	✓	5587.69	30.9597
Bi-LSTM [48]	0.7397 ± 0.0050	0.7560 ± 0.0079	×	419.74	151.5145
CNN + Mairesse [91]	0.7170 ± 0.0032	0.7364 ± 0.0060	✓	3347.04	118.52
APR_ConvLSTM (Ours)	**0.7901** ± 0.0030	**0.8056** ± 0.0063	×	247.24	3.5353

(CI) of each model. The results indicate a strong statistical significance between the Accuracy of our model and the SEPRNN [46], LMs + Attention mechanism [47], Attention-based Uni-LSTM [65], and CNN + Mairesse model [91] at the level of p-value < 0.01, and Bi-LSTM [48] at the level of p-value < 0.05.

Moreover, the t-test on the F-1 results finds a strong statistical significance between the performance of our model and LMs + Attention mechanism [47], CNN + Mairesse model [91] at the significance level of p-value < 0.01, and SEPRNN [46], Attention-based Uni-LSTM [65], and Bi-LSTM [48] at the level of p-value < 0.05.

Overall, this analysis demonstrates that the proposed model achieved significant improvement in terms of Accuracy (with 95% CI: 0.7860, 0.7943) and F1-score (with 95% CI: 0.7969, 0.8143) compared to all baselines. From these results, we reject the null hypothesis and accept the alternative hypothesis, which assumes a significant difference between the overall performance of our model and each baseline.

VI. DISCUSSION

This work introduced a unified end-to-end APR model, the APR_ConvLSTM, designed to effectively predict the Big Five personality traits from social media texts. In Section I, we formulated two main research questions to assess the success of the proposed model against machine learning methods and state-of-the-art deep learning models.

To answer **RQ1**, we conducted experiments on two real-world datasets (i.e., PAN-2015 and X-Big5), as well as a combined set of them to compare the APR_ConvLSTM performance with commonly used machine-learning methods. As we can see from Experiment 1 and Experiment 2, our model significantly outperforms all machine-learning methods in predicting all Big Five personality traits on both X-Big5 and the combined dataset. Additionally, it achieved the best performance results on the PAN-2015 dataset across all metrics compared to SVM, Decision Tree, and Naïve Bayes methods. Although almost similar results were achieved by the K-Nearest Neighbors method in predicting the four traits: Openness, Conscientiousness, Extraversion, and Agreeableness, our model outperforms it in predicting Neuroticism with an important increase in accuracy, recall, and F1-score metrics. Overall, although the performance of machine learning methods fluctuates from one dataset to another, our model remains the best performer at each personality dimension, regardless of the dataset used. This demonstrated the strong capacity of the APR_ConvLSTM and the advantage of deep neural networks in the APR task over traditional machine learning approaches.

To answer **RQ2**, we compared the performance of our model on the large dataset (i.e., combined dataset) with the state-of-the-art deep-learning APR models. As we have seen from Experiment 4 and Section V-C, our model significantly outperformed all baselines in predicting each personality trait with the best average accuracy and F1-score of 79.01% and 80.56%, respectively. Overall, the APR_ConvLSTM has two main advantages. First, it is a fast, efficient, and lightweight end-to-end model where all the personality dimensions are predicted simultaneously and independently from each other, instead of training multiple APR models, one for each dimension. Second, it extracts contextual features and complex linguistic patterns from texts without a need for a laborious feature engineering process.

Moreover, our proposed model has significant practical applications in a variety of areas. For example, in human-computer interaction [93], our model can improve user experience by adapting interfaces and interactions based on a user's personality traits inferred from their social media texts. In targeted marketing [17], businesses can optimize advertisements and recommendations by tailoring them to a user's personality traits, enhancing user engagement and satisfaction. Similarly, our model can assist human resource professionals in screening candidates whose personality traits align with specific roles, improving hiring decisions [18]. Additionally, it can be applied to a variety of downstream tasks, such as recommender systems and information seeking. For example, clustering people with the same personality traits and preferences enables industries to make accurate and personalized recommendations.

Apart from the valuable opportunities provided by the APR models, ethical issues of using social media data and personality information must be considered. One of the biggest concerns is the possible exploitation or misuse of users' data without their knowledge or consent [94], [95].

TABLE 14. Statistical analyses of our APR_ConvLSTM and state-of-the-art models in terms of accuracy and F1-score.

Model	Accuracy	95% CI	P-value	F-1 score	95% CI	P-value
SEPRNN [46]	0.7478 $\pm$ 0.0049	[0.7410, 0.7546]	0.0034**	0.7687 $\pm$ 0.0124	[0.7515, 0.7858]	0.0266*
LMs + Attention mechanism [47]	0.7481 $\pm$ 0.0065	[0.7391, 0.7571]	0.0015**	0.7652 $\pm$ 0.0091	[0.7526, 0.7778]	0.0047**
Attention-based Uni-LSTM [65]	0.7271 $\pm$ 0.0031	[0.7227, 0.7315]	0.0038**	0.7466 $\pm$ 0.0078	[0.7357, 0.7575]	0.0121*
Bi-LSTM [48]	0.7397 $\pm$ 0.0050	[0.7326, 0.7467]	0.0179*	0.7560 $\pm$ 0.0079	[0.7450, 0.7670]	0.0142*
CNN + Mairesse [91]	0.7170 $\pm$ 0.0032	[0.7126, 0.7214]	0.0008**	0.7364 $\pm$ 0.0060	[0.7280, 0.7448]	0.0065**
APR_ConvLSTM (Ours)	**0.7901** $\pm$ 0.0030	[0.7860, 0.7943]	-	**0.8056** $\pm$ 0.0063	[0.7969, 0.8143]	-

Asterisks indicate significance level where * corresponds to p-value < 0.05 and ** corresponds to p-value < 0.01.

To mitigate these concerns, it is essential to ensure data transparency, user consent, and promote the responsible use of APR models in real-world applications. Since addressing ethical concerns of APR models is beyond the scope of the current work, it is worthwhile to pay serious attention to them in future work.

VII. CONCLUSION

In this work, we presented a deep learning-based APR model, APR_ConvLSTM, that integrates the merits of CNN and Bi-LSTM to effectively recognize the Big Five personality traits from social media user-generated texts. The APR_ConvLSTM is a unified end-to-end APR model where all personality dimensions are recognized simultaneously without needing a cumbersome feature engineering process. Due to the lack of available datasets, we collected a new personality dataset from the X platform consisting of 31,417 tweets annotated with users' ground truth Big Five personality traits. Several pre-processing methods are applied to this dataset to clean it from noise and less informative data. Then, we performed extensive experiments on X-Big5 and PAN-2015 Author profiling datasets, including 1) comparison with traditional machine-learning methods, 2) investigating the effectiveness of the APR_ConvLSTM over large-scale dataset obtained from combining X-Big5 and PAN-2015 Author profiling datasets, 3) studying the impact of hyperparameters such as kernel sizes, kernel numbers and different word embedding methods on the performance of the APR_ConvLSTM, 4) comparing with state-of-the-art deep-learning APR models. The results show that our APR_ConvLSTM significantly outperforms all machine-learning methods and state-of-the-art baselines in detecting all the Big Five personality dimensions. Overall, it achieved the highest Accuracy and F-1 score of 79.51% and 86.54% on the PAN-2015 dataset and 87.95% and 81.35%, respectively, on the X-Big5 dataset. Moreover, it shows promising performance over its competitors, with the highest average Accuracy and F-1 score of 79.01% and 80.56%, respectively, on the combined dataset. Interestingly, the model reached competitive results in predicting Openness, Extraversion, Agreeableness, and Neuroticism traits with the highest F1 scores of 88.60%, 77.35%, 76.16%, and 74.52%, respectively, on the combined dataset. These results indicate the effectiveness of our model and the powerful learning capability of neural networks, from which the APR can benefit. Additionally, we found that increasing the convolution filter size results in a performance reduction due to noise impact, while increasing the number of filters leads to better recognition results. Glove and Word2Vec: Skip-gram were the best word embedding methods, providing better word representation for the APR_ConvLSTM.

A. LIMITATIONS

While our proposed model demonstrates promising performance in predicting personality traits from social media texts, it has some limitations that need to be considered in future work. First, the datasets used in this study are relatively small. In general, the text-based APR field lacks the availability of large, labeled social media personality datasets. This can be attributed to the cost of data annotations and user privacy issues. Second, we limited our experiments to datasets from the X platform; however, experimenting with datasets from other social networks, such as Facebook, will be more beneficial. Third,, our model is designed to predict the Big Five personality traits from users' texts; however, considering other modalities such as images and videos is worthwhile. Finally, in the related work section, we limited our review to those studies published after 2014, which may lead to excluding some relevant foundational studies.

B. FUTURE WORK

In terms of future work, there are several routes worth considering: (i) Enlarging the dataset collected in this study (i.e., the X-Big5 dataset) to include more representative and diverse instances. (ii) Exploring the model's applicability to other languages and various social media platforms sources (e.g., Facebook, Reddit) and text sources (e.g., essays, articles, news). (iii) Investigating the capacity of other deep-learning methods, e.g., attention mechanisms and various pre-trained language models, e.g., BERT, XLNet, and GPT, in predicting the Big Five personality traits from social media users' texts. (iv) Improving the model by including features from other user information such as audio, images, and video. (v) Enhancing the interpretability of the proposed model by incorporating explainability methods, such as SHAP and LIME, to identify the key linguistic features influencing the prediction of personality traits.

REFERENCES

[1] P. J. Corr and G. Matthews, *The Cambridge Handbook of Personality Psychology*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[2] G. W. Allport, *Pattern and Growth in Personality*. New York, NY, USA: Holt, Rinehart and Winston, 1961.

[3] I. B. Myers, *The Myers-Briggs Type Indicator*. Princeton, NJ, USA: Educational Testing Service, 1962.

[4] K. Lee and M. C. Ashton, "Psychometric properties of the HEXACO personality inventory," *Multivariate Behav. Res.*, vol. 39, no. 2, pp. 329–358, Apr. 2004.

[5] H. J. Eysenck, *The Structure of Human Personality* (Psychology Revivals). Evanston, IL, USA: Routledge, 2013.

[6] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annu. Rev. Psychol.*, vol. 41, no. 1, pp. 417–440, Jan. 1990.

[7] K. Harris and S. Vazire, "On friendship development and the big five personality traits," *Social Personality Psychol. Compass*, vol. 10, no. 11, pp. 647–667, Nov. 2016.

[8] M. Mund, C. Finn, B. Hagemeyer, and F. J. Neyer, "Understanding dynamic transactions between personality traits and partner relationships," *Current Directions Psychol. Sci.*, vol. 25, no. 6, pp. 411–416, Dec. 2016.

[9] P. Gridwichai, A. Kulwanich, B. Piromkam, and P. Kwanmuangvanich, "Role of personality traits on employees job performance in pharmaceutical industry in Thailand," *Syst. Rev. Pharmacy*, vol. 11, no. 3, pp. 185–194, Jan. 2020.

[10] R. Sartori, A. Costantini, A. Ceschi, and A. Scalco, "Not only correlations: A different approach for investigating the relationship between the big five personality traits and job performance based on workers and employees' perception," *Qual. Quantity*, vol. 51, no. 6, pp. 2507–2519, Nov. 2017.

[11] H. T. Bui, "Big five personality traits and job satisfaction: Evidence from a national sample," *J. Gen. Manage.*, vol. 42, no. 3, pp. 21–30, Apr. 2017.

[12] M. Kleć, "The influence of listener personality on music choices," *Comput. Sci.*, vol. 18, no. 2, p. 163, 2017.

[13] I. Cantador, I. Fernández-Tobías, and A. Bellogín, "Relating personality types with user preferences in multiple entertainment domains," in *Proc. UMAP Workshops*, Jan. 2013, pp. 1–16.

[14] A. L. Murray and T. Booth, "Personality and physical health," *Current Opinion Psychol.*, vol. 5, pp. 50–55, Apr. 2015.

[15] J. E. Strickhouser, E. Zell, and Z. Krizan, "Does personality predict health and well-being? A metasynthesis," *Health Psychol.*, vol. 36, no. 8, p. 797, 2017.

[16] G. Claridge and C. Davis, *Personality and Psychological Disorders*. Evanston, IL, USA: Routledge, 2013.

[17] S. Matz and J. B. Hirsh, "Marketing and personality," in *The Wiley Encyclopedia of Personality and Individual Differences: Clinical, Applied, and Cross-Cultural Research*. Hoboken, NJ, USA: Wiley, 2020, pp. 587–591.

[18] S. E. Bekhouche, F. Dornaika, A. Ouafi, and A. Taleb-Ahmed, "Personality traits and job candidate screening via analyzing facial videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1660–1663.

[19] J. Maria Balmaceda, S. Schiaffino, and D. Godoy, "How do personality traits affect communication among users in online social networks?" *Online Inf. Rev.*, vol. 38, no. 1, pp. 136–153, Jan. 2014.

[20] S. Dhelim, N. Aung, M. A. Bouras, H. Ning, and E. Cambria, "A survey on personality-aware recommendation systems," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 2409–2454, Mar. 2022.

[21] J. Lin, W. Mao, and D. D. Zeng, "Personality-based refinement for sentiment classification in microblog," *Knowl.-Based Syst.*, vol. 132, pp. 204–214, Sep. 2017.

[22] C. Esterwood and L. P. Robert, "Personality in healthcare human robot interaction (H-HRI): A literature review and brief critique," in *Proc. 8th Int. Conf. Hum.-Agent Interact.*, Nov. 2020, pp. 87–95.

[23] B. Zarouali, T. Dobber, G. De Pauw, and C. de Vreese, "Using a personality-profiling algorithm to investigate political microtargeting: Assessing the persuasion effects of personality-tailored ads on social media," *Commun. Res.*, vol. 49, no. 8, pp. 1066–1091, Dec. 2022.

[24] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 273–291, Jul. 2014.

[25] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2313–2339, Apr. 2020.

[26] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M. Moens, and M. D. Cock, "Computational personality recognition in social media," *User Model. User-Adapted Interact.*, vol. 26, nos. 2–3, pp. 109–142, Feb. 2016.

[27] D. Azucar, D. Marengo, and M. Settanni, "Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis," *Personality Individual Differences*, vol. 124, pp. 150–159, Apr. 2018.

[28] L. Yang, S. Li, X. Luo, B. Xu, Y. Geng, Z. Zeng, F. Zhang, and H. Lin, "Computational personality: A survey," *Soft Comput.*, vol. 26, no. 18, pp. 9587–9605, Sep. 2022.

[29] C. Stachl, R. L. Boyd, K. T. Horstmann, P. Khambatta, S. C. Matz, and G. M. Harari, "Computational personality assessment," *Personality Sci.*, vol. 2, p. e6115, Jul. 2021.

[30] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Personality predictions based on user behavior on the Facebook social media platform," *IEEE Access*, vol. 6, pp. 61959–61969, 2018.

[31] C. Yuan, J. Wu, H. Li, and L. Wang, "Personality recognition based on user generated content," in *Proc. 15th Int. Conf. Service Syst. Service Manage. (ICSSSM)*, Jul. 2018, pp. 1–6.

[32] S. Han, H. Huang, and Y. Tang, "Knowledge of words: An interpretable approach for personality recognition from social media," *Knowl.-Based Syst.*, vol. 194, Apr. 2020, Art. no. 105550.

[33] S. Poria, A. Gelbukh, B. Agarwal, E. Cambria, and N. Howard, "Common sense knowledge based personality recognition from text," in *Proc. Mex. Int. Conf. Artif. Intell.*, Dec. 2013, pp. 484–496.

[34] R. Mahajan, R. Mahajan, E. Sharma, and V. Mansotra, "'Are we tweeting our real selves?' Personality prediction of Indian Twitter users using deep learning ensemble model," *Comput. Hum. Behav.*, vol. 128, Mar. 2022, Art. no. 107101.

[35] R. L. Boyd and J. W. Pennebaker, "Language-based personality: A new approach to personality in a digital world," *Current Opinion Behav. Sci.*, vol. 18, pp. 63–68, Dec. 2017.

[36] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, "TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning," *Information*, vol. 9, no. 5, p. 127, May 2018.

[37] H. Ahmad, M. Z. Asghar, A. S. Khan, and A. Habib, "A systematic literature review of personality trait classification from textual content," *Open Comput. Sci.*, vol. 10, no. 1, pp. 175–193, Jul. 2020.

[38] K. N. P. Kumar and M. L. Gavrilova, "Latent personality traits assessment from social network activity using contextual language embedding," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 2, pp. 638–649, Apr. 2022.

[39] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 4–21, Jan. 2017.

[40] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," *Mach. Learn. Appl.*, vol. 6, Dec. 2021, Art. no. 100134.

[41] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 604–624, Feb. 2021.

[42] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[43] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. Cambridge, MA, USA: MIT Press, 1986.

[44] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[46] X. Xue, J. Feng, and X. Sun, "Semantic-enhanced sequential modeling for personality trait recognition from texts," *Appl. Intell.*, vol. 51, no. 11, pp. 7705–7717, Nov. 2021.
[47] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *J. Big Data*, vol. 8, no. 1, p. 68, Dec. 2021.
[48] M. A. Kosan, H. Karacan, and B. A. Urgen, "Predicting personality traits with semantic structures and LSTM-based neural networks," *Alexandria Eng. J.*, vol. 61, no. 10, pp. 8007–8025, Oct. 2022.
[49] 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.
[50] J. Yu and K. Markov, "Deep learning based personality recognition from Facebook status updates," in *Proc. IEEE 8th Int. Conf. Awareness Sci. Technol. (iCAST)*, Nov. 2017, pp. 383–387.
[51] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[52] M. Jolly, "Myers-briggs personality type dataset," 2017.
[53] F. M. R. Pardo, F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, "Overview of the 3rd author profiling task at pan 2015," in *Proc. CLEF Eval. Labs Workshop Work. Notes Papers*, 2015, pp. 1–8.
[54] H. Wei, F. Zhang, N. J. Yuan, C. Cao, H. Fu, X. Xie, Y. Rui, and W.-Y. Ma, "Beyond the words: Predicting user personality from heterogeneous information," in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, Feb. 2017, pp. 305–314.
[55] P.-H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, and V. S. Sinha, "25 tweets to know you: A new model to predict personality with social media," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 11, May 2017, pp. 472–475.
[56] N. H. Jeremy, C. Prasetyo, and D. Suhartono, "Identifying personality traits for Indonesian user from Twitter dataset," *Int. J. FUZZY Log. Intell. Syst.*, vol. 19, no. 4, pp. 283–289, Dec. 2019.
[57] V. Ong, A. D. S. Rahmanto, Williem, D. Suhartono, A. E. Nugroho, E. W. Andangsari, and M. N. Suprayogi, "Personality prediction based on Twitter information in bahasa Indonesia," in *Proc. Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, Sep. 2017, pp. 367–372.
[58] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our Twitter profiles, our selves: Predicting personality with Twitter," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 180–185.
[59] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: A review," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, Aug. 2020.
[60] R. Pandey, A. Kumar, J. P. Singh, and S. Tripathi, "Hybrid attention-based long short-term memory network for sarcasm identification," *Appl. Soft Comput.*, vol. 106, Jul. 2021, Art. no. 107348.
[61] S. Peng, L. Cao, Y. Zhou, Z. Ouyang, A. Yang, X. Li, W. Jia, and S. Yu, "A survey on deep learning for textual emotion analysis in social networks," *Digit. Commun. Netw.*, vol. 8, no. 5, pp. 745–762, Oct. 2022.
[62] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao, Z. Wu, X. Zhong, and J. Sun, "Deep learning-based personality recognition from text posts of online social networks," *Appl. Intell.*, vol. 48, no. 11, pp. 4232–4246, Nov. 2018.
[63] X. Sun, B. Liu, J. Cao, J. Luo, and X. Shen, "Who am I? Personality detection based on deep learning for texts," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.
[64] 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.
[65] J. Zhao, D. Zeng, Y. Xiao, L. Che, and M. Wang, "User personality prediction based on topic preference and sentiment analysis using LSTM model," *Pattern Recognit. Lett.*, vol. 138, pp. 397–402, Oct. 2020.
[66] S Dixon. (2022). *Number of Twitter Users Worldwide from 2019 to 2024*. Statistica. [Online]. Available: https://www.statista.com/statistics/303681/twitter-users-worldwide
[67] A. Shrivatava, S. Mayor, and B. Pant, "Opinion mining of real time Twitter tweets," *Int. J. Comput. Appl.*, vol. 100, no. 19, pp. 1–4, Aug. 2014.
[68] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLoS ONE*, vol. 8, no. 9, Sep. 2013, Art. no. e73791.
[69] G. Y. N. N. Adi, M. H. Tandio, V. Ong, and D. Suhartono, "Optimization for automatic personality recognition on Twitter in bahasa Indonesia," *Proc. Comput. Sci.*, vol. 135, pp. 473–480, Jan. 2018.
[70] C. J. Soto and O. P. John, "The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power," *J. Personality Social Psychol.*, vol. 113, no. 1, pp. 117–143, Apr. 2016.
[71] O. P. John, "The big five inventory-versions 4a and 54," Inst. Personality Social Res., University of California, Berkeley, Berkeley, CA, USA, Tech. Rep., 1991.
[72] N. A. Bowling, J. L. Huang, C. B. Bragg, S. Khazon, M. Liu, and C. E. Blackmore, "Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality," *J. Personality Social Psychol.*, vol. 111, no. 2, p. 218, 2016.
[73] C. J. Soto and O. P. John, "Short and extra-short forms of the big five Inventory–2: The BFI-2-S and BFI-2-XS," *J. Res. Personality*, vol. 68, pp. 69–81, Jun. 2017.
[74] C. W. Rudolph and H. Zacher, "Individual differences and changes in personality during the COVID-19 pandemic," *Social Personality Psychol. Compass*, vol. 17, no. 7, Mar. 2023, Art. no. e12742.
[75] A. R. Sutin, M. Luchetti, D. Aschwanden, J. H. Lee, A. A. Sesker, J. E. Strickhouser, Y. Stephan, and A. Terracciano, "Change in five-factor model personality traits during the acute phase of the coronavirus pandemic," *PLoS ONE*, vol. 15, no. 8, Aug. 2020, Art. no. e0237056.
[76] K. Krautter, M. Friese, A. Hart, and D. Reis, "No party no joy?—Changes in university students' extraversion, neuroticism, and subjective well-being during two COVID-19 lockdowns," *Appl. Psychol., Health Well-Being*, vol. 14, no. 4, pp. 1314–1332, Nov. 2022.
[77] Tweepy. *Tweepy, to Collect Tweets*. Accessed: Jun. 12, 2024. [Online]. Available: https://www.tweepy.org/
[78] NoSlang. *Internet, Gamer, & Text Slang Dictionary*. Accessed: Jul. 3, 2024. [Online]. Available: https://www.noslang.com/dictionary/
[79] *Natural Language Toolkit*. Accessed: Jul. 10, 2024. [Online]. Available: https://www.nltk.org/
[80] J.-I. Biel, V. Tsiminaki, J. Dines, and D. Gatica-Perez, "Hi YouTube!: Personality impressions and verbal content in social video," in *Proc. 15th ACM Int. Conf. Multimodal Interact.*, Dec. 2013, pp. 119–126.
[81] Y. Wang, Y. Hou, W. Che, and T. Liu, "From static to dynamic word representations: A survey," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 7, pp. 1611–1630, Jul. 2020.
[82] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*, 2014, pp. 1532–1543.
[83] M. Usama, B. Ahmad, E. Song, M. S. Hossain, M. Alrashoud, and G. Muhammad, "Attention-based sentiment analysis using convolutional and recurrent neural network," *Future Gener. Comput. Syst.*, vol. 113, pp. 571–578, Dec. 2020.
[84] H. Chen, L. Wu, J. Chen, W. Lu, and J. Ding, "A comparative study of automated legal text classification using random forests and deep learning," *Inf. Process. Manage.*, vol. 59, no. 2, Mar. 2022, Art. no. 102798.
[85] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.
[86] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark," *Neurocomputing*, vol. 503, pp. 92–108, Sep. 2022.
[87] D. Pedamonti, "Comparison of non-linear activation functions for deep neural networks on MNIST classification task," 2018, *arXiv:1804.02763*.
[88] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
[89] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 147–160, Apr. 2018.
[90] K. El-Demerdash, R. A. El-Khoribi, M. A. Ismail Shoman, and S. Abdou, "Deep learning based fusion strategies for personality prediction," *Egyptian Informat. J.*, vol. 23, no. 1, pp. 47–53, Mar. 2022.
[91] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 74–79, Mar. 2017.

[92] J. C. F. De Winter, ''Using the student's t-test with extremely small sample sizes,'' *Practical Assessment, Res., Eval.*, vol. 18, no. 1, p. 10, 2019.

[93] T. Alves, J. Natálio, J. Henriques-Calado, and S. Gama, ''Incorporating personality in user interface design: A review,'' *Personality Individual Differences*, vol. 155, Mar. 2020, Art. no. 109709.

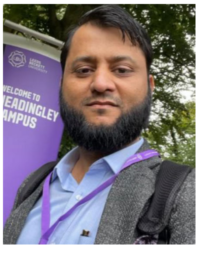
[94] L. V. Phan and J. F. Rauthmann, ''Personality computing: New frontiers in personality assessment,'' *Social Personality Psychol. Compass*, vol. 15, no. 7, Jul. 2021, Art. no. e12624.

[95] L. Alexander, E. Mulfinger, and F. L. Oswald, ''Using big data and machine learning in personality measurement: Opportunities and challenges,'' *Eur. J. Personality*, vol. 34, no. 5, pp. 632–648, Sep. 2020.

NOUF M. ALJUHANI received the B.S. degree in information systems from Taibah University, Madinah, Saudi Arabia, in 2018. She is currently pursuing the M.Sc. degree in information systems with King Abdulaziz University (KAU), Jeddah, Saudi Arabia. Her research interests include affective computing, social network analysis, text-mining, machine learning, deep learning, and natural language processing.

ABDULLAH AL-MALAISE AL-GHAMDI received the Ph.D. degree in computer science from George Washington University, Washington, DC, USA, in 2003. He is a Professor of software and systems engineering and AI with the Information Systems (IS) Department, Faculty of Computing and Information Technology (FCIT), King Abdulaziz University (KAU), Jeddah, Saudi Arabia. He has published two books and more than 70 research articles in well-known venues, such as the Web of Science, Springer, IEEE, and Taylor & Francis. His current researches under development include AI in digital education, a machine learning approach on Corona disease, and cognitive assessment on medical data. His areas of research interests include software engineering and systems, artificial intelligence, data analytics, business intelligence, IT-business-oriented architecture, business process re-engineering, data mining, and decision support systems.

HANAN S. ALGHAMDI received the Ph.D. degree in computer science from the University of Surrey, Guildford, U.K., in 2018. She is currently an Assistant Professor with King Abdulaziz University, Jeddah, Saudi Arabia. Regarding professional experience, she was the Strategic Communication Manager with the Savola Group, Saudi Arabia; and a Demonstrator with the University of Surrey, U.K. She has published several papers in the field of medical images analysis, e.g., Measurement of Optical Cup to Disk Ratio in Fundus Images for Glaucoma Screening; Ensemble Learning Optimization for Diabetic Retinopathy Image Analysis; Automatic Optic Disc Abnormality Detection in Fundus Images: A Deep Learning Approach; and Deep Learning Approaches For Detecting COVID-19 From Chest X-Ray Images: A Survey. Her current research interests include medical image analysis, explainable artificial intelligence, deep learning reinforcement learning, generative adversarial networks, evolutionary algorithms, and optimization. She is involved in several research activities and has participated as a reviewer for multiple journals and conferences locally and internationally, e.g., IEEE ACCESS, *Scientific Reports*, International Conference on Computational Intelligence and Intelligent Systems (CIIS), International Conference on Machine Learning and Human-Computer Interaction (MLHMI), Asia Conference on Algorithms, Computing and Machine Learning (CACML), and the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).



FARRUKH SALEEM received the Ph.D. degree in computer science from the University of Technology Malaysia. He is currently a Senior Lecturer with the School of Built Environment, Engineering, and Computing, Leeds Beckett University, Leeds, U.K. Previously, he was an Assistant Professor with the Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. He has also supervised students' research projects and theses for undergraduate, postgraduate, and executive M.S. with a wide range of research topics related to the computer science and machine learning fields. His research interests include machine learning, data science, artificial intelligence, decision support systems, business intelligence, and enterprise architecture.