Personality Trait Classification Using CNN-LSTM Model

Joffin George, Koshy M Varkey, Vidul Venogopalan, Rahul G
Ms.Neema George
Dept. of CSE,Mangalam College of Engineering
Ettumanoor,Kerala,India

Mail_id:joffingeorge10@gmail.com, koshymvarkey2000@gmail.com, vidulvenugopal2001@gmail.com, rahulgpalakuzhiyil@gmail.com, neema.george@mangalam.in

Abstract—Cognitive base sentiment analysis for understanding the behaviour of the user on the basis of tweets post by them in their social media has become a common technique nowadays. These techniques are not showing accurate result. In our paper we come forward with a hybrid based deep learning model called convolution neural network with LSTM to improve the efficiency of current technique. We classify the personality trait into 8 types they are introversion, extroversion,intuition,sensing,thinking,feeling,judging,perceiving .This type of method have shown better result in comparison with the existing system. This technology is helpful in recruiting people for various jobs. We also introduce the concept of audio to text conversion where user give their audio to the system and our system will tell the emotions based on given audio.

Keywords—Personality Trait,CNN,LSTM,Keras Tokenizer,Deep Learning

I. INTRODUCTION

Cognitive science deals with various emotions of the people on the basis of their tweets.Personality plays an important role in determining the individual. Personality of an individual can be judged from various parameters such as text,audio,video. CNN extracts the basic features of the sentences without storing the previous information. We introduce a technique which uses both CNN and LSTM technology to enhance the feature of existing system.The present system aims to classify user behaviour based on various deep learning models.We also introduce the concept of audio to text using the same CNN+LSTM model to improvise the existing system.

A.RESEARCH STUDY MOTIVATION

Various studies have been conducted by the research scientist for personality detection.The personality trait is a classification problem where the user gives the input tweet as text and he gets the various emotions corresponding to the text he has given.The aim of this paper is to build a strong model for personality trait detection.We have catagorized the various emotions as a pair like I-E,N-S,T-F,J-P. We classify the various tweets into this category.There is also provision for user to give their tweets in the form of audio and our system will detect the personality of individual based on the audio.

B. PROBLEM BACKGROUND

Cognitive-based SA applications have gained popularity in recent years among online communities as a way to learn about people's attitudes and personality traits towards various topics, laws, and other things. However, because. It takes a lot of time to analyze text using the present techniques to find personality traits in such content because of the diversity of social media information. Therefore, it has become essential to automatically classify personality traits for use in social media content extraction and analysis. We have all seen a lot of research in the fields of text-based SA, lexicon generation, cognition, aspect-based SA, and visual SA. However, further study on cognitive-based social media is needed, with an emphasis on extracting and classifying personality features. Our suggested method can resolve both issues, but the current system cannot handle both audio and text transformation.

C.RESEARCH PROBLEM

The present system for personality trait classification has limited number of models. These techniques uses old models and they need to be improved for improving the accuracy of the system. We treat personality trait as a classification problem and which need to be resolved for future. Furthermore we include the concept of speech to text where user will give his tweet in the form of audio and our training model will identify the emotion from that audio.Through this we show that our model is self sufficient to solve all kinds of transformation.

D. CONTRIBUTION

1.Exploring LSTM model and capturing information from text and training it using CNN and storing it using LSTM.
2.We used SVM and conducted various test like logistic regression,decision tree,k nearest neighbour.
3.our proposed system has showed very good performance and result in against of the existing system models.
4.The proposed system can help various companies to analyse the personality of their employees.
5.Furthermore we provide both speech to text as well as audio to text so there would be added benefit for users to use our system.

II. REVIEW OF LITERATURE

In this section, a comprehensive study of personality trait classification. The study discusses various practices and approaches that have address the problem of personality prediction, and the methods that have been used for this purpose. Specifically, the literature focuses on machine learning approaches to personality recognition, and the following studies are reviewed in detail :

A.Sentiment Analysis of Arabic Tweets from Twitter

Arabic sentiment analysis, however, poses challenges due to the informal and noisy nature of the language and its rich morphology. For Arabic sentiment analysis, they still require improvements in terms of accuracy and efficiency. In order to address this issue, an approach called corpus-based approach is used for Arabic sentiment analysis of tweets from Twitter. A Discriminative Multinomial Naïve Bayes (DMNB) method along with N-grams tokenizer, stemming, and term frequency-inverse document frequency (TF-IDF) techniques is utilized. The proposed approach is evaluated using a set of performance evaluation metrics on a public Twitter dataset. The experimental results shows the effectiveness of the presented approach, which outperformed related works and improved accuracy by 0.3%.

B. Recognizing Personality from reading text speech

This study depicts the relationship between an individual's reading text speech and the personality traits using the Five-Factor Model of Personality. This study involves 140 subjects whose reading text speech was determined with the help of Open SMILE toolkit and ComParE 2013 audio feature set that is used for feature extraction. Kernel SVM classifier was used for classification, along with five filter feature selection approaches and Principal Component Analysis. SVM models were trained individually for each trait using repeated cross validation and five different feature sets. One of the best achieved Unweighted Average Recall (UAR) ranges from 74% to 80% depending on the specific trait being analyzed. These results suggest that automatic identification of speaker personality based on reading text speech is a promising area for further research.

C. System for Personality and Happiness Detection

This study propose a platform for assessing personality and happiness based on Eysenck's theory. Their platform collects text messages from social media, specifically WhatsApp, and applies machine learning algorithms to classify them into distinct personality categories. Although the relationship between personality features and happiness is not yet clear, future correlations may emerge. The platform is described in detail, and various sources of messages are used as a proof of concept. Researchers have traditionally used both direct (e.g., the EPQ-R questionnaire) and indirect methods to gain insight into human personality, with written text being one of the latter. Because personality is thought to be consistent across situations and time, trained psychologists can infer a person's personality profile by observing their behavior. Based on existing research, it is reasonable to assume that individuals will exhibit unique written expression patterns that correspond to their distinct personalities.

D. Personality Detection of players in an educational game

This study discusses the use of Educational Data Mining (EDM) to model student behavior and personality in Intelligent Tutoring Systems (ITS). Specifically, the authors introduce an approach using data mining techniques and NLP to automatically detect student personality and behavior in an educational game. The framework relies on the classification of input excerpts into six different personality classes, using algorithms such as Naive Bayes, Support Vector Machine (SVM), and Decision Tree. Traditional techniques for detecting psychopathy, such as the Hare Psychopathy Checklist and the Psychopathy Checklist-Revised (PCL-R), rely on manual assessment. However textual content from social media can also be used to detect the personality traits of individuals. . Earlier studies have applied supervised machine learning approaches, such as SVM, NB, and DT, to identify personality traits in students during educational games. The results showed that using n-grams as features gave the finest performance as compared to other feature sets.

E. Predicting the Big Five Personality Traits Using Facial Images of Students

This study shows how to predict college students' personality characteristics with static facial images. It focuses on the relationship between self-reported personality characteristics and facial features.. To succeed this, they constructed a dataset that contains 13,347 data pairs composed of facial images and personality characteristics and trained a deep neural network with 10,667 sample pairs from the dataset and used the remaining samples to test (1335 pairs) and validate (1335 pairs) self-reported Big Five personalities... The results show that personality traits can be reliably predicted from facial images with an accuracy that exceeds 70%. In case of five-character tag classification, the recognition accuracy of neuroticism and extroversion was the most accurate, and the prediction accuracy exceeded 90%.

III. PROPOSED SYSTEM

A.MOTIVATION

Various deep learning models, including CNN, GRU, RNN, and LSTM, have been utilized for personality classification. However, all these models alone fail to capture semantic information effectively. The combination of deep learning models, such as CNN + LSTM, which allows us to take advantage of two models, CNN and LSTM, to capture context information more effectively. Moreover, using the LSTM model helps comprehend the context more efficiently by saving information in one direction. Our research project study aims to classify personality traits, such as 'I(Introversion)-E(Extroversion)', 'N(Intuition)-S(Sensing)', ' T(Thinking)-F(Feeling)' and ' J(Judging)-P(Perceiving)' from textual data. To achieve this goal, we propose implementing a deep neural network model called Convolutional Neural Network including Long Short-Term Memory (CNN+LSTM), which demonstrates great potential.

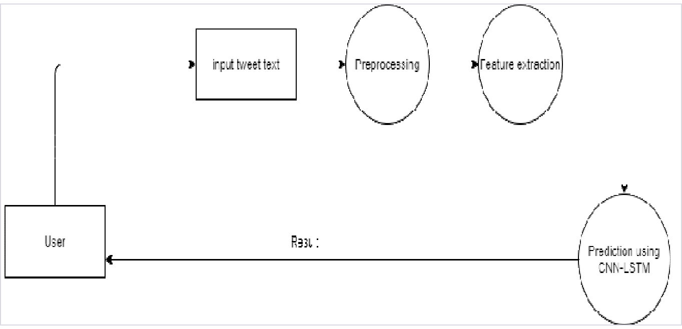

Fig 1: Implementation of a Deep Neural Network

Referring to Figure 1, the suggested approach for categorizing personality qualities from social media texts entails a number of modules, including input tweet text, data pre-processing, feature extraction, and the use of a deep neural network. Getting the required information is done in the first module, then pre-processing is done to get the social media reviews ready for analysis. After that, raw data are converted into a numerical representation as part of the feature extraction process. A deep neural network is then utilized in the final module to turn them into a machine-readable format represented by a real-valued vector. Using word embeddings, the words are mathematically encoded in this process and sent into the hidden layers, which use CNN and LSTM models. The LSTM model learns long-term knowledge to effectively identify user evaluations based on several personality qualities including "I-E," "N-S," "T-F," and "J-P."

The CNN model extracts the significant features from the input data. For a comprehensive description of the four modules, please see the following sections.

B. INPUT TWEET TEXT

We carried out classification trials on personality traits using the MBTI dataset, which has 8675 rows. According to the four personality characteristic categories "I-E," "N-S," "T-F," and "J-P" (see Table 1), each review in the dataset is given a distinct class. The Anaconda Jupyter notebook and the Python programming language were used to carry out the research. Using the ternary sets approach, the dataset was divided into three sets: validation data, test data, and train data (see Fig. 2).

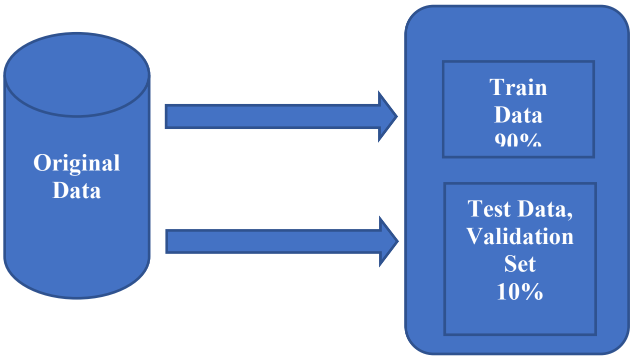

Fig 2. Original data partitioning

TABLE 1. MBTI dataset in detail

Dataset	URL	Description
MBTI	https://www.kaggle.com/datasnaek/mbti-type	Mitchell J developed the dataset, which contains 8676 instances of tweets and personality categories.The dataset is 25 MB in size.

1) TRAIN DATA

The train data is used to train the model by inputting the desired outcomes. To create the model, 90% of the data is used as the training set, although this can vary for different experiments. The model is fitted using the training data. An illustration of the train set, which is stored as a CSV file, is shown in Table 2.

TABLE 2. Train set examples that are related to personality trait classification

Review Text ID	Review Text	Label
1	I firmly believe in the ability to protect others and provide a voice to those who lack it. I hope you will watch this movie I've made with a kind attitude.	Thinking -Feeling
2	Sometimes I have the same thought. The reason for this is that, although I am an INFJ myself, I often find that I don't comprehend them.	Thinking -Feeling
3	The ISFPs I know are endearing, thoughtful individuals. They are souls.	Judging-Perception
4	I got 593. I am, however, a 953, according to what I've read about the Enneagram. I once read that many 9s mistakenly identify as 5s.	Intuition-Sensing
5	I apologise, but I still can't figure out the sequence; at the moment, I have you as XNFP leaning INFP. Could you please tell us about your upbringing? What kind of a kid were you? Did you always seek out new things?	Introversion-Extroversion

2) VALIDATION DATA

In the training phase, the model often displays excellent accuracy, but its effectiveness decreases during the testing phase. The use of a validation set is required to handle under- and over-fitting. A validation set of 10% was used for this investigation. Kera supports human data validation and automated data validation as two parameter tweaking strategies. In this study, we have opted to manually validate the data.

3) TEST DATA

The testing data set influences how well the model performs with fresh, untested data. After the model has been fully trained using the training and validation sets, this test is carried out. We used a 10% distinct sample of test data that wasn't part of the training set to assure objectivity. This data set was used to test the model, and Table 3 is an illustration of it. Using Scikit-Learn's train-test split technique, the data set was divided into 90:10 ratios, with 10% serving as the validation set. To adjust the hyperparameters and evaluate the model's design and performance, we used a validation set. Table 4 provides more information about the data set. After collecting the data, it is transferred to the pre-processing module.

TABLE 3: Test set examples related to personality trait classification

Review Text ID	Review Text	Label
5.	I was not the most clear with my example, you're right. Hmm... how about: I appreciate my trainer, but I wouldn't shed a tear if he got hit by a bus. vs I love my trainer, but I...	J(Judging)-P(Perception)
7.	Your post really makes a lot of sense to me because I also feel like I need some sort of real meaning that I can touch and feel within me. I want my actions in this life to matter in the long run and...	T(Thinking)-F(Feeling)

TABLE 4. Dataset description

Task	Dataset with title	Number of Review Text	Labels
Personality trait classification	DS: MBTI	8675	I(Introversion)-E(Extroversion) (6676-1999),J(Judging)-P(Perception) (5241-3434), N(Intuition)-S(Sensing) (7478-1197), T(Thinking)-F(Feeling) (4694-3981)

C. DATA PRE-PROCESSING

The suggested solution makes use of a second module that is concentrated on performing fundamental pre-processing operations on the given dataset. In order to remove any duplicate words from the file, these steps involve (i) lower-casing the reviews using a Python-based code, (ii) removing stop words like "the," "is," and "and," and (iii) tokenizing the reviews using the Keras tokenizer to separate them into tokens. With the help of this tokenizer, tokens may easily be transformed into numeric values.

D. FEATURE EXTRACTION

Selecting pertinent data features from the input data that best reflect personality qualities is the initial stage when feature extraction is used to classify personality traits using CNN and LSTM. These characteristics could be obtained from the type of input data. Raw data must be converted into a numerical representation that the CNN and LSTM networks can understand as part of the feature extraction process. This could entail methods for processing visual data, such as image processing, spectrogram analysis for audio data, or word embedding for text data. The CNN and LSTM networks are trained to categorise personality traits based on the recovered features after feature extraction. The LSTM network is in charge of capturing the temporal relationships in sequential data, whereas the CNN network is in charge of learning spatial properties in the data. The CNN and LSTM networks learn to categorise personality traits with a high degree of accuracy by utilising the representative features collected from the input data through cycles of training and testing.

E. IMPLEMENTATION OF A DEEP NEURAL NETWORK

In the last step, a deep neural network is used to categorise personality traits from the input text. Three layers make up the CNN+LSTM model: the input layer, the hidden layer, and the output layer. Here is an explanation of these several layers:

1) INPUT LAYER

The deep neural network's input layer is in charge of taking in incoming data. The embedding layer of Kera's library is used to convert words into real-valued vectors. Semantic information is captured by this numerical representation.

2) HIDDEN LAYER

Multiple CNN and LSTM layers make up the deep neural network's hidden layer. The following are the layers and components of the CNN model:

a. Convolutional Layer

With the use of a linear procedure known as convolution, this layer collects features from the incoming data. The input data is sent through the filter, and the resulting feature map is activated nonlinearly using a function like "Relu" to eliminate negative values.

b. Pooling Layer

The input from the previous layer is used to perform a downsampling procedure known as Maxpooling, which lowers the volume of the feature map after convolution.

c. LSTM Layer

The LSTM layer is added to learn long-term information. It incorporates input from the CNN model and keeps both recent and old data. It can also memorise long-term memories and retain knowledge for lengthy periods of time. The results of the LSTM layer are then added to the output layer.

3)OUTPUT LAYER

After feature extraction, downsampling, and long-term memory at the convolutional, pooling, and LSTM layers, respectively, the output layer of the deep neural network categorises the learned features. The input text, such as "I am finding the lack of me in these posts very alarming," is classified in this layer using the "softmax" function into four personality characteristic classes: "I-E," "N-S," "T-F," and "J-P." The target label (class), which is generated by the softmax function, is given to the input text by the class with the highest likelihood.

IV. EXPERIMENTS AND RESULTS

We tried various CNN+LSTM models with parameters having different values for the categorization of input text over different personality traits. We used a single layer with a variety of parameters to get a best output. As a result, optimising CNN+LSTM's parameters increases the classifier's effectiveness. We used various settings for the LSTM layer's "units" parameter.The various layers used in our model,their parameters and values are listed out in Table 5.

TABLE 5. Our proposed model parameters and their values

Proposed model Layers	Parameters and their values

CNN Layer	Kernal size =3x3
	Filters =64
	padding = same
	layers of pooling =2

Layer of units LSTM	50
Dropout Layer	rate=0.2

Dense Layer	classes=2
	activation =relu

Further parameters	length of input size=823
	epochs=5
	batch_size=32
	output_dimension=500

Experiment predicting Introversion-Extroversion

In this Experiment, we test the effectiveness of the proposed CNN+LSTM model to predict the accuracy whether the input text or the speech is an Introversion-Extroversion. The findings are shown in Table 6. The overall accuracy that we have got is 0.88%. After completing the experiment our model performed well for the Introversion personality characteristic because its F-measure is greater than the F-measure for Extroversion.

TABLE 6. Experiment results of the proposed model for the Introversion-Extroversion personality characteristics.

Personality Characteristic	F-measure	Accuracy
Introversion	0.92	0.88
Extroversion	0.72	

Experiment predicting Intuition-Sensing

In this Experiment, we test the effectiveness of the proposed CNN+LSTM model to predict the accuracy whether the input text or the speech is an Intuition-Sensing. The findings are shown in Table 7. The overall accuracy that we have got is 0.91%.After completing the experiment our model performed well for the Intuition personality characteristic because its F-measure is greater than the F-measure for Sensing.

TABLE 7. Experiment results of the proposed model for the Intuition-Sensing personality characteristics

Personality Characteristic	F-measure	Accuracy
Intuition	0.95	0.91
Sensing	0.62	

.

Experiment predicting Thinking-Feeling

In this Experiment, we test the effectiveness of the proposed CNN+LSTM model to predict the accuracy whether the input text or the speech is an Thinking-Feeling. The findings are shown in Table8. The overall accuracy that we have got is 0.85%.After completing the experiment our model performed well for the Feeling personality characteristic because its F-measure is greater than the F-measure for Thinking.

TABLE 8. Experiment results of the proposed model for the Thinking-Feeling personality characteristics

Personality Characteristic	F-measure	Accuracy
Thinking	0.84	0.85
Feeling	0.86	

Experiment predicting Judging-Perception

In this Experiment, we test the effectiveness of the proposed CNN+LSTM model to predict the accuracy whether the input text or the speech is an Judging-Perception. The findings are shown in Table 9. The overall accuracy that we have got is 0.80%. After completing the experiment our model performed well for the Perception personality characteristic because its f-measure is greater than the f-measure for Judging.

TABLE 9. Experiment results of the proposed model for the Judging-Perception personality characteristics

Personality Characteristics	F-measure	Accuracy
Judging	0.70	0.80
Perception	0.80	

V. CONCLUSION

In this method we predicted the personality of the individual using text and audio.By applying deep learning model and by concating CNN+LSTM we have helped achieve this target. With the help of max pooling layer we are able to extract basic feature of the in dividual.We are taking the data sequentially and with the help of LSTM we are able to get the information of prior data.This model has an average accuracy of 88%.

VI. REFERENCES

[1] 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.

[2] H. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. A. Hameed, ‘‘Applying deep learning technique for depression classification in social media text,’’ J. Med. Imag. Health Informat., vol. 10, no. 10, pp. 2446–2451, Oct. 2020, doi: 10.1166/jmihi.2020.3169.

[3] R. Katarya and P. Srinivas, ‘‘Predicting heart disease at early stages using machine learning: A survey,’’ in Proc. Int. Conf. Electron. Sustain. Com mun. Syst. (ICESC), Jul. 2020, pp. 302–305.

[4] Rahul and R. Katarya, ‘‘A review: Predicting the performance of students using machine learning classification techniques,’’ in Proc. 3rd Int. Conf. I-SMAC, Dec. 2019, pp. 36–41.

[5] G. Kou, Y. Xu, Y. Peng, F. Shen, Y. Chen, K. Chang, and S. Kou, ‘‘Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection,’’ Decis. Support Syst., vol. 140, Jan. 2021, Art. no. 113429.

[6] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, ‘‘Per sonality trait detection using bagged SVM over BERT word embedding ensembles,’’ 2020, arXiv:2010.01309.

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

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

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

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

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

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

[13] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and S. Khan, ‘‘Classification of poetry text into the emotional states using deep learning technique,’’ IEEE Access, vol. 8, pp. 73865–73878, 2020.

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

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