Jestr

Journal of Engineering Science and Technology Review 15 (4) (2022) 25 - 32

JOURNAL OF
Engineering Science and Technology Review

Research Article

www.jestr.org

Convolution Neural Network Regression Model to Predict Personality Scores

Mamta Bhamare[1,2,*] and K. Ashokkumar[1]

[1]*Department of Computer Science and Engineering,Sathyabama Institute of Science & Technology, Chennai, India*
[2]*School of Computer Engineering and Technology, MITWPU, Pune, India*

Received 4 February 2022; Accepted 30 July 2022

Abstract

The world has witnessed tremendous increase in the use of social media platforms in the last decade. Posting and updating on such platforms have become a way for people to express their opinions, experiences, feelings, views etc. Along with all these, they tend to post and share their personal Information. This Information can be used in advantageous way to understand their needs and requirements The approach, known as personality prediction, entails sepAarating digital input into attributes and mapping it according to a personality model. In this paper, we are interested to predict the personality of such users. Researchers have shown based on the social behavior towards friends, number of friends, groups joined etc. can help in predicting the personality trait from Facebook and Twitter. The text posts for the user are converted to be suitable representation using GloVe algorithm. And then feeding it to Convolution Neural Net, a popular Deep Learning architecture on the text posts that was implemented for the traits present in the Big Five trait theory. This method helped to achieve comparable outputs with other recent models.

Keywords: Personality; Personality Prediction; Social media; Big Five trait; Deep Learning; GloVe.

1. Introduction

For past decade there has been boom in the usage of social media sites. According to the survey conducted in July 2020, there were 3.96 billion users of various social networking sites like Facebook, twitter, etc. out of 7.79 billion populations. Thus, it accounts to be 51% of total world's population. Facebook has the most users, with 1.8 billion users, and roughly 800 million users spending about 40 minutes every day on it [7]. Earlier it was considered only as the way of entertainment for youths but later it had become a way of communication for the diversity of population. It has removed the barrier in communication that is Distance. Social media users tend to share their emotions, feelings their good and bad experiences while revealing personal details like age, profession, what are their likes and dislikes, expectations. Researchers have always been curious to find out the ways in which the personality can be predicted. It is often said that the words that are used by the person reflect their behavior and personality. So, the posts of such digital platforms can be used to predict the personality and study the behavioral psychology. Personality is often referred as the union of emotion, thinking pattern, etc.

Studies in the field of psychology showed that there is a correlation between personality and the linguistic behavior of a person's [26,27]. All the humans are born with their distinct set of all these characteristics. These characteristics can be studied, and many inferences and insights can be drawn from them. It has wide applications in various fields like Medical, Marketing etc. In medical, the text can help to find the mental health for a person as there is relation between the mental health and the way of expressing thoughts. For marketing, the prediction of personality can help to build recommender systems and target marketing using these digital platforms. Some other applications include eligibility for job, relationship success etc. Personality psychology is the branch that studies the difference between the individual's behavior and actions in some situations. It is believed that personalities are long-term, stable and not readily altered [20,28]. It had been studied for several thousands of years and is still researched. There are three criteria that are used to characterize personality traits:

- Consistency: - This means that person must be consistent across situations that are related to the trait. That is example; if a person is talkative at home, they tend to be talkative in college.
- Stability: - This is related to the time, Example if a person is talkative at age of 30, he will be talkative at age of 40 also.
- Individual differences: - Different Individuals have different behaviors.

Earlier the prediction was done just by observing the people, identifying the differences in the behavior in varied situations. But this was a long process and could help in studying small subset of people, but for large group this way of getting personality traits is insufficient. Then, personality assessment was done through rating scales. In this there are group of Questions that are related to particular trait and few choices for each question. The answer chosen by the user is rated based on the proportionality of that question with the trait. This technique for prediction is accurate when the person knows himself. But there are people who do not know about themselves, so this was short coming of this method, and it was very tedious way to predict for a very large group of people. So, there was a need to automate the prediction process and use the latest technology for prediction. Increasing use of social media sites to interact in the virtual world had led to increase in number of researchers to predict the personality. There came

*E-mail address: mamta.bhamare@mitwpu.edu.in
ISSN: 1791-2377
doi:10.25103/jestr.154.04

into existence use of Machine learning and Deep learning approaches to automate the process of prediction. To predict the personality using this approach, there are two necessities: Trait Model and Algorithm.

- Theories help to organize the various characteristics in people and help to get in-sights about patterns in which other people behave.
- Algorithms mean the models that are used for automation.

These theories are of many varieties. Mainly two theories are used Type and Trait theory. In type theory it is believes that personality is made up of few clearly defined and distinct categories. On other hand, trait theory identifies individual differences. According to this theory, person's personality is result of some combined characteristics. Most researchers have employed Big five and MBTI (Myers Briggs Type Indicator) personality models to predict personality. Some researchers have employed the DISC (Dominance, Influence, Compliance, and Stability) framework. Personality prediction is a task that identifies information about an individual's personality trait from a given collection of facts. Various studies have been conducted to far in order to predict a person's personality based on publicly available information on social media. Researchers employed the information available in online social networking accounts in the form of status updates to forecast user personalities using standard PRT methodologies [21,22,23].

This paper first discusses the Big Five Factor model and the related work done in personality prediction from social media post. We discuss the dataset used, preprocessing, word embedding and Convolutional Neural Network. The next section discussed the experimental setup that is used for the proposed architecture. In the last section results are discussed. The contributions of this work can be summarized as follows:

1. With the advances in the use of Deep learning for text classification, we propose ConvNet based approach to predict the personality which helps to increase the prediction accuracy.
2. Unlike the other technique, we encouraged merging different sources of social media data to enhance the number of datasets for better classification.
3. We assessed the model's performance and compared it to previous research algorithms that provide the best performance in predicting personality.

We demonstrated that our methods outperform prior studies in terms of predicting personality traits.

2. Related Work

Psychology, the study of Personality earlier based on survey or questionnaire. Such way to get personality of every individual is impossible task. So, there is need to automate this process. Social sites have large number of contents in the form of Text, Videos, Images, and Audio. Extensive literature survey [2] has been done to predict the personality on basis of the information from the social media using numerous techniques like Machine Learning, Deep Learning, and Data Mining etc. Figure 1 shows the approaches that are used in the literature for prediction of personality.

2.1. Approaches Used

- Questionnaire: Earlier method to predict the personality was by using a set of questions in multiple choice form. Users were asked to answer those questions and based on the answer chosen, the score was given. For each Big Five Trait model, there were few related questions.
- Semantic approach: This is one of the simplest approaches towards predicting the personality. User text is represented in the form of some numerical vector. Semantic metrices then are used to measure the similarity between the generated vector and the vector that represents the personality traits based on the model used.
- Machine Learning: Classical approaches cannot handle vast amount of data [29]. Machine Learning algorithm can find out the hidden patterns from the data after extracting the features. Classification is used to classify the low and the high personality characteristics and regression to predict the personality score for everyone.
- Deep Learning: As for most visual, deep learning exercises, the most widely used and advanced personality recognition method is the Convolutional Neural Networks (CNN). Deep learning tries to mimic Human Brain learning process that is learning with help of examples. They are based on ANN that can have many hidden layers it can up-to 150.

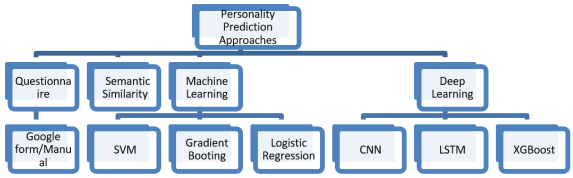

Fig. 1. Taxonomy of various approaches.

2.2. Literature work

The use of Facebook and Twitter datasets to predict personality is not novel. Authors [10, 11, 12, 13] conducted study using an open-source Facebook personality dataset called myPersonality, which comprises of 250 individual's status data and attributes and maps to the big five personality model. The major feature extraction approach is Linguistic Inquiry and Word Count (LIWC), which is a linguistic analytical tool that aids in the analysis of quantitative texts by calculating the number of words that contain the meaning of categories based on a psychological lexicon. As evidenced by the growing use of various analytical tools,

Using human way of understanding the languages and interpreting them called as Semantic approach [1] was used for personality prediction. It was done after converting the text in the form of vectors and then finding the similarity between the words that are related to personality traits. Using the brands pages that are on the social media sites, posts or comments related to that Brand can be extracted and algorithms such as SVM, Gradient Boosting and CNN can be used to automate the process of prediction. Wiu et.al [3] extracted the semantic vectors by proposing a new kind of algorithm called as AttRCNN, a 2 – layer hierarchical model of deep learning approach. Deep semantic features are combined with statistical linguistic features extracted directly from text posts and fed into traditional regression algorithms

to predict the real-valued Big Five personality score. Deep semantic features are fed into regression algorithms in order to increase the prediction accuracy of personality detection systems even further. Pearson's correlation [4] was implemented to get the relationship that exists between the text words, additionally, algorithms like XGBoost, GB, SVM and Logistic Regression were used to train model on the collected dataset based on five factor model. Using individual SNA feature sets, the XGBoost machine learning approach produced the best personality prediction. The best accuracy was obtained for extraversion as the trait most frequently indicated by Facebook features. Because the experiment relied on a small number of items from the myPersonality sample dataset (250 users, 9917 status updates), the accuracy of the results was somewhat constrained.

Xue et al. [5] used microblogging site called as Sina Weibo to collect the status updates of the users. A set of questionnaires was given to the users who participated that has 44 questions, from this their personality scores were obtained. Features were extracted of various categories like Static, dynamic and content features. LDL algorithms like Random Forest, LR, K nearest neighbor, SVM, SVR, and MLP were implemented. Optimal feature space extraction can be performed on bigger data sets using deep learning algorithms to increase the prediction accuracy of LDL approaches in personality recognition. N. Majumder [6] presented a model based on a CNN features extractor. Sentences from the essays are fed into convolution filters, which produce the sentence model as n-gram feature vectors. After preprocessing, this document is fed into a vector, which is then fed into a fully connected neural network with one hidden layer. Additional features extraction and preprocessing are both possible. Long Short-Term Memory (LSTM) recurrent network can be used to construct both the sentence vector and the document vector from a succession of sentence vectors. As a strategy for extracting characteristics from social media data sources. The previous study also revealed another reliable open vocabulary feature extraction method known as the National Research Council (NRC) emotion lexicon database. This corpus was created by the National Research Council of Canada and contains approximately 14,000 English words as well as the associations of these words with eight common emotions, namely anger, fear, anticipation, trust, surprise, sadness, joy, and disgust, as well as the sentiments of each of these words, which can be positive or negative [18]. Furthermore, in order to use such an open vocabulary feature extraction approach, the dataset must first be translated into English before the feature extraction method can be used. In these experiments, an algorithm was utilized.

The proposed system by the author [19] was trained using three well-known machine-learning algorithms: a Naive Bayes classifier, a Support Vector Machine, and a Multilayer Perceptron neural network. The algorithm was used to predict the personality of Tweets from three datasets accessible in the literature, and it produced an approximately 83 percent accurate prediction, with some personality qualities performing better than others in terms of individual categorization rates. Gjurkovic [19] launched a large-scale dataset marked with MBTI kinds, obtained and analyzed a wealthy collection of characteristics from this dataset, trained and assessed benchmark models for predicting personality. Three distinct classifiers supporting vector machine (SVM),2-regulated logistic regression (LR), and a multilayer perceptron of three layers were used.

Pre-trained models were employed by Acheampong FA [8] to handle various NLP issues such as text-based emoticon categorization and poisonous comment classification [9]. When RoBERTa and XLNet are combined with additional NLP features such as TF-IDF and sentiment analysis, accuracy improves. Deep learning has been widely used in recent technologies to increase performance in predicting a person's personality. As in the experiment conducted by [14, 15, 16, 17] utilizing a different dataset, namely personality Café, where there are differences in personality modelling, the Myers Briggs Type Indicator (MBTI) technique is used. The use of LIWC is not without its own set of constraints. LIWC was created to investigate the language used by people while writing about traumatic situations. As a result, its applicability in any generic situation is dubious. Furthermore, the authors and their colleagues created the LIWC categories somewhat randomly. Other textual elements of a user's profile on a social network can be studied to extract some previously unknown traits and examine their effect on personality identification. A new set of traits can be extracted and evaluated in terms of their usefulness in determining a user's personality [25].

3. Methodology

3.1. Data Collection

The data set used was provided by a project that was developed in 2007 based on Face- book. This data set was under the project called myPersonality. This project was done by David Stillwell and Michael Koniski. It was basically a platform that allowed the users to take psychometric test, based on Big Five trait model. Depending on the answer chosen and proportionality of the question with the trait score was given. Around 6 million of the users participated amongst which 40% were willing to share their Facebook data. The data set had various columns like Author Id, Status, Categories of the personality traits whether the user belongs to that trait, Scores of personality trait ranging from 1 to 5. It had some other columns like connectivity, betweenness etc. Another data set that was used is collected from twitter. This data set is extracted and is unlabeled. This is used for testing purpose. We here used the Transfer learning approach where we trained the model on different problem and used it to solve different but related problem. Twitter is microblogging site and that is popular, here what the user share are called as tweets. Twitter allows users to mine or extract the tweets for any users along with other information. To collect the Twitter data for use, Tweepy was used. It is an open-source library in python that gives access to Twitter API. Tweets were collected for the user and given to the model created. Because this is a regression problem, we only used columns with continuous values for personality traits. Figure 2 depicts the various statistics performed on the Five Factor Model. The minimum, maximum, and average values were discovered in order to gain insights into the data.

1.65 Min of AGR	3.62 Average of AGR	5.00 Max of AGR
1.45 Min of CON	3.47 Average of CON	5.00 Max of CON
1.33 Min of EXT	3.35 Average of EXT	5.00 Max of EXT
1.25 Min of NEU	2.61 Average of NEU	4.75 Max of NEU
2.25 Min of OPN	4.13 Average of OPN	5.00 Max of OPN

Fig. 2. Statistics for Traits in Dataset.

3.2. Workflow

The figure 3 shows the flow of the personality prediction system implemented. Dataset used is myPersonality, as discussed in data collection after removing the unnecessary column like betweenness, connectivity etc. Also, some preprocessing steps as discussed in further section was done. To give input to convolution layer, first we need to convert it into vector or numerical form so that the DL could process it. This is done with the help of word embedding algorithm. The CNN will then have the input with a convolutional layer with many kernel sizes, bundling layers that extract just the key aspects of the text. We have trained 5 different models on the same architecture, one for each Big Five trait to get better prediction of personality as the dataset is less in amount. We have retrieved the Twitter data as we stated before and we use the same preprocessing and embedding techniques on this extracted dataset as on the trained model. This will give the output for one of the traits. All this will be elaborated in further section.

3.3. Pre-Processing

Preprocessing is one of the important tasks in natural language processing, it transforms text into predictable form. There are various sub steps in preprocessing. One of the basic steps in implementing any machine learning model is to clean the data. Columns that are no longer needed in implementation, such as Betweenness, are eliminated. Using NLTK library various cleaning steps were done on the Status column in the dataset like removing punctuation marks, usernames, converting to lowercase, lemmatization, stop word removal, tokenization, remove URLs and html tags, replacing the Slang words using slang dictionary. Next stage is a numerical or vector conversion.

3.4. Embedding

Human language can be easily understood by the humans. But to make machine learning models to understand the language we need to convert it into numerical values. Text embedding is the mathematical way of representing the text data. Analysis is done on the text and represent each word or entire sentence into a very dense or high-dimensional space. Word's embeddings are the model that maps the words in human language in form of vectors. It improves the ability of the system to understand natural language and thus improve the ability to learn from it. For Word embedding, GloVe a popular way of Word embedding is used. GloVe stands for Global Vectors for Word representations [4] developed by Pennington et.al in 2014 at Stanford. Unlike word2vec it allows to incorporate word statistics in global, that is word-occurrence to get the vectors along with the local statistics related to that document or context. GloVe is beneficial to get the semantic relation between the matrix that are cooccurring. Cooccurring matrix gives the details about how the words occur together, each value represents how the corresponding row and column are occurring together. n-dimensional vector is created for every word. For a word as W mathematical equation will be

$$W = g(w),\ if\ w\ \in Gl \qquad 0,\ w\ \notin \mathrm{Gl} \tag{1}$$

Here Gl means the GloVe Dictionary and the g(w) means the vector in GloVe that it will return if the word is present. It will return zero if the word is not present that eventually helps to preserve the word sequence in the Status. Text document is initially indexed then embedding algorithm is applied and then the Embedding matrix is formed. Here each sentence is represented as dimensional vector.

$$S * E \tag{2}$$

Where S is the sentence and E is the Embedding size. Embedding size chosen for the experiment was 50 and 100.If a word is present in the GloVe document, the corresponding size of embedding vector is formed. The output of word embedding is given to the convolution layer.

3.5. Convolution Neural Network

After embedding is done next step is to apply Convolution neural network. The figure 4 shows detail system architecture with various layers in the model. The overall architecture can be explained in Four blocks from input to output from Input to Output. First Block is to convert the input text to the corresponding vector form as explained in above section. Next Block is Convolution layer and third in Pooling Layer that does the task of feature mapping and extraction of important features. Last block consists of Dense layers, these will give the output for Regression that is one value that is in Big Five Personality trait. Initially we have Data set that is processed and then these words need to be converted to the form that can be readable by model using GloVe, it has various embedding or dimensionality options like 50,100,200 and 300.

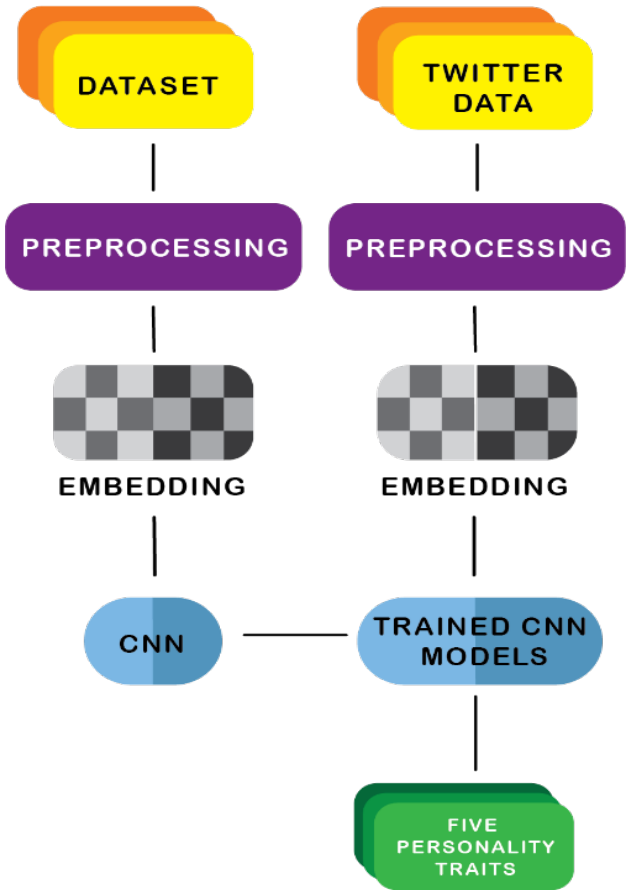

Fig. 3. Workflow for Personality Prediction.

This means the researcher can decide on the embedding size to be considered. Here chosen dimension is 50 and 100. In the data set after preprocessing, we got maximum number of words is 64, but we considered 65. So, for a row in Status column, we will get 65*50 embedding matrix. This complete the embedding next layer is the convolution layer, which is

the first block in the CNN. In convolution, we have used filter kernels of height or sizes as 1, 3 and 5. This is done to create features map. Size chosen means the number of words is considered. Feature maps capture the result that is obtained after applying the kernels. 32 features maps are obtained after each of these sizes. After convolution, next is the Pooling layer. Pooling layer is added to reduce the size of our feature maps obtained in the convolution layer. This is also called as down-sampling. This layer can improve the efficiency of the algorithm and in turn improve the performance of the model. We have applied Max-Pooling to those takes largest value from the feature map. This will help extract the words that are important to the sentence and will discard all the less important words.

In the next part, all these features map after applying Max pooling, can be concatenated in one feature map. This is done to make the CNN to learn different representation or features map no matter what the height of Kernel is.

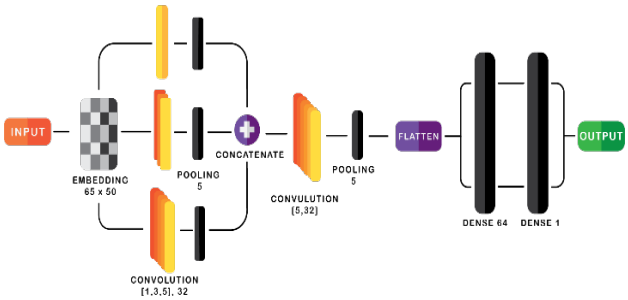

Fig 4. Architecture of Proposed System.

Then again convolution layer is applied, this step helps model to learn more and more complex features that could not be learned earlier. As this is regression problem, the activation function use Rectified Linear Unit (ReLu). After this pooling layer is applied again with max-pooling. Before giving these learned features maps as input to dense layers that is considered as the fully connected neural nets, these should be converted to 1-D array. This array is achieved by applying flatten layer to the output of pooling layer. The output of flatten layer is given as input to the dense layer that has 64 neurons. At the end, the output of the dense layer is given another dense layer that has 1 neuron that will give the output. One output means it represents one of the traits of Five Factor model. The above model is just for one trait, this is done for all the Five traits.

3.6. Transfer Learning to Predict Personality from Twitter

Figure 5 shows how the transfer learning is used. In this, the model was trained on myPersonality dataset, that has Facebook posts, we used this trained model on the twitter tweets extracted. User was asked to give any screen name of choice and then the tweets were extracted from that screen name. These files were stored in csv file and then preprocessed. Most of the preprocessing carried out was using either NLTK library or by writing the regular expressions. To give the model input it was required to convert it to vectors form, this was done by using the GloVe as word embedding algorithm described earlier. We created Five models one for each trait present in the Big Five Trait theory, these models were saved in a pickle file. This pickle file serializes the model in form of byte stream and when deserialized can be converted back to normal form. For every model it was done. The extracted data after embedding was passed to the pickle file with saved model that gave output for that particular trait. Thus, as output we got the traits for all tweets present in the embedding. As these tweets were extracted for the single person, we calculated the mean for each trait and that was the final of personality.

4. Experiments and Results

We trained and tested the model on the dataset after pre-processing and embedding it, with architecture as discussed in figure 5.2. The training was instantiated on two batch sizes as 64 and 128 for four different epochs ranging from 5 to 20. The experiment for both batch size and the epochs were conducted for 3 different spilt ratios for training and testing, they were as 70:30,80:20 and 90:10. Each of these spilt has 10% of validation data. Embedding is one of the most important for text-based CNN. Differing the embedding size results in two variance of the results. So, to find the accurate model two different embedding were chosen as 50d and 100d.

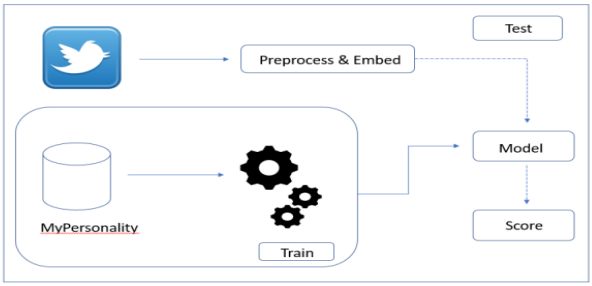

Fig 5. Predicting Personality from Twitter.

All the inclusive the experiments were various combinations based on three splits ratio each had two embedding dimensions comprised of two batch sizes each had 4 training epochs. Evaluating the RMSE and training for all experiments depends on the correct use of optimizer as discussed in 5.4.3. We used three optimizers namely SGD, Adam, RMSprop for all three combinations. To find the best model that gave the lowest accuracy, it was to decide on the optimizer. For this, as shown in table 1, we focused on 80-20 spilt of the data and focused on embedding size of 50, then checked for the RMSE values for 3 optimizers can been that for both of the batch size , RMSprop shows very high RMSE values and there were much fluctuations in the value, whereas the SGD has lower values but the Loss graph showed that it had much higher losses for the training data that could further lead to wrong predictions on unseen data. So, the Adam was chosen as the optimizer.

Table 1. Comparing Root Mean Square error for Optimizers.

			Regression - RMSE					
			Split-Ratio	80-20				
			Optimizers	SGD		ADAM		RMSprop	
				Validation	Testing	Validation	Testing	Validation	Testing
50	64	5		0.48	0.45	0.49	0.47	0.45	0.47
		10		0.49	0.44	0.5	0.49	0.53	0.53
		15		0.46	0.47	0.51	0.52	0.49	0.49
		20		0.45	0.44	0.56	0.5	0.52	0.52
	128	5		0.46	0.45	0.46	0.44	0.65	0.68
		10		0.45	0.45	0.46	0.48	0.57	0.46
		15		0.44	0.43	0.52	0.5	0.51	0.52
		20		0.46	0.45	0.54	0.53	0.5	0.53

After finalizing Adam as the optimizer next step was to find the best spilt ratio for the model, so considered Embedding size as 50 and Optimizer as Adam as shown in 2. We found out the mean value for all the five traits for all the spilt and it was seen that the 70:30 spilt had less RMSE values, but the loss of each trait was much higher as compared yo other two spilt. And training loss was less than validation as the data set had very less data, it could not generalize well for the unseen data.90:10 gave much higher peaks in loss that would result in wrong predictions further. Whereas it was seen that 80:20 had minimal loss and minimum RMSE value as well. So, the best ratio for spilt was 80:20.

Table 2. Comparing Root Mean Square Error for Spilt ratio.

Split-Ratio			70-30		80-20		90-10	
Optimizers			ADAM		ADAM		ADAM	
			Validation	Testing	Validation	Testing	Validation	Testing
Embedding Size	Batch Size	Epochs						
50	64	5	0.48	0.45	0.49	0.47	0.46	0.46
		10	0.5	0.5	0.5	0.49	0.49	0.51
		15	0.51	0.53	0.51	0.52	0.5	0.52
		20	0.54	0.54	0.56	0.5	0.52	0.56
	128	5	0.46	0.45	0.46	0.44	0.43	0.48
		10	0.47	0.48	0.46	0.48	0.49	0.47
		15	0.51	0.5	0.52	0.5	0.51	0.49
		20	0.49	0.52	0.54	0.53	0.58	0.51

Next step was to finalize with the Batch size as it is also one of the hyperparameter for the training the convolution neural network. This was gain done on Embedding size 50, Optimizer as Adam and spilt ratio as 80:20 as shown in Table 3 This was compared for all the traits with same combination of other values. And could be clearly seen that the RMSE values were minimal for 128 batch sizes rather than 64. And we need RMSE low as possible, so the choice was 128 batch sizes. After fixing for optimizer as Adam, Spilt ratio as 80:20, and Batch size as 128.We now need to decide for Embedding Size for the text and Number of epochs. So again, we considered the values for all the traits with Optimizer Adam, Spilt ratio as 80:10 and Batch size as 128 as in table 2. Increase in embedding size, slightly increased the RMSE values. Average of RMSE for all the models for 50 dimension was less than average of RMSE for 100 dimensions. In short, it was seen that there was slight difference in the embedding size 50d and 100d. This minute difference between them helped to choose 50 dimensions as the final size. Increasing the dimension increased the computation time and the results were not much impressive so 100d was discarded. For all the traits, on 50 as the embedding size, 80:20 as spilt ratio, Adam as optimizer, and batch size as 128 was compared. It was seen that epoch 5 gave lower RMSE values in average. For most of the traits it was best even. Increasing the epoch increased the computation time as well as it increased the loss, so they were discarded.

As discussed above the model chosen has optimizer as Adam, spilt as 80:20, batch size as 128, Embedding dimension as 50 and Epoch as 5. For this model, we trained myPersonality dataset for all the 5 traits and found the Loss graphs. Figure below show the Training and Validation loss for the all the models created for Big Five personality traits.

Table 3. Comparing RMSE for Embedding size and Finalizing Model.

Traits			Openess		Conscientiousness		Neuroticism		Agreeableness		EXtraversion	
Optimizers			ADAM		ADAM		ADAM		ADAM		ADAM	
			Validation	Testing	Validation	Testing	Validation	Testing	Validation	Testing	Validation	Testing
Embedding Size	Batch Size	Epochs										
50	128	5	0.46	0.44	0.57	0.59	0.62	0.61	0.57	0.56	0.68	0.7
		10	0.46	0.48	0.62	0.64	0.68	0.68	0.6	0.58	0.75	0.74
		15	0.52	0.5	0.68	0.67	0.69	0.69	0.63	0.6	0.77	0.79
		20	0.54	0.53	0.67	0.7	0.73	0.68	0.61	0.6	0.76	0.8
128	128	5	0.45	0.45	0.58	0.58	0.61	0.62	0.58	0.56	0.7	0.7
		10	0.47	0.48	0.63	0.64	0.66	0.66	0.56	0.58	0.76	0.75
		15	0.52	0.52	0.67	0.64	0.68	0.68	0.59	0.59	0.75	0.79
		20	0.52	0.53	0.7	0.68	0.69	0.68	0.6	0.59	0.78	0.77

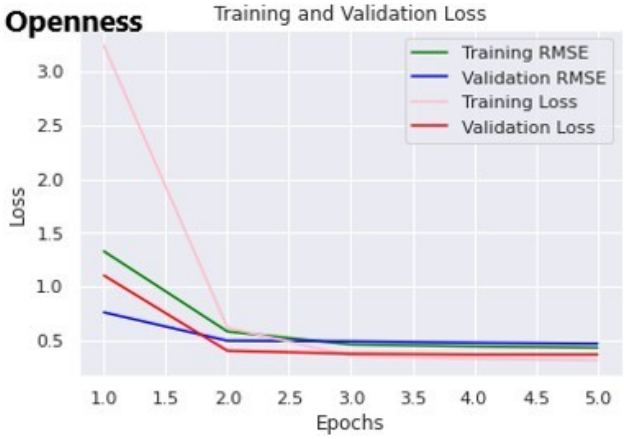

Fig 6. Training and Validation Loss for Openness.

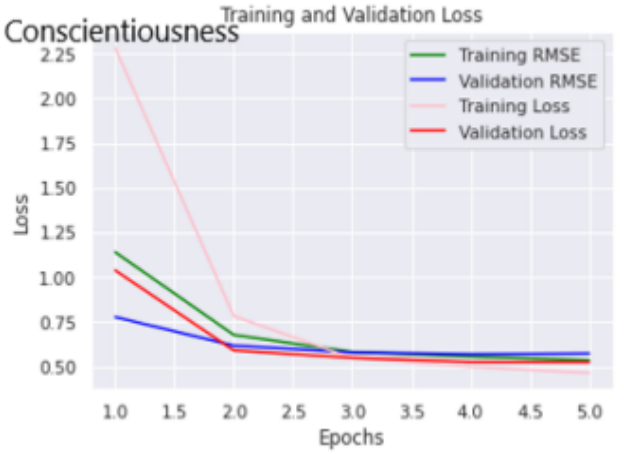

Fig 7. Training and Validation Loss for Conscientiousness.

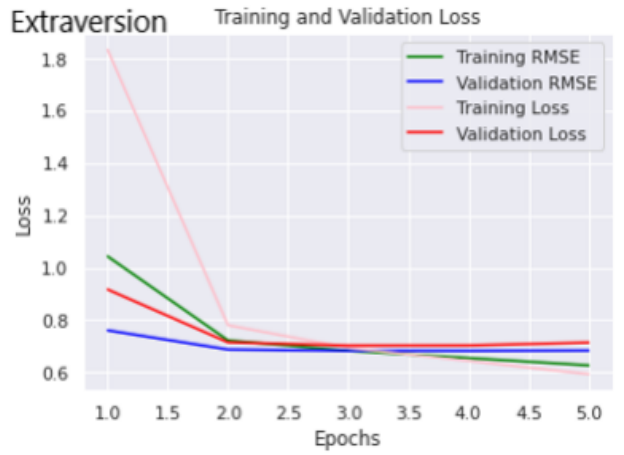

Fig 8. Training and Validation Loss for Extraversion.

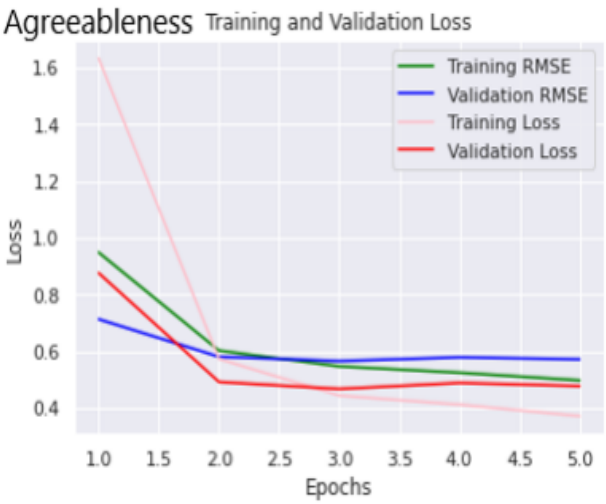

Fig 9. Training and Validation Loss for Agreeableness.

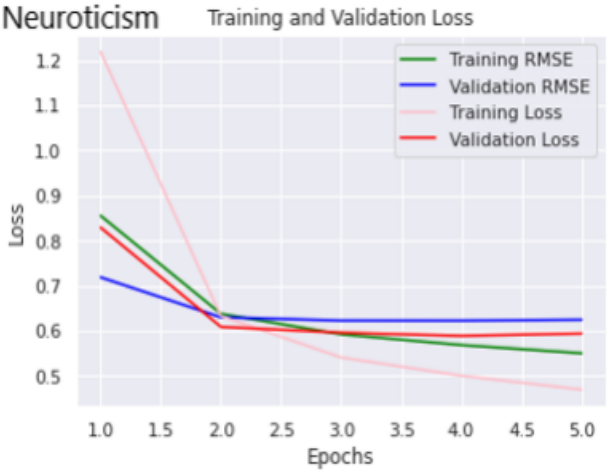

Fig 10. Training and Validation Loss for Neuroticism.

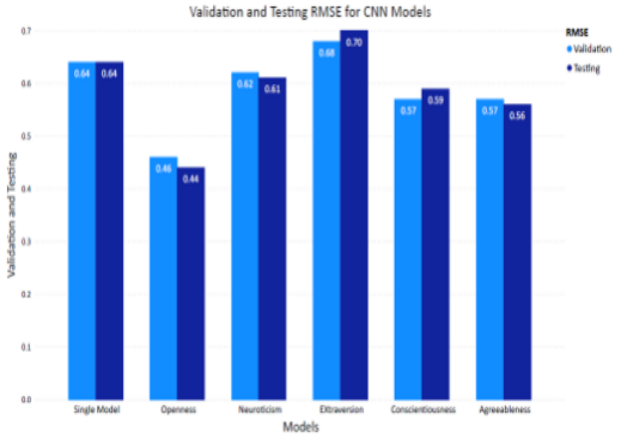

Fig 11. Training and Validation RMSE for Big Five CNN models.

5. Conclusion

Personality prediction can be done using deep learning, a subset of machine learning algorithms. In this paper, a popular Deep learning architecture called CNN is used to predict the personality. The proposed architecture can be effectively used to predict the values of the traits and can produce comparable results with other state-of-art. Training was done on myPersonality dataset, a regression problem was solved. Models were created one for each trait present in Big Five Trait model. We then applied already trained model on the twitter data that was extracted. In this twitter posts were considered separately and then predictions were made and as the result these were averaged as these were related to one person's tweets. As a part of future scope, we could incorporate images as well with the text data to make predictions more accurate. Apart from this using other embedding algorithms could also be considered. We also can make deeper networks.

References

1. Hayat, M.K., Daud, A., Alshdadi, A.A., Banjar, A., Abbasi, R.A., Bao, Y., and Dawood, H. (2019) Towards Deep Learning Prospects: Insights for Social Media Analytics. *IEEE Access*, **7**, 36958–36979, doi:10.1109/access.2019.2905101.
2. Mamta Bhamare, Dr K.Ashok Kumar, Personality Prediction from Social Networks text using Machine Learning", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019.
3. Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. 2018. Deep learning-based personality recognition from text posts of online social networks. Applied Intelligence 48, 11 (November 2018), 4232–4246. DOI:https://doi.org/10.1007/s10489-018-1212-4.
4. Tadesse, Michael M. et al. "Personality Predictions Based on User Behavior on the Facebook Social Media Platform." IEEE Access 6 (2018): 61959-61969.
5. D. Xue et al., "Personality Recognition on Social Media With Label Distribution Learning," in IEEE Access, vol. 5, pp. 13478-13488, 2017, doi: 10.1109/ACCESS.2017.2719018.
6. N. Majumder, S. Poria, A. Gelbukh and E. Cambria, "Deep Learning-Based Document Modeling for Personality Detection from Text," in IEEE Intelligent Systems, vol. 32, no. 2, pp. 74-79, Mar.-Apr. 2017, doi: 10.1109/MIS.2017.23.
7. Bachrach Y, Kosinski M, Graepel T, Kohli P, Stillwell D. Personality and patterns of Facebook usage. In 4th Annual ACM Web Science Conference; 2012. p. 24-32
8. Acheampong FA, Nunoo-Mensah H, Chen W. Transformer models for text-based emotion detection: a review of BERT-based approaches. Artif Intell Rev. 2021. https://doi.org/10.1007/s10462-021-09958-2.
9. Maslej-kreš V, Sarnovský M, Butka P. Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. Appl Sci. 2020. https://doi.org/10.3390/app10238631.
10. Howlader P, Pal KK, Cuzzocrea A, Kumar SDM. Predicting facebook-users' personality based on status and linguistic features via flexible regression analysis techniques. Proc ACM Symposium Appl Comput. 2018. https://doi.org/10.1145/3167132.3167166.
11. Tandera T, Hendro S, D., Wongso, R., & Prasetio, Y. L. . Personality prediction system from facebook users. Procedia Comp Sci. 2017;116:604–11. https://doi.org/10.1016/j.procs.2017.10.016.
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in Neural Information Processing Systems, 2017-Decem(Nips), 2017. pp. 5999–6009.

13. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: generalized autoregressive pretraining for language understanding. ArXiv, NeurIPS; 2019. pp. 1–18.
14. Cui B (n.d.). Survey analysis of machine learning methods for natural language processing for MBTI Personality Type Prediction. http://cs229.stanford.edu/proj2017/final-reports/5242471.pdf.
15. Hernandez and Knight. (n.d.). Predicting MBTI from text.
16. Keh SS, Cheng I-T. Myers-Briggs personality classification and personality-specific language generation using pre-trained language models. July. 2019. http://arxiv.org/abs/1907.06333.
17. Yuan C, Wu J, Li H, Wang L. Personality recognition based on user generated content. 2018 15th International Conference on Service Systems and Service Management, ICSSSM 2018; 2018. pp. 1–6. https://doi.org/10.1109/ICSSSM.2018.8465006.
18. Lima, Ana Carolina & De Castro, Leandro. (2014). A multi-label, semi-supervised classification approach applied to personality prediction in social media. Neural Networks. 58. 10.1016/j.neunet.2014.05.020.
19. Gjurkovic, Matej and Jan Snajder. “Reddit: A Gold Mine for Personality Prediction.” PEOPLES@NAACL-HTL (2018).
20. Funder, David & Fast, Lisa. (2010). Personality in Social Psychology. 10.1002/9780470561119.socpsy001018.
21. S. Adal and J. Golbeck. 2012. Predicting personality with social behavior. In Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.
22. R. Wald, T. Khoshgoftaar, and C. Sumner. 2012. Machine prediction of personality from facebook profiles. In Proceedingsof the 2012 IEEE 13th International Conference on Information Reuse and Integration (IRI‟12). 109–115.DOI:10.1109/IRI.2012.6302998
23. G. Farnadi, S. Zoghbi, M. F. Moens, and M. De Cock. 2013. Recognizing personality traits using facebook status updates. In Proceedings of Workshop on Computational Personality Recognition (WCPR‟13).
24. Kaushal, Vishal & Patwardhan, Manasi. (2018). Emerging Trends in Personality Identification Using Online Social Networks—A Literature Survey. ACM Transactions on Knowledge Discovery from Data. 12. 1-30. 10.1145/3070645.
25. Alam F, Stepanov EA, Riccardi G. Personality traits recognition on social network—Facebook. AAAI Workshop—Technical Report, WS-13-01, 2013. pp 6–9.
26. Mairesse F, Walker M, Mehl M, Moore R. Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of artificial intelligence research. 2007; 30: p. 457-500. 3
27. Fast L, Funder D. Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. Journal of personality and social psychology. 2008; 94(2): p. 334.
28. https://courses.lumenlearning.com/intropsych/chapter/what-is-personali ty/