2022 7th International Conference on Data Science and Machine Learning Applications (CDMA) | 978-1-6654-1014-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDMA54072.2022.00042

Students Personality Assessment using Deep Learning from University Admission Statement of Purpose

Salma Kulsoom
School of Electrical Engineering and Computer Science
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
email: skulsoom.mscs17seecs@seecs.edu.pk

Seemab Latif
School of Electrical Engineering and Computer Science
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
email: seemab.latif@seecs.edu.pk

Tanzila Saba
College of Computer and Information Sciences
Prince Sultan University
Riyadh, 11586Saudi Arabia
Email: tsaba@psu.edu.sa

Rabia Latif
College of Computer and Information Sciences
Prince Sultan University
Riyadh, 11586Saudi Arabia
Email: rlatif@psu.edu.sa

***Abstract*—Statement of Purpose (SOP) plays a vital role in the university admissions process as reviewers assess the personality of the students by reading their SOPs. In past, the Big Five personality traits of the students are assessed to predict their future academic performance. An exciting application of machine learning is the personality assessment using personality traits and behavior. In this paper, our focus is on developing a deep learning-based personality assessment model for the detection of Big Five Personality traits from SOP and mapping them to speculate a student's academic performance at the university. Our proposed model uses Long-Short Term Memory (LSTM), Convolutional Neural Network (CNN) and Bi-Directional LSTM (Bi- LSTM) architectures to extract features and predict ratios of Big Five traits in the SOP. The proposed model has been trained and tested on an essays' dataset and 400 students' SOP collected from computer science undergraduate students. Maximum accuracy achieved for essays dataset is 88.2% and for student's personal statement is 67.0 % with FastText Embedding.**

Keywords- Big Five Personality Traits, Long-Short Term Memory, Convolutional Neural Network, Deep Learning, Statement of Purpose (key words)

I. INTRODUCTION

Personality is a combination of emotions, behavior, motivation, and thought patterns possessed by an individual. Personality has a great impact on success. There is a strong support in literature for hypothesis that personality traits are predictor of individual's job performance [1-3] and students' academic performance and motivation [4]. Student's academic performance is a key factor to determine for educational institutes and universities. Besides good academic background and motivation, certain personality traits are required by the universities to ensure that the candidates to whom they are offering admissions are capable to complete their education and will not quit. For this reason, they get a written statement from each candidate at time of taking admission known as "Statement of Purpose SOP" or "Personnel Statement". SOP is an important aspect of admission procedure that tells about the candidate, why he/she is applying for the program, why he/she is a good candidate and what are the future plans. Due to thousands of applicants, it becomes impossible for application reviewers to read each SOP and analyze it. Therefore, a mechanism is required to automatically assess the personality traits from SOPs.

Automatic detection and assessment of personality traits from text has lot of practical applications. It can be very useful for students' career counseling, offering admissions in subjects according to personality type, selecting suitable candidates for job, recommending products to customers according to their personality type, understanding reviews on basis of type of personality etc. Personality is a very broad subject. In past, models have been developed to categorize personality into different types. The most widely used model is known as "Big Five Model" [5-6]. The five personality traits in Big Five model are:

- **Agreeableness (AGR):** How people react to others opinion? People high in this trait have morality, self-sacrifice, cooperation, modesty, and sympathy.
- **Conscientiousness (CON):** How disciplined, punctual, and hard working the person is? People high in this trait are responsible, reliable, dutiful and goal oriented.
- **Extraversion (EXT):** How talkative, energetic, social, and fun liking the person is? People high in this trait are very social, talkative, and easily stimulated by their external environment.
- **Neuroticism (NEU):** How nervous and sensitive person is? People high on this trait are depressed, angered, anxious and emotionally unstable.
- **Openness (OPN):** How inventive, creative, and curious a person is? People high on this trait are very creative, adventures and intelligent.

Computational Personality is a research field that includes artificial intelligence and personality psychology. In this field, personality is studied and analyzed using computational intelligence techniques from text, social media, and other multimedia contents. Prior scientific work in field of computational personality is based on bag of words, syntactic and semantic representation of language, Open vocabulary developed on basis of social media data [7], rich psycho-linguistic lexicon like Linguistic Inquiry and

Word Count lexicon (LIWC) [8-10]. Major drawback of lexicon-based approaches is that they require rich feature engineering and domain knowledge because such linguistic features are highly dependent on domain The relationship between these linguistic features and Big Five personality traits is not stable over different context of writing. Furthermore, LIWC categories are conceptually too broad due to which they tend to generalize [10].

Personality Traits are not only dependent on choice of words, but they are embedded deep in semantics and structure of the language. Therefore, extraction of big five personality traits require a deep neural network model which can automatically extract specific features related to each trait and can preserve structure and semantics of text over long range. Purpose of this research is to develop a lexicon and context independent deep neural network model to extract big five personality traits by preserving structure and semantics of language, and to develop a model which can predict probability of each trait in each text rather than classifying it into one trait as binary classification.

Rest of the paper is structured as follows: section 2 presents previous approaches used for computational personality recognition and their limitations. Section 3 presents our proposed methodology and dataset used for training. Section 4 provides comparison and analysis of results and Section 5 presents conclusion.

II. LITERATURE REVIEW

In literature, research has been carried out to identify relationship between big five traits and writing styles and auto detection of big five traits from text. Some very earlier work in this domain was initiated by Pennebaker and King [11]. They carried out deep research on building and identifying linguistic features correlated to big five personality traits. They compiled essays dataset written in stream of conscientiousness mode by 2000 students. They developed their own Psycho-linguistic lexicon known as Word Count and Linguistic Inquiry (LIWC) for identifying correlation between big five personality traits and language. Although they did not propose any model for automatic detection of traits from text, but they identified several important linguistic cues related to personality traits and proved that writing style is an important individual difference.

Bergold and Steinmayr reported that high consciousness score refers to strong association between academic performance and intelligence [12]. They used Neuroticism Extraversion Openness Five Factor Inventory (NEO FFI) for personality assessment and GPA was used for academic performance assessment. Hierarchical regression and moderation analyses were used for identifying correlation between traits and performance. They concluded that neuroticism is negatively associated with academic performance in high school and college students [12].

Novikova and Vorobyeva carried out research to identify the correlation between big five personality traits and academic performance of the students [13]. They concluded that consciousness is strongly and positively associated with high academic performance. Openness scored high in subjects requiring creativity, neuroticism has a slightly negative impact while agreeableness did not show any correlation.

A. *Relationship between Big Five Personality Traits and Lingusitic Styles*

Literature shows correlation between LIWC and Medical Research Council (MRC) features with big five personality traits. All LIWC and MRC features are not useful in extracting personality traits. There are few important linguistic markers for each trait that helps in identification of that particular trait from text. Table I presents summary of important linguistic markers for each personality trait.

TABLE I. RELATIONSHIP BETWEEN BIG FIVE PERSONALITY TRAITS AND LINGUISTIC STYLE

Personality Traits	*Linguistic Markers*	*Feature Set*
Extravert	Words related to Anger, affect, swearing, positive and negative emotions [11].	LIWC
Neuroticism	Self-references, Frequent use of words related to negative emotions and anxiety, words associated with negative affect [11], [13].	LIWC
Agreeableness	Frequent use of tentative words like perhaps, maybe etc. Use longer words and short sentences, while disagreeable people use words related to swearing, negation and anger [11], [13].	LIWC and MRC
Conscientiousness	Words associated with insight, longer words, words related to negative, words with feelings features appraisal and affect and fewer appreciation appraisal words [11], [13]. They concentrate more on their own feelings whereas, unconscientiousness candidates use words related to swearing, anger and negative emotions.	LIWC, MRC and Appraisal Features
Openness to experience	Frequent use of articles and long words. Non open candi- dates tend to talk about themselves and their occupation and use second person pronoun [11], [13].	LIWC and MRC

B. *Analyzing Realtionship between Personality Traits and Linguistic Styles using Machine and Deep Learning*

Mairesse et al. performed detailed analysis on relationship between linguistic styles and personality traits both in text and conversation [8]. They used their own features like average length of word in text, age of acquisition of words, image ability of words along with LIWC features. They applied Naïve Bayes, Ada boost, SVM and ranking models with combination of different features set. Maximum accuracy achieved with SVM for Neuroticism (Emotional stability) was 57.35%. They also concluded that personality traits are not mutually exclusive and traits classification is not a binary classification problem.

Liu et al. developed language independent compositional model for extraction of big five traits from short tweets [14]. They used Bi-Directional Gated Recurrent Units (GRU) to construct words from character N-grams, which were fed into another GRU to construct sentences from word N-grams. Output of this network was feed into multi-layer perceptron to predict personality traits. Their dataset consisted of 14,166 tweets from 52 users. Their model works only for short texts and was a binary classification model.

Majumder et al. presented a novel deep neural network-based model for extraction of big five personality traits from text [15]. Input to model was Word2Vec representation of words. They applied convolutional layer to construct sentences from word N-grams and another convolutional layer to construct documents from sentences. This representation of document was then combined with LIWC and MRC features and fed into two-layer perceptron with softmax classification function. They trained five different classifiers all with same under lying architecture for each trait. Maximum accuracy achieved through this model was 62.68% for openness. Early work on computational personality recognition used lexical and grammatical approaches like LIWC, MRC, POS tags with SVM and Naïve Bayes classifiers. All models presented so forth in literature are binary classifiers or One vs. Rest Classifiers but in realty big five traits are not mutually exclusive they are overlapping, and people possess these traits to a certain degree. Deep Neural Network based approaches have been recently used for extraction of big five traits from text.

Personality traits recognition is multi label probabilistic text classification problem and traits are buried deep in structure and semantics of words. Simple look up table or bag of words is not a good representation for this purpose. It requires more sophisticated technique which can capture semantic similarity of words according to context. Common practice to obtain such representation is to use word embedding techniques like Word2Vec, GloVe and FastText. Word Embedding techniques can capture semantic and syntactic similarity of words in context of document. Furthermore, extraction of big five personality traits require a deep neural network model which can preserve structure and semantics of words both within and across the sentence.

Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) based architectures have been successfully applied in many Natural Language Processing (NLP) tasks to capture both local and global semantics across long sentences and have shown much better accuracies than SVM and Naïve Bayes [16-18]. CNN can extract local dependencies among neighboring words in a sentence and they can extract features in parallel. N-grams features can be easily extracted from different regions of text using filter regions of different sizes. But CNN cannot preserve relationship among words over long sentences.

LSTM can learn long range dependencies and temporal relationship among words even when the sentence is too large. LSTM is a variation of Recurrent Neural Network (RNN) with an extra memory cell yield for the self-looping associations and can recall inputs about 1000 times steps away. The Bidirectional LSTM (BiLSTM) is an enhancement for the LSTM. BiLSTM scans input text from both forward and reverse direction; hence is more robust. It has shown better performance in many text classification problems [19-21].

CNN cannot capture long range dependencies among words, LSTM cannot extract features in parallel, but they can capture long range dependencies among words. Therefore, combining CNN and LSTM results in much better performance for text classification [21, 26]. By Feeding output of CNN into LSTM, both local and temporal relationship between words over very long text can be retained. Combined CNN-LSTM architecture have been used in many text classification problems like sentiment analysis, reviews classification, named entity recognition, PoS tagging and have outperformed shallow machine learning classifiers [16-18], [22].

III. Proposed Proposed CNN-LSTM Approach

Our proposed model incorporates different variations of LSTM and CNN models for improving accuracy of the results over the given problem. Fig I show the proposed architecture. We have applied basic pre-processing steps on the dataset using standard NLTK library. Pre-processing includes splitting of text into tokens, removing punctuation marks and numbers and stemming words. Stop words were not removed from text because use of articles and stop words are also an important indicators of specific personality type [9], [23].

Raw text cannot be fed directly into deep neural network model. It needs to be represented in vector form. Traditional word representation techniques are based on bag of words representation or sparse vector representation where words are treated as independent units. Sparse representation of words cannot incorporate relationship and semantic similarity between words due to which a lot of rich information is lost. On the other hand, word embedding techniques can capture rich syntactic and semantic similarity of the words. It is also able to incorporate relationship between words according to the context. As personality traits are hidden in syntax, semantics, and composition of the words [17] therefore, we used word embedding techniques for representing our text data rather than shallow bag of words model. There are three different types of word embeddings used in literature Word2Vec, FastText and GloVe. We have used pretrained FastText and GloVe embeddings in our approach.

A. LSTM Model

We have applied LSTM model with variation of different layers.

- LSTM with Dropout Layer - Dropout layer is used for regularization and to avoid overfitting.

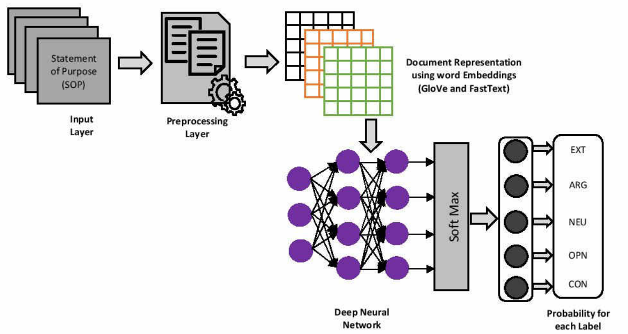

Figure 1. Proposed Model

- LSTM with Max pooling – to generate single feature map and retain most important features of the text as shown in Fig II.
- LSTM with CNN - to capture both local and long-range dependencies among words, as shown in Fig III.

Hyper-parameters used for all the LSTM variations are given in table II.

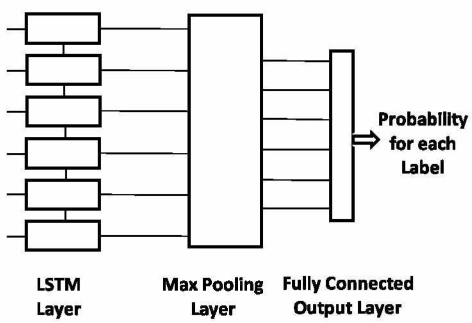

Figure 2. LSTM with Max Pooling layer

B. CNN Model

We have applied one dimensional CNN with filter regions of different sizes. Hyper-parameters of the CNN model are given in table II. We implemented two CNN model with above mentioned specifications and varying filter region of size 3 and 5. CNN alone does not work well therefore, we combined one dimensional CNN with LSTM and Bi-LSTM which resulted in much better performance.

C. Bi-LSTM Model

We have applied Bi-LSTM Model with variations of different layers. Hyper-parameters of the Bi-LSTM model are given in table II. Bi-LSTM scans input text both in forward and reverse direction and takes both left and right context of a word into account while predicting target word. Thus, it can preserve rich syntactic and semantic meaning of words that are very important indicator of personality traits.

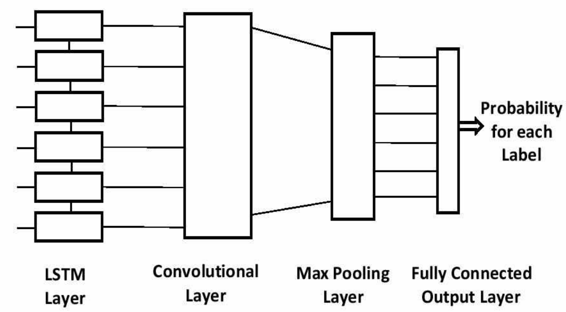

Figure 3. Combined LSTM-CNN model

To avoid over fitting of the model, we have added a dropout layer for regularization. Then we appended Max Pooling layer on top of Bi-LSTM which resulted in improved accuracy, Fig IV. Applying Max Pooling layer on top of Bi-LSTM reduced the variation and helped in improving generalization.

TABLE II. HYPER-PARAMETERS FOR LSTM, CNN AND BI-LSTM MODELS

Hyper-parameter	*Value*
Activation function inner layers	Relu
Classification function at output layer	Sigmoid
Loss function	Binary cross Entropy
Learning rate	0.001
Batch Size	16 (LSTM & Bi-LSTM) 32 (CNN Model)
Epochs	20 (LSTM & Bi-LSTM) 50 (CNN Model)
Embedding	glove.6B.300sd GloVe and crawl 300d−2M.vec FastText

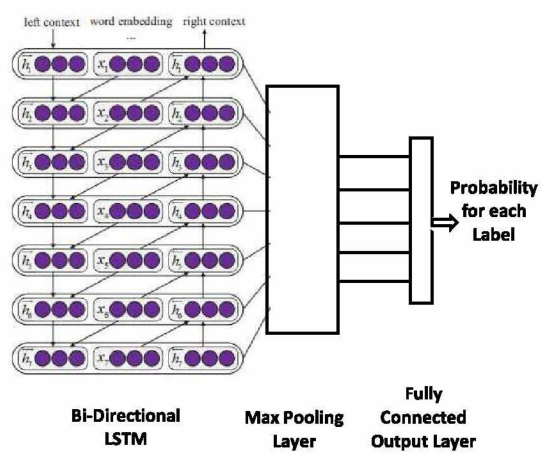

Figure 4. Bi-LSTM with Max Pooling Layer

As LSTM layer on top of one-dimensional CNN has been widely used for many NLP applications, we tried this architecture with Bi-LSTM. We have appended Bi-LSTM layer on top of one-dimensional CNN, but it resulted in reduced accuracy. Then we appended one-dimensional convolutional layer with Global Max pooling layer on top of Bi-LSTM layer which resulted in better, as given in Fig V.

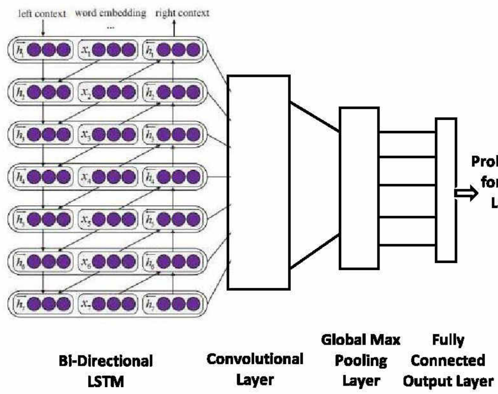

Figure 5. Bi-LSTM with 1D CNN and Global Max Pooling Layer

D. *Exploratory Analysis of the Dataset*

To develop a personality detection and assessment model from students' personal statements, we required dataset of personal statements. For this purpose, we have collected 400 personal statements from undergraduate students of first, third and fourth year studying at the university in Pakistan. Students were asked to write personal statements consisting of minimum 500 words in English language. Their corresponding personality traits were assessed through self-assessment tests available online. Two different types of tests were used for assessment to get more reliable results. To get true personality traits, students were asked to complete these tests twice: Once at beginning of semester and once during the semester. Details of test are as follows:

- Truiy Online Personality Test based on questionnaire provided by [24-25].
- Five Factor Personality Test is based on International Personality Item Pool representation of the Five Factor Model of personality. It was developed by the University of Westminster under their research project to be used in online psychological research. Test is consisted of 41 questions.

After collection of personal statements and big five personality scores, personal statements were annotated with their corresponding personality labels. We have collected 400 personal statements and their corresponding online tests. As our collected dataset is not sufficient for deep learning model training therefore, we used Essays dataset provided by Pennebeker and King [11]. This dataset contains 2000 essays written in stream of conscientious mode and annotated with personality trait labels of big five model. We can see almost uniform distribution of all five traits in the dataset therefore, we did not normalize data for equal distribution.

IV. Results and Analysis

Maximum accuracy has been achieved with Bi-LSTM-max pooling model. Maximum accuracy for essays dataset is 88.2% and for students' personal statement dataset is 67.0 % with FastText Embedding. Bi-Directional LSTM preserves text structure from both forward and backward direction and hence it can capture rich structural features of long text documents which are key indicators of big five personality traits.

TABLE III. ACCURACY OF MODELS

Model	*GolVe Embedding*		*FastText Embedding*	
	Essays Dataset	***Personal Statements***	***Essays Dataset***	***Personal Statements***
LSTM Max Pooling	71.3%	53.7%	72.8%	54.0%
LSTM-CNN	72.6%	57.9%	75.4%	61.3%
1D CNN with Kernel Size =5	70.5%	51.8 %	73.7%	54.3%
1D CNN with Kernel Size =3	75.0%	52.1 %	78.2%	54.8 %
Bidirectional LSTM with 1D CNN and global Max pooling	82.1%.	62.0%	**84.6%**	**69.8%**
Bidirectional LSTM with Max pooling layers	86.6%	65.7%	**88.2%**	**67.0%**

Performance of combined Bi-LSTM with 1D CNN and global max pooling model is comparable to Bi-directional LSTM. Accuracy with FastText Embedding for essays dataset is 84.6% and for students' personal statement dataset is 69.8%. To best of our knowledge, this is maximum accuracy achieved so forth for Big Five Personality traits from text. Model also performs efficiently when tested on corpus of personal statements collected for this research. Hence, it can be utilized successfully to predict proportion of specific personality traits from personal statements of students.

Proposed approach presents multiple probabilistic deep recurrent neural models for extracting personality traits are not dependent on any psycholinguistic lexicon. These models can extract rich features buried in structure and composition of the language. Such features are more significant than LIWC features explored in previous research. Proposed model predicts probability for each one of the five traits to which a document belongs to rather than classifying in binary manner as personality traits are not mutually exclusive. An individual to a certain degree possesses more than one personality traits. All models presented so forth in literature are binary classifiers. To best of our knowledge, this proposed model is first in domain of personality traits extraction which does not rely on any psycholinguistic lexicon. It does not require any sort of hand-crafted features for identification of personality traits and focuses on structure of language rather than selection of words. This model is domain independent and has been tested against real time data of SOP showing promising results.

V. CONCLUSION

Personality traits have deep impact on student's academic performance besides good academic background. Universities asks for Statement of Purpose from students to assess their personality and predict their success in the academic program. Reading statements from thousands of applicants is neither feasible nor unbiasedly quantifiable. Proposed research presents recurrent deep neural network model for automatic extraction of personality traits from SOP. Traditional approaches for extraction of personality traits from text are based on use of hand-crafted features and bag of words with shallow machine learning classifiers. Our proposed approach focuses on preserving structural, compositional, and semantic features of language. This model is based on variations of recurrent neural networks LSTM and CNN which can preserve rich syntactic and semantic features. Proposed methodology shows state of the art performance by achieving 88.2% accuracy which is maximum attained accuracy so forth for extraction of Big-Five personality traits from plain text documents and personal statements. Primary contribution by this research is the proposed model being independent from prerequisites of feature engineering, lexicon, and vocabulary. This domain independent model has been tested on personal statements written in real life by students and has shown 67% accuracy.

ACKNOWLEDGMENT

This work was supported by the Artificial Intelligence and Data Analytics (AIDA) Lab CCIS, Prince Sultan University, Riyadh, Saudi Arabia. Authors are thankful for the support.

REFERENCES

[1] D. A Bonebright, 40 years of storming: A historical review of tuckman's model of small group development. Human Resource Development International, 2010, 13(1), p. 111–120.

[2] R.R. McCrae, Costa Jr., and Paul T. The five-factor theory of personality, 2008.

[3] Gönül Kaya Özbag˘. The role of personality in leadership: five factor personality traits and ethical leadership. Procedia-Social and Behavioral Sciences, 2016, p. 235–242.

[4] S. Hakimi, E. Hejazi, and M.G. Lavasani. The relationships between personality traits and students' academic achievement. Procedia-Social and Behavioral Sciences, 2011, 836–845.

[5] A. E. Poropat. Ameta, Analysis of the five-factor model of personality and academic performance. Psychological Bulletin, 2009, 135(2), p. 322–338.

[6] E.C. Tupes and R.E. Christal. Recurrent personality factors based on trait ratings. Journal of personality, 1992, 60(2), p. 225–251.

[7] H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, L. Dziurzynski, S.M. Ramones, and et al. Agrawal, M. Personality, gender, and age in the language of social media: The open-vocabulary approach. PLoS ONE, 2013, 8(9).

[8] F. Mairesse, M.A. Walker, M.R. Mehl, and R.K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research, 2007, 30(1), p. 457–500.

[9] J. Oberlander and S. Nowson. Whose thumb is it anyway? classifying author personality from weblog text. In Proceedings of the COLING/ACL 2006, 2006, p. 627–634.

[10] J.W. Pennebaker, R.L. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[11] J.W. Pennebaker and L.A. King. Linguistic styles: Language use as an individual difference. Journal of personality and social psychology, 1999, 77(6), p. 1296–1312.

[12] S. Bergold and R. Steinmayr. Personality and intelligence interact in the prediction of academic achievement. Journal of Intelligence, 2018, 6(2), p. 27–45.

[13] I. A. Novikova and A. A. Vorobyeva. Big five factors and academic achievement in russian students. Psychology in Russia, 2017, 10(4), p. 93-106.

[14] F. Liu, J. Perez, and S. Nowson. A language-independent and compositional model for personality trait recognition from short texts. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, p. 754–764.

[15] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria. Deep learning-based document modeling for personality detection from text. IEEE Intelligent Systems, 2017, 32(2), p. 74–79.

[16] A. Hassan and A. Mahmood. Deep learning approach for sentiment analysis of short texts. In Proceedings of the 3rd international conference on control, automation and robotics (ICCAR), 2017, p. 705–710.

[17] N.Kalchbrenner,E.Grefenstette,andP.Blunsom.Aconvolutionalneural network for modelling sentences. 2014, p. 655–665.

[18] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. A C-LSTM neural network for text classification. arXiv preprint arXiv:1511.08630, 2015.

[19] F. Mohammad. Is preprocessing of text really worth your time for online comment classification? arXiv preprint arXiv:1806.02908v2, 2018.

[20] S. Srivastava, P. Khurana, and V. Tewari. Identifying aggression and toxicity in comments using capsule network. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, p. 98–105.

[21] Mehta, Y., Majumder, N., Gelbukh, A. *et al.* Recent trends in deep learning based personality detection. Artificial Intelligence Review, 2020, p. 2313–2339.

[22] X. Sun, B. Liu, J. Cao, J. Luo and X. Shen, "Who Am I? Personality Detection Based on Deep Learning for Texts," *2018 IEEE International Conference on Communications (ICC),* 2018, p. 1-6,

[23] F. Iacobelli, A.J. Gill, S. Nowson, and J. Oberlander. Large scale personality classification of bloggers. In Proceedings of the International Conference on Affective Computing and Intelligent Interaction, 2011, p. 568–577.

[24] J. W. Pennebaker, and L. A. King, Linguistic Styles: Language use as an individual differences, Journal of Personality and Social Psycology, 2000, 77(6), pp. 1296–1312.

[25] F. Mairesse, M. M. Walker, M. Mehl and R. Moore, Using Linguistic cues for automatic recognition of personality in conversation and Text, Journal of Artificial Intelligence Research, 2007, p. 457-500.

[26] S. Latif, S. Bashir, M. M. A Agha, R. Latif, Backward-Forward Sequence Generative Network for Multiple Lexical Constraints, In Proceedings of the 16th Artificial Intelligence Applications and Innovations AIAI, 2020, p. 39-50.