iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** VI **Month of publication:** June 2024

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue VI June 2024- Available at www.ijraset.com

Author Profiling Approach: Predicting Personality Traits on Twitter Data using Combined BERT and SimCSE Embeddings

Kottu Divya Jyothi[1], I Lakshmi Pradeepa[2], Karunakar Kavuri[3]
[1, 2, 3]*Department of Computer Science and Engineering, Swarnandhra College of Engineering and Technology, Narasapur, AP, India*

Abstract: ***Author profiling involves predicting various characteristics of an author from their writing style, including age, gender, native language, and personality traits. The PAN2015 shared task concentrated on author profiling within social media, challenging participants to predict the personality traits of Twitter users based on their tweets. In recent years, deep learning methods have risen to prominence in author profiling. Researchers frequently employ several notable models such as Word2Vec, doc2vec, GloVe, and FastText for generating word embeddings. These models have consistently shown effectiveness across various natural language processing tasks. For the PAN2015 task, participants employed a range of deep learning models to generate word embeddings, aiming to predict the age, gender, and personality traits of Twitter users. In this study, our focus was on enhancing the accuracy of personality traits classification using the PAN2015 dataset, a renowned benchmark corpus for author profiling. We employed pre-trained deep learning models, namely BERT and SimCSE, to generate word embeddings and sentence embeddings. For classification, we utilized Long Short-Term Memory (LSTM) and Convolution Neural Network (CNN) classifiers. Our findings revealed that the LSTM model, integrated with combined BERT and SimCSE embeddings, achieved an accuracy of 87.53% for personality traits classification, while the CNN model, similarly equipped, attained 80.48%. Additionally, utilizing BERT alone with LSTM yielded an accuracy of 78.45%, and with CNN, 75.32%. Our findings highlight the versatility of these models in addressing a range of natural language processing tasks, indicating their potential utility in diverse author profiling applications.***

Keywords: Author Profiling, Personality Traits Prediction, PAN2015, Word Embeddings, Sentence Embeddings, LSTM, CNN, BERT, SimCSE.

I. INTRODUCTION

The author profiling (AP) task aims to identify author demographics, such as age, gender, personality traits, or native language, through text analysis [1]. This research area has seen a significant surge in interest in recent years, contributing to various fields including marketing, security, terrorism prevention, and forensics. Machine learning approaches treating AP as a multi-class, single-label classification problem model it as assigning class labels (e.g., male, female) to texts. Input data for machine learning algorithms is often represented as fixed-length feature vectors, utilizing methods like bag-of-words or bag-of-n-grams [2].

From a deep learning perspective, author profiling involves training models to predict predefined author attributes based on their writing style. This typically employs neural network architectures such as MLP, convolution neural networks (CNNs), and recurrent neural networks (RNNs) to extract relevant features from text data [3]. For instance, in gender classification, a deep learning model learns from a dataset with known gender labels to recognize male or female writing styles. Similarly, in age classification, the model predicts author age based on writing style patterns associated with different age groups. Deep learning models excel in various profiling tasks including gender, age, language variety, personality traits, and sentiment analysis. Leveraging complex neural networks, these models discern subtle text patterns often overlooked by traditional machine learning, achieving state-of-the-art performance on benchmark datasets.

In this article, we utilized BERT and SimCSE pre-trained models to acquire document embeddings for the AP task. We applied LSTM and CNN classifiers to these embeddings and conducted experiments on a single genre of data. Dataset preparation involved splitting it into training and validation sets with an 80:20 ratio. Prior to training, text data underwent pre-processing steps including removal of stop words, stemming, lemmatization, and tokenization [4]. Subsequently, we employed pre-trained word embeddings to convert the text data into fixed-length vectors for input into our classification models. For personality traits classification; we trained a CNN classifier on the text data using pre-trained BERT and SimCSE embeddings.

Training involved utilizing the binary cross-entropy loss function and the Adam optimizer. Additionally, we trained an LSTM classifier on the text data with the same embeddings. LSTM, a recurrent neural network variant capable of capturing long-term dependencies in text data, was employed. Training this classifier utilized the categorical cross-entropy loss function and the Adam optimizer.

BERT and SimCSE are well-known techniques for producing word embeddings and sentence embeddings, which represent words as dense, low-dimensional vectors. While both methods are popular, they each have unique approaches and strengths. BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art natural language processing model developed by Google. It revolutionized the field by introducing a new approach to pre-training language representations. Unlike previous models that processed words in a left-to-right or right-to-left manner, BERT utilizes a bidirectional Transformer architecture, enabling it to consider the context of a word by looking at both preceding and succeeding words simultaneously. This bidirectional context understanding allows BERT to generate high-quality word embeddings that capture rich semantic information [5]. BERT's pre-training objectives, such as masked language modelling and next sentence prediction, enable it to learn deep contextualized representations of words, which can then be fine-tuned for various downstream tasks, such as text classification, named entity recognition, and sentiment analysis. As a result, BERT embeddings have become widely adopted and have demonstrated remarkable performance across a range of natural language processing applications. SimCSE (Similarity-based Contrastive Learning of Sentence Embeddings) is a recent advancement in the field of natural language processing that focuses on learning high-quality sentence embeddings. Unlike traditional methods that rely solely on supervised learning or self-supervised learning objectives, SimCSE introduces a novel similarity-based contrastive learning framework [6]. It aims to learn sentence embeddings by maximizing the similarity between augmented versions of the same sentence while minimizing the similarity between augmented versions of different sentences. By leveraging contrastive learning, SimCSE encourages the model to capture semantic similarities between sentences while emphasizing their differences, leading to more robust and informative embeddings. These embeddings have shown promising results in various downstream tasks such as semantic textual similarity, paraphrase detection, and text classification. Moreover, SimCSE's simple yet effective training approach makes it computationally efficient and easily applicable to a wide range of NLP tasks, making it a valuable addition to the toolkit of researchers and practitioners in the field. Moreover, one of them excels in capturing semantic relationships between words and another excels in capturing semantic relationship between sentences. This article is divided into seven sections. Section 2 provides an overview of existing approaches to author profiling. Section 3 describes the characteristics of the dataset utilized. In Section 4, we delve into the BERT word embedding, SimCSE sentence embedding methods, as well as the LSTM and CNN classifiers employed in this study. Section 5 elaborates on the proposed method. The experimental results are presented in Section 6. Finally, Section 7 concludes the work and discusses potential future enhancements.

II. LITERATURE SURVEY

The PAN evaluation campaign, held annually since 2013, serves to foster research in author profiling and associated tasks. Successful approaches from each edition have leveraged diverse feature representations, encompassing lexical, stylistic, content-based, and deep learning-based features [7]. These methodologies have consistently demonstrated high performance in predicting author attributes such as age, gender, native language, personality traits, and the use of hateful language across various textual datasets [6].

In the PAN 2014 edition, participants were tasked with identifying the author's age and gender using an expanded dataset comprising blog posts, tweets, and hotel reviews in both English and Spanish. The winning approach for age prediction employed a combination of lexical, stylistic, and content-based features for both English and Spanish texts [7]. Gender prediction, also for English and Spanish, relied on a diverse set of features, including lexical, morphological, and syntactic features.

The PAN 2015 edition introduced new author profiling tasks, including predicting the author's native language and personality traits, across datasets consisting of blog posts, tweets, and Face book posts in multiple languages [8]. For native language prediction, the winning approach utilized character n-grams and lexical features, while the top-performing method for personality trait prediction combined lexical and stylometric features.

In PAN 2016, the tasks were further expanded to include prediction of native language, gender, personality traits, and detection of hateful language, across Twitter and Face book posts in multiple languages [9]. The winning approach for native language prediction incorporated character n-grams and word embeddings, while gender and personality prediction utilized a combination of lexical, stylometric, and topic-based features. For hateful language detection, the winning approach employed a deep learning model with character-level embeddings and bidirectional LSTMs [10].

The winning approaches have included ensemble-based classification, content-based and style-based features, and second order representations. In 2015, the task was extended to four languages, and in 2016, the focus shifted towards cross-genre age and gender identification. The best performing system used combinations of stylistic features and the second order representation. The use of distributed representations of words, such as word2vec embeddings, has been limited in AP research. The doc2vec algorithm, which learns neural network-based document embeddings, has shown promise in previous research. This paper evaluates different parameters of the doc2vec algorithm and compares its performance with traditional feature representations. The evaluation includes both single- and cross-genre AP settings.

III. DATASET CHARACTERISTICS

Multilingual corpora were provided by task organizers. Corpora contain 14166 tweets from 152 English authors, 9879 tweets from 100 Spanish authors, 3687 tweets from 38Italian authors and 3350 tweets from 34 Dutch authors. Tweets were balanced by gender and unbalanced by age. The dataset provided this year consisted of tweets in Spanish, English Italian and Dutch. Regard to gender, the corpus was balanced in all four languages (50% of tweets were label as “female” and the other half as “men”).

TABLE I
FEMALE AND MALE DISTRIBUTION OF THE CORPUS

Language	Female		Male		Total samples
	Samples	Percentage	Samples	Percentage	
Spanish	50	50%	50	50%	100
English	76	50%	76	50%	152
Italian	19	50%	19	50%	38
Dutch	17	50%	17	50%	34

There were five personality traits to predict: extroverted, stable, open, conscientious and agreeable; each one of them with a possible value between -0.5 and +0.5. It is important to mention that the samples for personality traits were totally imbalanced. For example: in Italian, for the conscientious personality trait there were just 5 labels of the 11 possible ones (-0.5, -0.4, …, +0.4, +0.5), and the number of samples of these existing labels varied a lot.

TABLE IIIII
NUMBER OF SAMPLES PER LABEL IN EACH PERSONALITY TRAITS

		-0.5	-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5
Spanish	Extroverted			3		5	5	28	32	9	9	9
	Stable			2	10	26	9	12	19	10	10	2
	Agreeable				3	16	6	16	40	12	2	5
	Conscientious				2		21	7	20	12	21	17
	Open					7	10	37	15	9	14	8
English	Extroverted			1	4	10	17	41	37	20	13	9
	Stable			11	5	22	9	19	37	19	18	12
	Agreeable			5	2	12	19	44	46	13	7	4
	Conscientious				1	4	30	38	27	33	12	7
	Open					2	1	47	39	23	19	21
Italian	Extroverted						8	13	9		3	5
	Stable				1	3	3	8	4	12	5	2
	Agreeable					1	3	11	9	7		7
	Conscientious						3	18	6	5	6	
	Open						1	14	9	2	7	5
Dutch	Extroverted						3	5	11	7	6	2
	Stable				1	5	3	3	4	6	8	4
	Agreeable				2	1	5	10	10	2	4	
	Conscientious					2	4	15	6	5	2	
	Open							4	11	4	12	3

Classification attribute was created for each arff file with respect to each personality trait. We experimented with only English data set as shown in table 3.

TABLE IVVVI
ENGLISH DATASET

Gender	**Male**								
	76								
Age	**18-24**			**25-34**			**34-49**		**50+**
	58			60			25		12
Extroverted	**-0.3**	**-0.2**	**-0.1**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
	1	4	10	17	41	37	20	13	9
Stable	**-0.3**	**-0.2**	**-0.1**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
	11	5	22	9	19	37	19	18	12
Agreeable	**-0.3**	**-0.2**	**-0.1**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
	5	2	12	19	44	46	13	7	4
Conscientious	**-0.3**	**-0.2**	**-0.1**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
	0	1	4	30	38	27	33	12	7
Open	**-0.3**	**-0.2**	**-0.1**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
	0	0	2	1	47	39	12	19	21

IV. WORD EMBEDDINGS AND CLASSIFICATION ALGORITHMS

Word embeddings and classification algorithms play crucial roles in natural language processing (NLP). Word embeddings represent words as dense vectors of real-valued numbers, capturing latent linguistic characteristics and semantic relationships within text. By converting words into these numerical vectors, word embeddings facilitate the input of natural language data into various machine learning models [14]. Classification algorithms, on the other hand, are used to categorize text data into predefined classes. When combined with word embeddings, these algorithms can effectively process and analyse large datasets. The synergy between word embeddings and classification algorithms enhances the accuracy and efficiency of NLP applications, allowing for more sophisticated and nuanced understanding of language.

A. Word Embeddings

Natural language processing (NLP) has become increasingly popular in both research and business over the past few years. The advancement of computing power now allows for large-scale text processing, enabling the quantification of millions of words within hours. By quantifying text through language modelling, natural language can be input into statistical models and machine learning techniques. A common approach to text quantification involves representing each word in the vocabulary with a vector of real-valued numbers, known as a word embedding [11]. These numbers reflect scores of latent linguistic characteristics, and the trained word embeddings capture the similarities between words in text data.

1) *BERT Embeddings:* Bidirectional Encoder Representations from Transformers (BERT) builds upon transformers by utilizing a bidirectional approach. Similar to ELMo, BERT's training process is divided into two main phases: pre-training and fine-tuning. The pre-training phase consists of two tasks: predicting masked words from the input using a classifier and predicting the next sequence of words. Word masking addresses the issue in bidirectional models where transformer layers enable words to "see themselves." This is mitigated by masking 15% of input words during pre-training, either with a special token indicating the mask or by replacing them with different tokens. Transfer learning, which involves training an AI system for one task and then applying that knowledge to another task is exemplified in BERT's fine-tuning process. Depending on the specific task, such as sentence classification, question answering, or named entity recognition, the network adjusts its inputs and outputs accordingly. BERT is available in two versions: Small BERT and Large BERT. Essentially a trained Transformer Encoder stack, Small BERT includes 12 encoder layers, 12 attention heads, and 768 hidden units, while Large BERT comprises 24 encoder layers, 16 attention heads, and 1024 hidden units
2) *SimCSE Sentence Embedding:* Pre-trained models, such as BERT, GPT, and SimCSE, are increasingly utilized to provide high-quality sentence embeddings in natural language processing. These models are trained on vast corpora of text, capturing intricate patterns, semantics, and syntactic structures of language. When used for sentence embedding, these pre-trained models transform sentences into dense, fixed-dimensional vectors that encapsulate their meaning and context [15]. This process involves passing the sentence through the pre-trained model, which outputs an embedding that reflects the semantic content of the sentence. These sentence embeddings can then be employed in greatly enhancing the performance and accuracy of these applications due to the rich linguistic features encoded in the embeddings.

The specific details such as the number of nodes, layers, and activation functions used in neurons in the SimCSE pre-trained model depend on the underlying pre-trained transformer model it is built upon. Since SimCSE can be applied to various pre-trained language models, these details will align with the chosen base model. Here, we provide an example based on BERT, one of the commonly used models with SimCSE, you are using SimCSE with BERT-base, the model would have 12 layers, each with 768 hidden units, and it uses the GELU (Gaussian Error Linear Unit) activation function.

B. Classification Algorithms

Deep learning classification algorithms for text data, such as Convolution Neural Networks (CNNs), Recurrent Neural Networks (RNNs) like LSTM and GRU, Transformer models like BERT and GPT, as well as attention mechanisms and ensemble methods, have revolutionized natural language processing. These algorithms leverage deep neural networks to automatically learn complex patterns and representations from text, achieving high accuracy in different types of text applications. Transfer learning with pre-trained models further enhances performance by leveraging large-scale text data for fine-tuning on specific tasks, making deep learning a cornerstone in modern text classification solutions for various applications.

1) *Long Short-Term Memory (LSTM):* In the context of text classification, Long Short-Term Memory (LSTM) networks are pivotal due to their ability to effectively capture dependencies and sequential patterns within textual data. Unlike traditional recurrent neural networks (RNNs), LSTMs mitigate the vanishing gradient problem and facilitate the retention of relevant information over longer sequences. This capability is crucial for tasks such as sentiment analysis, where understanding the context and temporal dependencies of words or phrases within a sentence is essential for accurate classification. LSTMs excel in learning from sequences of text data, enabling them to model intricate relationships and nuances inherent in natural language, thereby enhancing the accuracy and performance of text classification systems.
2) *Convolution Neural Networks (CNN):* Convolution Neural Networks (CNNs) are adept at text classification tasks by leveraging their ability to capture local patterns and hierarchical representations within sequences of text. In CNN-based text classification, the input text is typically represented as a sequence of word embeddings or one-hot encoded vectors. The convolution layers in the CNN then apply filters of varying sizes over the input embeddings, detecting specific features or patterns (such as n-grams) at different levels of granularity. Max-pooling or average-pooling layers follow the convolutional layers to extract the most salient features from the feature maps generated by the convolutions. These pooled features are then fed into fully connected layers for final classification into predefined categories. CNNs excel in capturing spatial dependencies in images, and when adapted for text, they effectively capture local dependencies among neighbouring words, making them suitable for Author profiling task. Their ability to automatically learn hierarchical representations of text data makes CNNs a powerful choice for text classification tasks, achieving state-of-the-art performance.

V. PROPOSED METHOD

In this study, we present an approach for predicting personality traits using word embeddings and deep learning algorithms. The Figure 1 shows the proposed approach when the word embeddings are generated through BERT model. In this approach, the pre-trained small cased BERT (Bidirectional Encoder Representations from Transformers) model is used for personality trait prediction. The BERT model is a bidirectional transformer pre-trained using a combination of Masked Language Modelling (MLM).The small cased BERT model produces a 768-dimensional weighted vector for each word. In this approach, create a document for each author by combining all 100 tweets of individual authors. Apply pre-processing techniques like URL' removal, removal of punctuation marks, removal of stop words on personality traits dataset.

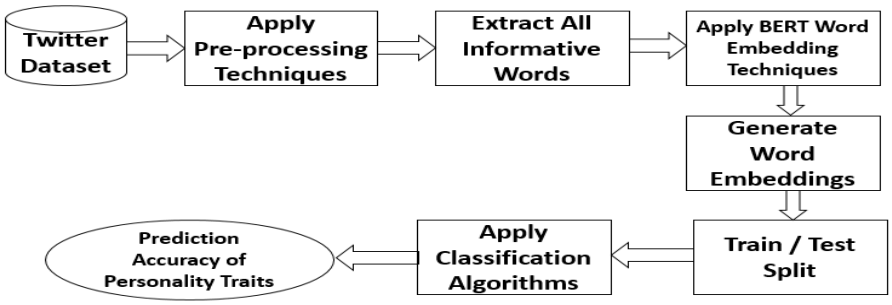


Fig. 1 Proposed Model using BERT Embeddings

Tokenize the text using BERT's Word Piece Tokenizer and divide the tokens of each document into groups of 500 tokens. If the last group contains fewer than 500 tokens, employ padding by adding zeroes to ensure each group consists of exactly 500 tokens. Each group of 500 tokens serves as input to a small cased BERT pre-training model, which produces 768-dimensional vectors for each token group. Utilizing max pooling, these vectors are consolidated into a single 768-dimensional vector representing each author[12]. The BERT model generates 768-dimensional vectors for each author in the dataset, which are subsequently used as input for machine learning algorithms during training. These algorithms evaluate the accuracy of the proposed approach.

Our enhanced proposed method integrates the BERT model for generating word embeddings and SimCSE for creating high-level sentence embeddings. These embeddings are then combined and utilized by two deep learning architectures, CNN and LSTM, to construct a classification model. As illustrated in Figure 2, our approach begins with preprocessing a personality traits Twitter dataset, involving steps such as tokenization, stop word removal, stemming, and the elimination of punctuation marks, hashtags, emojis, and retweets. Post-preprocessing, all terms are fed into the BERT model to generate 768-dimensional word vectors. These vectors form the basis for creating document vectors, which are further enriched with sentence embeddings.

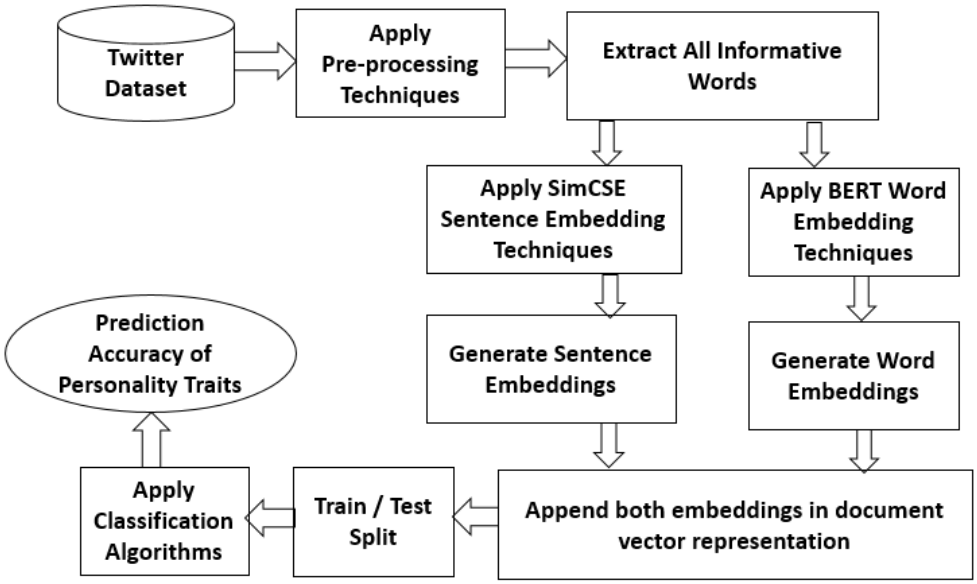


Fig. 2 Proposed Model for Personality Traits using BERT and SimCSE

The resulting document vectors are input to CNN and LSTM models, which internally construct the classification model and evaluate its accuracy. Deep learning techniques have proven effective in enhancing text classification accuracy by circumventing the need for manual feature identification, a crucial step in traditional machine learning approaches.

VI. EXPERIMENTAL RESULTS

In this study, we propose a method for predicting personality traits using word embeddings and deep learning classification algorithms. Table IV presents the experimental results of the proposed approach.

TABLE VIIv
THE ACCURACIES OF PROPOSED METHOD FOR PERSONALITY TRAITS PREDICTION

Profile / Embedding Techniques	Personality Traits Prediction using BERT	Personality Traits prediction using Combined BERT and SimCSE
CNN	75.32	78.45
LSTM	80.48	87.53

The proposed approach using the BERT + SimCSE model achieved the highest accuracies of 87.53 and 78.45 for personality trait prediction when the classification model was generated with LSTM and CNN, respectively. The combined BERT and SimCSE embeddings outperformed the BERT model embeddings alone in terms of accuracy for personality trait prediction. Among the classifiers, the LSTM demonstrated superior performance compared to the CNN classifier.

The results of the proposed method can be illustrated through graphical representations as shown in Fig. 3

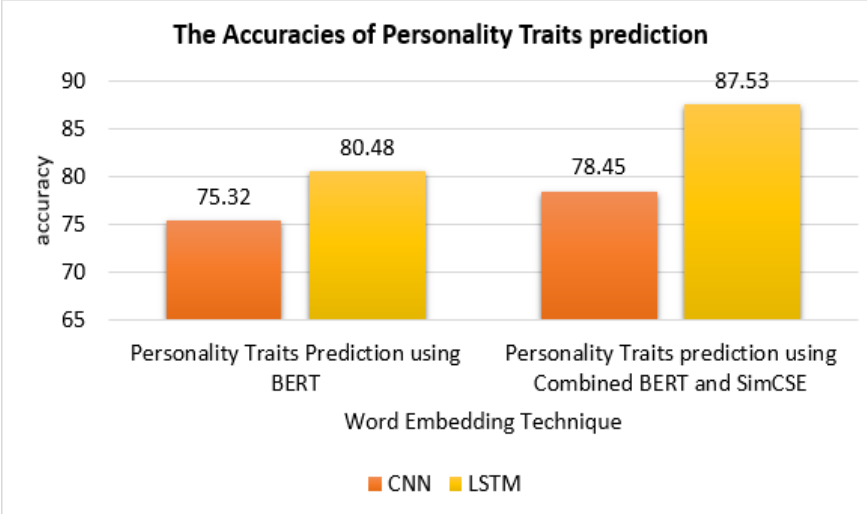


Fig. 3 Experimental Results analysis of Proposed Model for Prediction of Personality Traits using BERT and SimCSE

VII. CONCLUSIONS

In this study, we utilized the contrastive learning framework of SimCSE and applied the BERT pre-trained language model to extract text features for solving the PAN 2015 Author Profiling task. Our experimental results indicate that applying contrastive learning to natural language processing tasks, such as author profiling, can yield satisfactory outcomes. Additionally, the results demonstrate the powerful capability of the BERT model in text vector representation.

In future work, we plan to further refine our methods and aim to achieve improved results in open-domain author profiling tasks.

REFERENCES

[1] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "A Survey on Author Profiling Techniques", International Journal of Applied Engineering Research, March 2016, Volume-11, Issue-5, pp. 3092-3102.

[2] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Profile specific Document Weighted approach using a New Term Weighting Measure for Author Profiling ", International Journal of Intelligent Engineering and Systems, 9 (4), pp. 136-146, Nov 2016.

[3] Roy KhristopherBayot, Teresa Gon_calves, Multilingual Author Profiling using LSTMs Notebook for PAN at CLEF 2018

[4] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017). Enriching word vectors with subword information, Transactions of the association for computational linguistics 5: 135–146.

[5] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," en, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162. [Online]. Available: http://aclweb.org/anthology/D14-1162 (visited on 01/28/2021)

[6] Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In CLEF Conference on Multilingual and Multimodal Information Access Evaluation, CELCT, pp. 352-365 (2013).

[7] Kavuri, K., Kavitha, M. (2020). A Stylistic Features Based Approach for Author Profiling. In: Sharma, H., Pundir, A., Yadav, N., Sharma, A., Das, S. (eds) Recent Trends in Communication and Intelligent Systems. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-0426-6_20

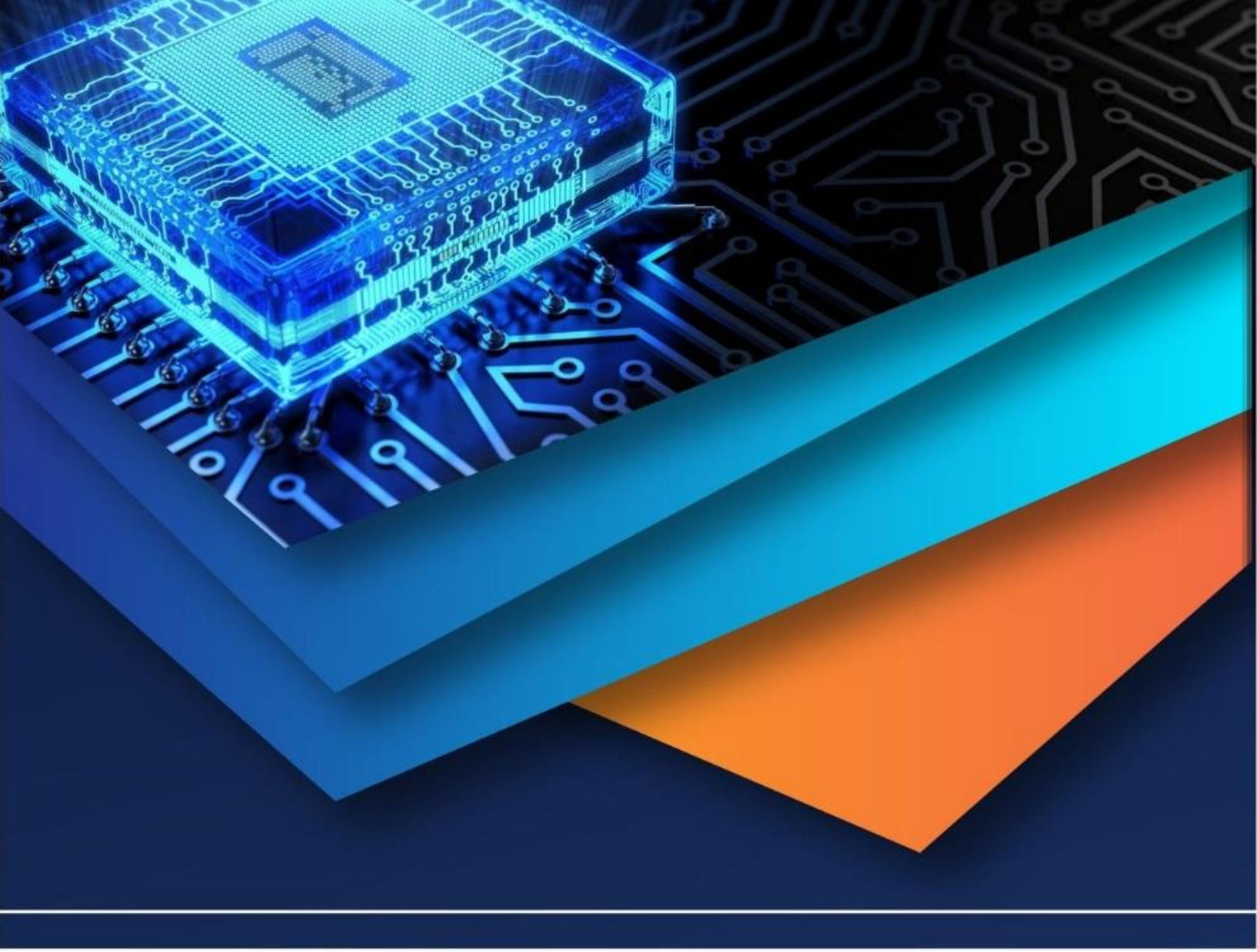
[8] Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B.,Daelemans, W.: Overview of the 2nd author profiling task at pan 2014. In CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014, pp. 1-30 (2014).

[9] Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In CLEF p. 2015 (2015).

[10] F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, and B. Stein, "Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations," CEUR Workshop Proc., vol. 1609, pp. 750–784, 2016.

[11] Kavuri, K. ., & Kavitha, M. . (2023). A Word Embeddings based Approach for Author Profiling: Gender and Age Prediction . International Journal on Recent and Innovation Trends in Computing and Communication, 11(7s), 239–250. https://doi.org/10.17762/ijritcc.v11i7s.6996.

[12] Kavuri, K., Kavitha, M. (2024). A Feature Selection Technique–Based Approach for Author Profiling Using Word Embedding Techniques. In: Lin, F.M., Patel, A., Kesswani, N., Sambana, B. (eds) Accelerating Discoveries in Data Science and Artificial Intelligence I. ICDSAI 2023. Springer Proceedings in Mathematics & Statistics, vol 421. Springer, Cham. https://doi.org/10.1007/978-3-031-51167-7_72
[13] Kavuri, Karunakar & Kavitha, M. (2022). A Term Weight Measure based Approach for Author Profiling. 275-280. 10.1109/ICESIC53714.2022.9783526.
[14] K Divya Jyothi, Kavuri. Karunakar et al, "A Word Embedding Techniques based Approach for Celebrity Profiling", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.10, Issue 6, page no.f202-f208, June-2023, Available :http://www.jetir.org/papers/JETIR2306529.pdf
[15] Qiu, Yong, Haoliang Qi, Yong Han and Kaicheng Huang. "Authorship Verification Based on SimCSE." *Conference and Labs of the Evaluation Forum* (2023).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY