iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.48280

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 10 Issue XII Dec 2022- Available at www.ijraset.com

Language Variety Prediction using Word Embeddings and Machine Leaning Algorithms

Chennam Chandrika Surya[1], Karunakar K[2], Murali Mohan T[3], R Prasanthi Kumari[4]
[1]*Pg-Student,* [2]*Associate Professor,* [3]*Professor,* [4]*Assistant Professor, Computer Science and Engineering, Swarnandhra Institute of Engineering and Technology, Narsapur, WG(Dt), Andhra Pradesh, India.*

Abstract: ***Author Profiling is a technique of predicting demographic characteristics like gender, age, location, nativity language, educational background etc., of an author by analysing their written texts. Author profiling is used in several text processing applications like forensics analysis, marketing, security. The author profiling techniques identify the stylistic differences among the author writing styles to identify the demographics of authors. Researchers experimented with various stylistic features like lexical features, content-based features, syntactic features, semantic features, domain specific features, structural features, readability features etc., to identify the stylistic differences among different author's texts. The dataset plays an important role to analyse the stylistic differences of authors. PAN is one competition organizes different types of tasks in every year to encourage the participants around the globe for providing solutions to different types of text classification problems like plagiarism detection, authorship attribution, authorship verification, authorship profiling, celebrity profiling, style change detection, fake news spreaders detection, hate speech spreaders detection etc. The author profiling task was introduced in 2013 by the organizers of PAN competition. The organizers carefully gather the datasets and make available to the researchers for providing solutions to the problems. Every year the organizers conduct competitions on different sub-tasks of author profiling and provides datasets in different languages and in different genres. In 2017 competition, PAN introduces a task of predicting the language variety of an author. They release the dataset in four languages. In this work, we proposed an approach for English language dataset of language variety prediction. The proposed approach used the word embeddings generated by the Word2Vec model and BERT (Bidirectional Encoder Representations from Transformers) model. The word embeddings are used for generating the document vectors by combining the word embeddings of words those contain in documents. The document vectors are trained with two machine learning algorithms such as support vector machine and random forest. The Random Forest attained best accuracy of 96.87 for language variety prediction when experiment conducted with BERT embeddings.***
Keywords: Author Profiling, Word Embeddings, Word2Vec, BERT, Machine Learning Algorithms

I. INTRODUCTION

The internet is increasing tremendously with full of text through social media, blogs, reviews etc. The crimes are also increasing with the text like harassing messages, threatening mails etc. There is a need of one research technique required to identify the basic information of authors who wrote the text. Author profiling is one such text classification technique, which is used to identify profiles like gender, age, location, native language and the personality traits of the author. In author profiling, linguistic features are used to determine the profile of an author and the most common techniques that are used are different kind of machine learning techniques.

Author profiling is an important task in many different applications like forensic analysis, marketing, security. For instance, from a marketing perspective, companies may be interested in knowing more about anonymous reviewers written reviews on various product review sites. In forensic linguistics, author profiling can be used to determine the linguistic profile of an author of a suspicious text. This is something that can be valuable for evaluating suspects and as a support in investigations. From an intelligence perspective, author profiling is used to gain more information about a possible suspect - this could for example be a potential violent lone actor that reveals an intention of committing targeted violence in a online setting, something that research has shown is the case in previous attacks. Author Profiling is used to extract as much information as possible about an author may increase law enforcement's chances to advance in their investigations.

The PAN competition starts competitions on author profiling in 2013 and provided good dataset for experimentation in every year. In this paper, we worked on the dataset of PAN 2017 for predicting the language variety of authors. The organizers released a Twitter dataset in four different languages. In this work, we proposed an approach for language variety prediction by using word embeddings and machine learning algorithms.

Two word embedding techniques such as Word2Vec model and BERT models are used for representing the documents as vectors. Two machine learning algorithms such as Support Vector Machine (SVM) and Random Forest (RF) are used for producing classification model by training on these document vectors.

This paper is structured in 6 sections. The existing work in Author Profiling for nativity language prediction is explained in section II. The dataset descriptions are presented in section III. Section IV discuss about the proposed approach, word embeddings, evaluations measures and machine learning algorithms. Section V analysed the experimentation results. The conclusions with future score are presented in section VI.

II. RELATED WORK

The researchers provided different types of methods for differentiating the writing styles of authors. Angelo Basile et al., participated [1] in the PAN 2017 shared task on Author Profiling for identifying authors' language variety for English, Spanish, Arabic and Portuguese. The aim of authors was to create a single model for all language varieties. The best-performing system (on cross-validated results) in their work is a linear support vector machine (SVM) with word unigrams and character 3- to 5-grams as features. A set of additional features, including POS tags, additional datasets, geographic entities, and Twitter handles, hurt, rather than improve, performance. Results from cross-validation indicated high performance overall and results on the test set confirmed them, at 0.86 averaged accuracy, with performance on sub-tasks ranging from 0.68 to 0.98.

Alina Maria Ciobanu et al., presented [2] a computational approach to author profiling for language variety prediction. They applied an ensemble system with the output of multiple linear SVM classifiers trained on character and word ngrams. The authors evaluated the proposed system using the dataset provided by the organizers of the 2017 PAN lab on author profiling. The proposed approach achieved 97% accuracy on language variety identification for Portuguese.

Don Kodiyan et al., described [3] an approach for the Author Profiling Shared Task at PAN 2017. The aim was to classify the language variety of a Twitter user solely by their tweets. Author Profiling can be applied in various fields like marketing, security and forensics. Twitter already uses similar techniques to deliver personalized advertisement for their users. PAN 2017 provided a corpus for this purpose in the languages: English, Spanish, Portuguese and Arabic. To solve the problem, the authors used a deep learning approach, which has shown recent success in Natural Language Processing. The submitted model consists of a bidirectional Recurrent Neural Network implemented with a Gated Recurrent Unit (GRU) combined with an Attention Mechanism. They attained an average accuracy of 85.22% over all languages for language variety classification.

Matej Martinc et al., presented [4] the results language variety identification performed on the tweet corpus prepared for the PAN 2017 Author profiling shared task. The proposed approach consists of tweet preprocessing, feature construction, feature weighting and classification model construction. They experimented with a Logistic regression classifier, where the main features are different types of character and word n-grams. Additional features include POS n-grams, emoji and document sentiment information, character flooding and language variety word lists. Proposed model attained the best results of 0.9838 on the Portuguese test set for language variety prediction tasks. The worst accuracy was achieved on the Arabic test set.

Eric S. Tellez et al., described [6] an approach to cope with the Author Profiling task on PAN17 which contain the task of language identification for Twitter's users. They used MicroTC (µTC) framework as the primary tool to create classifiers. µTC follows a simple approach to text classification and it converts the problem of text classification to a model selection problem using several simple text transformations, a combination of tokenizers, a term-weighting scheme, and finally, it classifies using a Support Vector Machine. Proposed approach reaches accuracies of 0.8275, 0.9004, 0.9554, and 0.9850 for language variety prediction in Arabic, English, Spanish, and Portuguese languages, respectively.

A. Pastor López-Monroy et al., described [7] the system for participating at CLEF-PAN 2017. They addressed the Author Profiling (AP) task by exploiting the corpus as a knowledge base. showing strong evidence of the usefulness of the representation to determine language variety. Furthermore, this representation can be seen as a natural extension to Second Order Attributes, which could be combined in future works in order to expose finer details about user relationships.

III. DATASET DESCRIPTION

In this work, PAN17-twitter dataset was used for experimentation which was released in 2017 [8]. This dataset consists of Twitter posts labelled with specific variation of their native language English (Australia (AS), Canada (CN), Great Britain (GB), Ireland (IL), New Zealand (NZ), United States (US)). The distribution of the classes is balanced. In total, PAN17-twitter consisted of 3,60,000 posts with a distribution of 60,000 posts for each class. Table 1 shows the characteristics of the dataset.

Table I
The Characteristics Of Pan 17 Twitter English Dataset For Nativity Language Prediction

Number of Authors	Number of Posts	Class Labels	
		Country Name	Number of Tweets
3600	360000	Ireland	60000
		Canada	60000
		Great Britain	60000
		New Zealand	60000
		United States	60000
		Australia	60000

IV. PROPOSED APPROACH

In this work, we proposed an approach for language variety prediction using word embeddings and machine learning algorithms. In the proposed approach, two word embedding techniques such as Word2Vec model and BERT model are used for generating word embeddings and two machine learning algorithms such as SVM and RF for generating classification model. The Figure 1 shows the proposed approach when word embeddings generated by the Word2Vec model. In this approach, first, the language variety dataset pre-processed by using different pre-processing techniques such as tokenization, stop word removal, stemming, removal of punctuation marks, removal of hash tags, emojis and re-tweets. After pre-processing the data from dataset, extract all terms from the dataset. All terms are forwarded to Word2Vec model for generating word embeddings for words. In this approach, each word is represented with 300 dimensional word vectors. These word vectors are used to create a document vector by combining the word embeddings of that document. All document vectors are passed to machine learning algorithms. These algorithms generate the classification model internally and compute the accuracy of proposed approach. In recent times, the deep learning techniques are more successful to improve the accuracy of text classification. The deep learning techniques avoid the features identification that is very important step in machine learning based approaches. The Figure 2 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 language variety 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 language variety dataset.

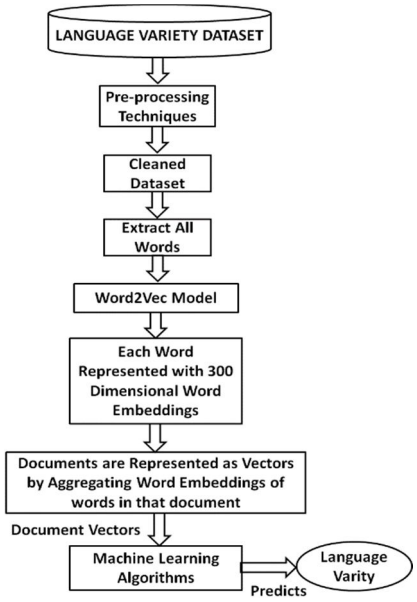


Figure 1: Proposed Model using Word2Vec Embeddings

Tokenize the text by using BERT's Word Piece Tokenizer. Split the tokens of each document into groups of 500 tokens. If last group contain tokens of less than 500, use padding technique of adding zeroes to make 500 tokens.

Each group of 500 tokens are given as input to small cased BERT pre-training model. This model generates 768 dimensional vectors for each group of 500 tokens. Max pooling technique is used to generate the 768 dimensional vectors for an author by consolidating different 768 dimensional vectors. The BERT model finally creates 768 dimensional vectors for each author in the dataset. These vectors are given to machine learning algorithms for training. The machine learning algorithms predicts the accuracy of proposed approach.

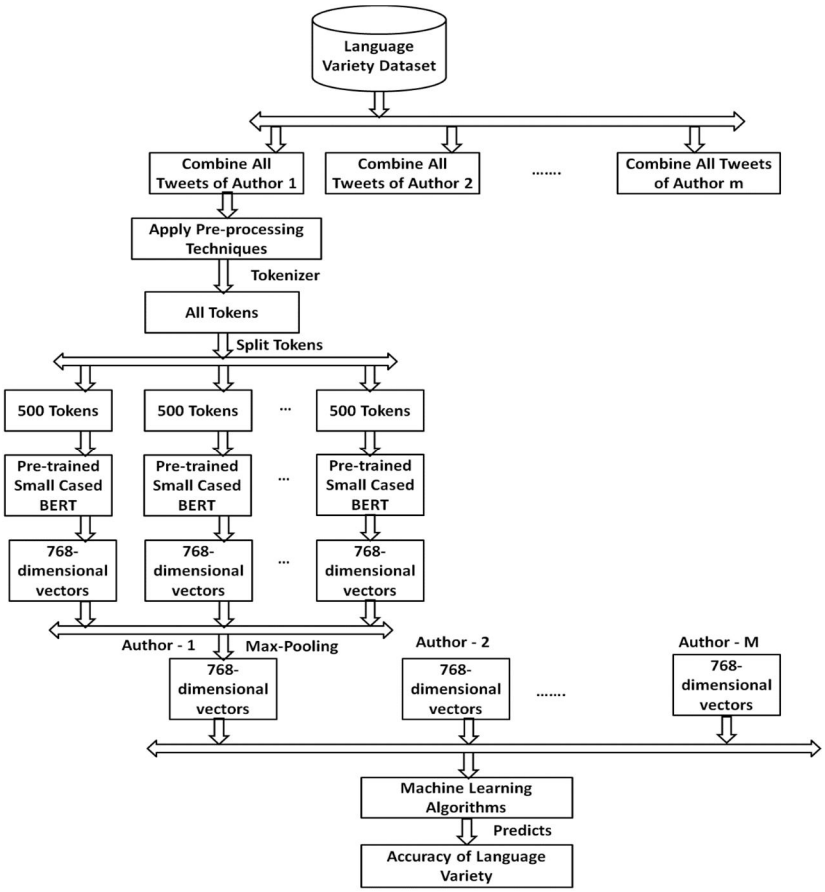


Figure 2: Proposed Model using BERT Embeddings

A. Word Embeddings

Natural language processing has gained popularity in both research and business over the past years. The increase of computing power enables one to process text on a large scale, quantifying millions of words within hours. Language modelling by the quantification of text allows users to feed natural language as input to statistical models and machine learning techniques. A popular approach to quantify text is to represent each word in the vocabulary by a vector filled with real-valued numbers, called a word embedding. The numbers in these word embeddings represent scores of latent linguistic characteristics. The trained word embeddings capture similarities of words in text data.

1) Word2Vec Model

A word embedding is based on the words surrounding the word of interest. These surrounding words are the so-called context words. The estimation procedure of word2vec is based on the prediction power of words to predict other words in the neighbourhood, the so-called local context window, of those words in the text. Word2Vec model has two variants such as CBOW (Continuous Bag Of Words) and Skip-Gram model. The CBOW variant, on the other hand, compares each word with the average representation of the surrounding words and the word in the center of the context window is predicted given its context. The process of CBOW model is represented in Figure 3.

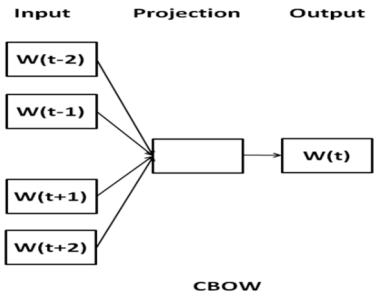


Figure 3: CBOW Model

The Skip-Gram variant of word2vec uses word by word similarity comparisons and tries to predict a word in the local context window given the word in the middle of this context window. The process of Skip-Gram model is represented in Figure 4.

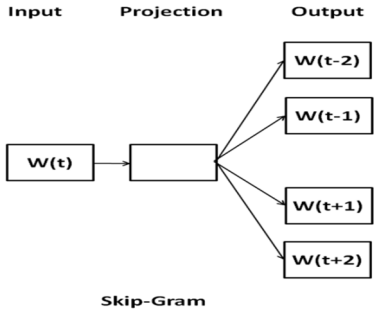


Figure 4: Skip-Gram Model

2) BERT Model

Bidirectional Encoder Representations from Transformers (BERT) follows up on transformers in a bidirectional manner [20]. BERT also follows ELMo in that training involves a split between two tasks, pre-training and fine-tuning. Pre-training for BERT splits into a further two tasks, the first will be predicting masked words from an input with a classifier and the second task is to predict the next sequence of words. Word masking is done to solve an issue of a bidirectional model, transformer layers enable words to be able to see "see themselves". This is solved by masking 15% of input words during the pre-training process, with either a token indicating that token is masked, or replace with another token. Transfer learning is the process of learning an AI system for a particular task, then applying what has been learned to another task. The fine tuning process of BERT is an example of transfer learning. Depending on the task such as sentence classification, question answering, and named entity recognition, the network will change the input and outputs accordingly. BERT is released with two varieties small BERT and Large BERT. BERT is basically a trained Transformer Encoder stack. Small BERT contains 12 encoder layers, 12 attention heads and 768 hidden units. Large BERT contains 24 encoder layers, 16 attention heads and 1024 hidden units.

B. Evaluation Measures

The outcomes of machine learning algorithms are represented with different evaluation measures like recall, precision, F1-score and accuracy. To represent these measures, a confusion matrix is required that is mentioned in Table II.

Table II
Confusion Matrix For Evaluation Measures

		Predicted label	
Actual Label		C_p	C_n
	C_p	TP	FN
	C_n	FP	TN

Confusion matrix is a visualization of the predicted and actual classification results in the form of table with size n x n, where n is a number of classes. The confusion matrix represents the way the model is confused when making predictions.

In this matrix, C_p and C_n are the positive class and negative class respectively. TP and TN are number of positive class and negative class documents are correctly predicted as positive class and negative class respectively. FN and FP are number of positive class and negative class documents are incorrectly predicted as negative class and positive class respectively.

Precision is a ratio among number of positive instances correctly predicted and number of instances predicted as positive. Equation (1) is used to determine the precision.

$$\Pr ecision = \frac{TP}{(TP + FP)} \tag{1}$$

Recall is the ratio among number of positive instances correctly predicted and number of instances in positive class. Equation (2) is used to compute the recall measure.

$$\mathrm{Re}\, call = \frac{TP}{(TP + FN)} \tag{2}$$

F-measure is the weighted average of recall and precision. Equation (3) is used to calculate the F-measure.

$$F - measure = \frac{2 * \Pr ecision * \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \tag{3}$$

Accuracy is a popular measure which is used in different fields of research to evaluate the performance of proposed approaches. Accuracy is the ratio among the number of instances predicted correctly and number of instances in the dataset. Equation (4) is used to compute the accuracy measure.

$$Accuracy = \frac{Number\ of\ Correct \Pr edictions\ (TP + TN)}{Total\ Number\ of\ Examples\ (TP + TN + FP + FN)} \tag{4}$$

In this work, accuracy measure is used for presenting the experimental results of proposed approach.

C. Machine Learning Algorithms

ML is a set of methods for automatically detecting patterns in existing data and making predictions about future data. There are two types of ML methods such as supervised (predictive) and unsupervised (descriptive). In case of supervised ML, building ML model consists of training the model on data containing set of observations and automatically predicting target output based on previously learned knowledge. In case of unsupervised ML, the samples of dataset are divided into different clusters based on the similarity among samples. Classification is a type of supervised ML method for mapping a set of unlabelled inputs (data instances) to the categorical output variable called as class. Regression is predicting the target value of continuous output variable that is real-valued like weight, price of the house, etc.

The machine learning algorithms generates the classification model by training the machine with training data. The classification model is used to predict the class label of test documents as well as to determine the efficiency of the proposed methods. In this work, two ML algorithms such as SVM, and RF are used to evaluate the performance of proposed approaches for language variety prediction.

1) Support vector Machine (SVM)

The SVM is popularly used in various research domains like text classification, sentiment analysis, pattern recognition etc., to handle both regression and classification problems. The SVM classifier was proposed [9] by C., Vapnik, V., 1995. In SVM classifier, hyperplanes were also called as support vectors that are identified to increase marginal difference between numerous classes. Support vectors utilize the data points that are closest to the decision boundary that separates two classes, which are the points that impacts on the hyperplane's orientation. Hyperplanes are used to categorize and divide the data into different classes. To handle, multiple categories of data, different types of kernels such as sigmoid kernel, RBF kernel and linear kernel are developed in SVM.

2) Random Forest (RF)

Random Forest is widely used and successful in various research domains like information retrieval, sentiment analysis, text classification etc. The random forest algorithm constructs a collection of N decision trees based on N bootstrap samples with replacement of samples from the original sample. Each decision tree constructed with a different set of samples which are randomly selected from the training dataset samples.

The RF classifier used majority voting technique to assign a class label to a new sample. The new sample was given to all decision trees and collects the decisions of all decision trees based on the path that new sample reaches from root to a class label. The class label is assigned to a new sample based on majority decisions of decision trees [10].

V. EXPERIMENTAL RESULTS

In this work, we proposed an approach for language variety prediction by using word embeddings and machine learning algorithms. The Table 2 shows the experimental results of proposed approach.

Table III
The Experimental Results Of Proposed Approach For Language Variety Prediction

Word Embeddings / Machine Learning Algorithms	Support Vector Machine	Random Forest
Word2Vec	87.63	90.51
BERT	92.76	96.87

The proposed approach with BERT model attained best accuracies of 92.76 and 96.87 for language variety prediction when classification model generated with SVM and RF respectively. The BERT embeddings produce best accuracy for language variety prediction when compared with the word embeddings of Word2Vec model embeddings. The RF classifier shows best performance when compared with the performance of SVM classifier.

VI. CONCLUSIONS AND FUTURE SCOPE

The author profiling techniques are heavily used in various social media platforms to identify the basic information of suspected authors. In this work, we conducted experiment for predicting the language variety of authors. We proposed an approach for language variety prediction by using word embeddings and machine learning algorithms. The Random Forest classifier with BERT embeddings attained an accuracy of 96.87% for language variety prediction.

In future work, we are planning to implement large case BERT model for generating word embeddings and also planned to implement other machine learning algorithms to improve the accuracy of language variety prediction.

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, "A Document Weighted Approach for Gender and Age Prediction", International Journal of Engineering -Transactions B: Applications, Volume 30, Number 5, pp. 647-653, May 2017.
[3] Dr. T. Murali Mohan, Dr. T. Raghunadha Reddy, Dr. A. Balakrishna, T V Satya Sheela, "Stylistic features based Approach for Bot Detection", Design Engineering, Issue: 7, 2021, Pages: 12699 – 12712
[4] Para Upendar, T Murali Mohan, S. K. LokeshNaik, T Raghunadha Reddy, "A Novel Approach for Predicting Nativity Language of the Authors by Analyzing their Written Texts", SPRINGER 6th International Conference on Innovations in Computer Science and Engineering, Guru Nanak Institute of Technology, Hyderabad, Telangana, 17-18, August 2018.
[5] Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim, N-GrAM: New Groningen Author-profiling Model Notebook for PAN at CLEF 2017
[6] Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, Liviu P. Dinu, Including Dialects and Language Varieties in Author Profiling Notebook for PAN at CLEF 2017
[7] Don Kodiyan, Florin Hardegger, Stephan Neuhaus, and Mark Cieliebak, Author Profiling with Bidirectional RNNs using Attention with GRUs Notebook for PAN at CLEF 2017
[8] Matej Martinc, Iza Škrjanec, Katja Zupan, and Senja Pollak, PAN 2017: Author Profiling - Gender and Language Variety Prediction Notebook for PAN at CLEF 2017
[9] Sebastian Sierra, Manuel Montes-y-Gómez, Thamar Solorio3 and Fabio A. González, Convolutional Neural Networks for Author Profiling Notebook for PAN at CLEF 2017
[10] Eric S. Tellez, Sabino Miranda-Jiménez, Mario Graff, and Daniela Moctezuma, Gender and language-variety identification with MicroTC Notebook for PAN at CLEF 2017
[11] A. Pastor López-Monroy, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, and Thamar Solorio, Social-Media Users can be profiled by their Similarity with other Users Notebook for PAN at CLEF 2017
[12] https://pan.webis.de/clef17/pan17-web/author-profiling.html#data

[13] T. Raghunadha Reddy, P. Vijayapal Reddy, T Murali Mohan, Raju Dara, "An Approach for Suggestion Mining based on Deep Learning Techniques", International Conference on Computer Vision, High Performance Computing, Smart Devices and Networks(CHSN-2020), 28-29 December, 2020, JNTUK, Kakinada, Andhra Pradesh.

[14] Raghunadha Reddy T, Vishnu Vardhan B, GopiChand M, Karunakar K, "Gender prediction in Author Profiling using ReliefF Feature Selection Algorithm", Proceedings in Advances in Intelligent Systems and Computing, Volume 695, PP. 169-176, 2018.

[15] Swathi Ch, Karunakar K, Archana G, T. Raghunadha Reddy, "A New Term Weight Measure for Gender Prediction in Author Profiling", Proceedings in Advances in Intelligent Systems and Computing, Volume 695, PP. 11-18, 2018.

[16] P Buddha Reddy, Dr. T Murali Mohan, Dr. P Vamsi Krishna Raja and Dr. T Raghunadha Reddy, "A Novel Approach for Authorship Verification", SPRINGER 3rd International Conference on Data Engineering and Communication Technology (ICDECT), Stanley College of Engineering and Technology for Women, Abids, Hyderabad, Telangana, India, 15 – 16 March, 2019.

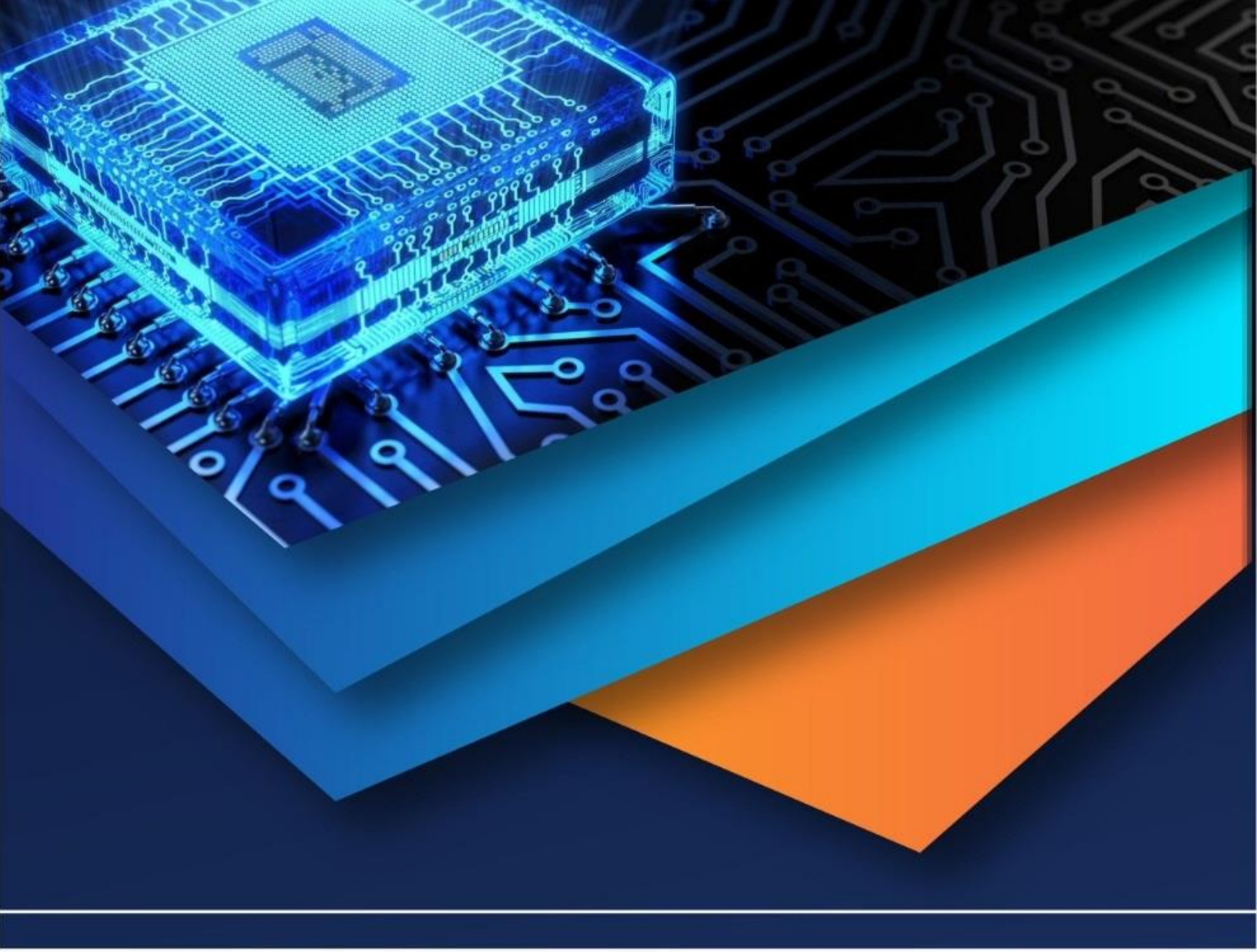
[17] Srikanth Reddy G, Murali Mohan T, Raghunadha Reddy T, "Author Profiling Approach for Location Prediction", first international conference on Artificial Intelligence and Cognitive computing conducted by MLR Institute of Technology, Dundigal, Hyderabad, 02-03 February, 2018**.**

[18] Cortes, C. & Vapnik, V. Machine Learning (1995) 20: 273. https://doi.org/10.1023/A:1022627411411

[19] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32, https://doi.org/10.1023/A:1010933404324

[20] Karunakar Kavuri, 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

[21] Karunakar. Kavuri and M. Kavitha, "A Term Weight Measure based Approach for Author Profiling," 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), 2022, pp. 275-280, doi: 10.1109/ICESIC53714.2022.9783526.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY