JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Word embedding Technique with Similarity Measures based Approach for Author Profiling

[1]Anusha Villuri, [2]Yugandhar Bokka, [3]P. Haritha, [4]R Prasanthi Kumari

[1]PG-Student, [2, 3] Associate Professor, [4] Assistant Professor

[1, 2, 3, 4] Department of CSE, Swarnandhra Institute of Engineering and Technology, Narasapur, AP, India.

Abstract : Author Profiling (AP) is a type of text classification and information extraction technique, which extracts the author's hidden information from their written texts. This technique extracts the author's demographic features like gender, age, location, nativity language, educational background, personality traits etc., by analysing their written texts. The author profiling techniques are used in various applications like marketing, security, forensic analysis etc. Researchers proposed several methods for differentiating the writing style of authors by using stylistic features, machine learning techniques and deep learning techniques. In this work, we proposed two approaches based on word embedding techniques with machine learning algorithms and word embedding techniques with similarity measures for author profiling. We are concentrating on prediction of gender and age group of authors. The PAN competition 2014 dataset is used in this work for experimentation. In the first proposed approach, we used word embedding techniques for representing words as vectors. These word vectors are used for representing the documents as vectors. The machine learning algorithms are used for training these document vectors. These algorithms develop the classification model to predict the accuracy of gender and age prediction. In the second approach, we used word embedding techniques for representing words as vectors and similarity measures are used to find the similarity among the documents. For gender and age dimensions, the second approach attained good accuracies than first approach. The proposed approaches attained best accuracies for gender and prediction when compared with other popular approaches to author profiling.

IndexTerms **- Author Profiling, Gender Prediction, Age Prediction, Word2Vec, Doc2Vec, Similarity Measures.**

I. INTRODUCTION

The web is growing constantly with a huge amount of text mainly through blogs, twitter tweets, reviews and other social media. Most of this text is written by various authors in different contexts. The challenges with the availability of text faced by the researchers and information analysts are to develop automated tools for information analysis. Authorship analysis is one such area attracted by several researchers to extract information from the anonymous text. Authorship analysis is a procedure of finding the authorship of a text by inspecting its characteristics. Authorship Analysis is classified into three categories such as Authorship Attribution, Authorship Verification and Author Profiling [1].

Authorship Attribution predicts the author of a given anonymous document by analyzing the documents of multiple authors [2]. Authorship Verification compares multiple pieces of written text of a single author and checks whether it is written by the same author or not without identifying the author [3]. Author Profiling is a type of text classification task, which predicts the profiling characteristics such as gender, age, occupation, location, native language, personality traits and educational background of the authors by analysing their writing styles [4].

Author Profiling is an important technique in the present information era which is used in several applications like forensic analysis, educational domain and marketing [4]. Social web sites are an integral part of our lives through which, crimes are cropping up like public embarrassment, fake profiles, defamation, blackmailing and stalking in the form of textual data. Forensics is a field to analyse the style of writing, signatures, documents, and anonymous letters. In this aspect, Author Profiling techniques are useful in crime investigation and forensic analysis to identify the perpetrator by analysing the text characteristics. In the marketing domain, the consumers were provided with a space to review the product. Most of the reviewers were not comfortable in revealing their personal identity like gender, age and location. Author Profiling techniques are helpful to analyse these anonymous reviews and classify the consumers based on their age, gender, occupation, nativity language and country. Based on the classification results, companies try to adopt new business strategies to serve their customers. Author Profiling is also beneficial in educational domain in estimating exceptional talent and the suitable level of knowledge of each student or a student group in the educational forum by analyzing the posts of the student.

Author Profiling techniques mainly depends on the writing styles of the authors. A general assumption made by observing various datasets is that the female are more expressive and their emotional involvement in their writings is considerably more than that of males. Female writings contain more positive and negative words than male. Another assumption is that male are tend to tell stories by focusing on what happened, while female focus more on how they felt and when these stories happened, instead of focusing on the story itself [1].

According to Koppel et al. [4], woman use more pronouns in their writings and men use more number of determiners and quantifiers. Similarly the female authors write about topics like shopping, kitty parties and beauty whereas the male authors concentrate more on topics related to politics, sports and technology. Prior works [1, 4] found that the male authors use more number of prepositions in their articles and blog posts when compared to female authors. Generally the writing styles of the authors vary based on the selection of topics and the writing styles like grammar rules and choice of words. In an observation [5], the female writings were observed with wedding styles and male write more about politics and technology. Further females use more adverbs and adjectives than male authors.

The content based features are more useful to distinguish the writing styles of female and male authors. The occurrence of words like my husband, pink and boyfriend increases the chances of text written by female and the occurrence of words like cricket and world cup increases the chances of text written by male [1]. Females are more likely to include pronouns, negations, verbs, friends, words related to home, family and various emotional words. Males tend to use more number of prepositions, articles, numbers and longer words [6].

In this work, we proposed two approaches by using machine learning algorithms and similarity measures. In both approaches, word embedding techniques are used for representing the words as vectors. In the first approach, the documents are created by using word vectors and are trained with machine learning algorithms. Machine learning algorithms predict the accuracy of age and gender prediction. In the second approach, the similarity measures are used to find the similarity among documents. The accuracy is predicted based on the similarity among the test and training documents.

This paper is organized in 7 sections. The section 2 describes the previous works proposed for author profiling. The dataset characteristics are presented in section 3. The proposed approach 1 is presented and discussed in section 4 with the required components for implementing the proposed approach. The section 5 explains the proposed approach 2 with the components used in the proposed approach. The experimental results are presented and discussed in section 6. The section 7 concludes this work with future enhancements.

II. Related Works

Author profiling is technique of predicting demographic features of authors by processing their writing styles. The PAN evaluation campaign has been organized annually since 2013 to promote studies on author profiling and related tasks. The winning approaches of each edition have used a variety of features and feature representations, including lexical, stylistic, content-based, and deep learning-based features, and have achieved high performance on predicting the author's age, gender, native language, personality traits, and use of hateful language in various types of text data [7].

In the competition of PAN 2014 edition, the task was to identify the author's age and gender, but with an expanded dataset including blog posts, tweets, and hotel reviews in both English and Spanish. The winning approach for the age prediction task on English and Spanish was based on a combination of lexical, stylistic, and content-based features [8]. For gender prediction on English and Spanish, the winning approach relied on a variety of features, including lexical, morphological, and syntactic features.

In the competition of PAN 2015 edition [9] introduced new author profiling tasks, including predicting the author's native language and personality traits. The dataset consisted of blog posts, tweets, and Facebook posts in multiple languages. For the native language prediction task, the winning approach used a combination of character n-grams and lexical features. For the personality trait prediction task, the best-performing approach was based on a combination of lexical and stylometric features.

In the competition of PAN 2016 [10], the task was again expanded to include prediction of the author's native language, gender, and personality traits, as well as the use of hateful language. The dataset included Twitter and Facebook posts in multiple languages. The winning approach for native language prediction used character n-grams and word embeddings, while the best-performing approach for gender and personality prediction relied on a combination of lexical, stylometric, and topic-based features. For the hateful language detection task, the winning approach was based on a deep learning model using character-level embeddings and bidirectional LSTMs.

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.

Author profiling is the process of analysing textual data to extract information about various personality traits of the author. It has both commercial and social implications. Many approaches have been proposed in past to increase the accuracy of the extracted information. In [11], researchers applied natural language processing and machine-learning approach for author profiling. NLP techniques like Tokenization, lemmatization, word and char n-grams are used in integration with machine learning classifiers like logistic regression (LR), random forest (RF), decision tree (DT) and support vector machine (SVM). The proposed method obtained an accuracy of 81.2%, 79.8%, 63.2% and 88.0% for the four classifiers respectively for gender prediction and 72.5%, 68.1%, 53.7% and 81.0% respectively for age prediction i.e. SVM outperformed the other classifiers with an accuracy of 88.0% for gender prediction and 81.0% for age prediction.

III. Dataset Characteristics

In The corpus used in this work for gender and age prediction was collected from PAN 2014 hotel reviews corpus. Table 1 depicts the characteristics of the reviews dataset for gender and age dimension.

Table 1: Dataset characteristics of gender and age profiles

S. No	Age Group	Documents Count	Male Documents Count	Female Documents Count
1	18_24	360	180	180
2	25_34	1000	500	500
3	35_49	1000	500	500
4	50_64	1000	500	500
5	65+	800	400	400
Total Count of Documents		4160	2080	2080

The corpus was constructed carefully to ensure its quality with regard to text cleanliness and annotation accuracy. In order to make this dataset applicable to Author Profiling and to ensure its quality, reviews containing less than 5 lines of text were excluded from our dataset and the reviews written by the authors whose gender was given in their user profile. In this work, two author profiles such as gender, and age were considered for analysis. The corpus is balanced in gender dimension, but unbalanced in case of age dimension.

IV. Word Embedding Techniques based Approach for Author Profiling using Machine Leaning Algorithms

The proposed word embedding techniques based approach for author profiling using machine learning algorithms are displayed in Figure 1. In the proposed approach, we used the PAN 2014 competition reviews dataset. The dataset has cleaned by using different pre-processing techniques such as lowercase conversion, punctuation marks removal, stop-word removal, and lemmatization. From the cleaned dataset, extract all informative words. All these informative words are forwarded to two modules such as embedding technique of word2vec for generating word embedding vectors and doc2vec module to generate the document vectors. Word2vec based word vectors are used to generate document vectors. These 2 modules document vectors are trained with machine learning algorithms. The machine learning algorithms predicts the accuracy of age and gender prediction.

4.1 Word2Vec

Word2vec is a combination of models used to represent distributed representations of words in a corpus C. Word2Vec (W2V) is an algorithm that accepts text corpus as an input and outputs a vector representation for each word. The vectors we use to represent words are called neural word embeddings, and representations are strange. One thing describes another, even though those two things are radically different. So, a neural word embedding represents a word with numbers. It's a simple, yet unlikely, translation. Word2vec is similar to an autoencoder, encoding each word in a vector, but rather than training against the input words through reconstruction, as a restricted Boltzmann machine does, word2vec trains words against other words that neighbour them in the input corpus.

There are two flavours of this algorithm namely such as Continuous Bag Of Words (CBOW) and Skip-Gram. Given a set of sentences (also called corpus) the model loops on the words of each sentence and either tries to use the current word *w* in order to predict its neighbours (i.e., its context), this approach is called "Skip-Gram", or it uses each of these contexts to predict the current word *w,* in that case the method is called "Continuous Bag Of Words" (CBOW). To limit the number of words in each context, a parameter called "window size" is used.We use the latter method because it produces more accurate results on large datasets.

4.2 Doc2Vec

Numeric representation of text documents is a challenging task in machine learning and such a representation is used in many applications such as document retrieval, web search, spam filtering, topic modelling etc. The goal of doc2vec is to create a numeric representation of a document, regardless of its length. Doc2Vec is an extension of Word2vec that encodes entire documents as opposed to individual words. Doc2Vec vectors represent the theme or overall meaning of a document. In this case, a document is a sentence, a paragraph, an article, or an essay, etc. In Doc2Vec, the name of the document, like file name or file ID will be the input, and the sliding window of the words from the document is the output.

4.3 Machine Learning Algorithms

Machine Learning Algorithms also known as classifiers usually have three main parts. First, an algorithm that will create a model for the data which can be called as modelling, then a training part which entails training our model with training data, and finally a testing part. The training phase of a classifier is responsible for creating the training model. Training model loop through the training data and calculates necessary probabilities to create the model. The testing phase of a classifier is responsible for testing the training model and finding the success rate of it. In this work, K-fold cross-validation testing method was used to test our classifier. K-fold cross-validation has K iterations. In every iteration, one random unit datum is selected for testing and the remaining K-1 units are used for training. This process is repeated K times while each randomly selected unit is used exactly once. The next sections explain two classifiers such as Random Forest and support vector machines which are used in our experimentation.

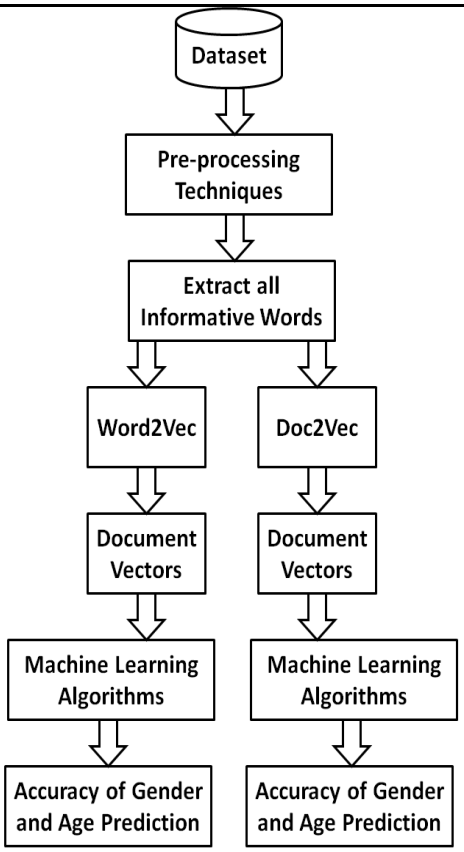

Figure 1: Word Embedding Techniques based Approach for Author Profiling using Machine Leaning Algorithms

2.1. Random Forest Classifier

Ensemble classification is a function of ensemble learning to increase the efficiency of classification. Ensemble classification is more accurate than individual classifiers because it uses more number of classifiers as an ensemble. Then a voting method is used to predict the class label of unknown document. Majority voting [12] is a simple and effective voting scheme followed in ensemble classification. In majority voting, every classifier in the ensemble is participated to predict the class label of an unknown document. Once all the classifiers in the ensemble submitted their vote on class label of an unknown document, the ensemble classification method assigns a class label based on which class label gained greatest number of votes. Random forest classifier is an ensemble learning method developed by [13]. It is used for regression and classification and it combines the random selection of features and the bagging approach. The random selection of features constructs a set of decision trees with controlled deviation. Bagging is used to construct a decision tree from the training data by using a sampling with replacement technique. Each decision tree in the ensemble operates as a base classifier to predict the class label of an unknown document. This is done through majority voting where each classifier give its vote to decide the class label of unknown document, then the class label with most of the votes are assigned to the unknown document.

2.2. 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 [14] 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.

V. Word Embedding Techniques based Approach for Author Profiling using Similarity Measures

The proposed word embedding techniques based approach for author profiling using similarity measures are displayed in Figure 2. In the proposed approach, we used the PAN 2014 competition reviews dataset. The dataset has cleaned by using different pre-processing techniques such as lowercase conversion, punctuation marks removal, stop-word removal, and lemmatization. From the cleaned dataset, extract all informative words. All these informative words are forwarded to an embedding technique of word2vec for generating word embedding vectors. The dataset is divided into training documents and testing documents. Word2vec based word vectors are used to generate training documents and testing documents as vectors. The class label is decided for test document based on the similarity among the test document and training documents. The predicted class label is compared with the original known class label of test document. The accuracy of gender and age is predicted based on the number of test documents are correctly identified their class label.

5.1. Similarity Measures

A similarity measure is a function which determines the degree of similarity between a pair of textual objects [15]. Similarity measures are very important for document clustering and Text-Mining. In general, these measures were used to calculate similarity between two queries, two documents and one document and one query. Similarity measures also used to rank the

documents based on the similarity scores between the document and query. In this work the experimentation carried out with cosine similarity measure. The next subsection explains the cosine similarity measure.

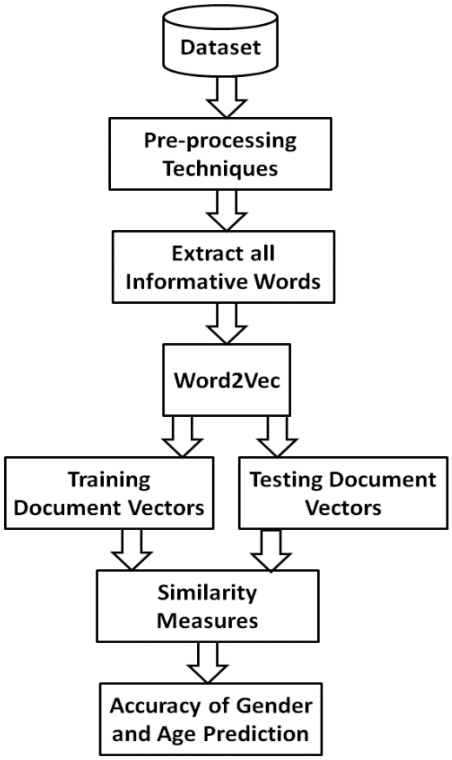

Figure 2: Word Embedding Techniques based Approach for Author Profiling using Similarity Measures

5.1.1 Cosine Similarity Measure (CSM)

In VSM, the sets of documents and queries are viewed as vectors. Cosine similarity measure is a popularmethod for calculating the similarity value between the vectors [16]. With document and queries beingrepresented as vectors, similarity signifies the proximity between the two vectors. Cosine similarity measurecomputes similarity as a function of the angle made by the vectors. If two vectors are close, the angle formedbetween them would be small and if the two vectors are distant, the angle formed between them would be large.The cosine value varies from +1 to -1 for angles ranging from 0 to 180 degrees respectively, making it the idealchoice for these requirements. A score of 1 evaluates to the angle being 0^o, which means the document aresimilar. While a score of 0 evaluates to the angle being 90^o, which means the documents are entirely dissimilar.The cosine weighting measure is implemented on length normalized vectors for making their weights comparable. Equation (1) gives the formula for Cosine Similarity.

$$C\,SM\left(D_p, D_q\right) = \frac{\sum_{k=1}^{n} WT\left(T_k, D_p\right) \times WT\left(T_k, D_q\right)}{\sqrt{\sum_{k=1}^{n} \left(WT\left(T_k, D_p\right)\right)^2} \times \sqrt{\sum_{k=1}^{n} \left(WT\left(T_k, D_q\right)\right)^2}} \quad (1)$$

Where, $WT(T_k, D_p)$ and $WT(T_k, D_q)$ are the weights of the term $T_k$in documentsD_p and document D_q respectively.

VI. Experimental Results

In this work, the experiment performed for predicting the accuracies of gender and age prediction. Two approaches are proposed for predicting gender and age.

6.1. Evaluation Measures

A number of different conventional performance measures are available for evaluating the effectiveness of Author Profiling Approaches. The definition of almost all measures is based on the same 2×2 contingency table that is constructed as shown in Table 2.

Table 2: Contingency table for category c_i.

Category ci		Original labels of documents	
		YES is correct	NO is correct
Label by the system	Predicted YES	TPi	FPi
	Predicted NO	FNi	TNi

In this Table 2, 'YES' and 'NO' represent a binary decision given to each document *dj* under category *ci*. Each entry in the table indicates the number of documents of the specified type. TP_iisthe number of true positive documents that the system predicted were YES, and were in fact in the category c_i.$FP_i$is the number of false positive documents that the system predicted were YES, but actually were not in the category c_i.$FN_i$is the number of false negative documents that the system predicted were NO, but was in fact in the category c_i.$TN_i$is the numbers of true negative documents that the system predicted was NO, and actually were not in the category c_i.

In this work accuracy measure is used to measure the effectiveness of the Author Profiling system. The accuracy measure is represented in Equation (2).

$$Accuracy = \frac{TP_i + TN_i}{TP_i + FP_i + FN_i + TN_i} \quad (2)$$

In this context, Accuracy is the ratio of number of test documents correctly predicted their gender and age to the total number of documents in the test corpus. The Equation (3) shows the accuracy measure

$$Accuracy = \frac{Number\ of\ test\ documents\ predicted\ their\ profiles\ correctly}{Total\ number\ of\ documents} \quad (3)$$

6.2. Results of word embeddings based approach using machine learning algorithms

The experimental results of proposed word embeddings based approach for author profiling using machine learning algorithms are displayed in Table 3.

Table 3: The gender and age accuracies of proposed approach using machine learning algorithms

	Gender Prediction Accuracy		Age Prediction Accuracy	
	SVM	RF	SVM	RF
Word2Vec	92.1	93.1	54.4	85.3
Doc2Vec	80.4	87.4	63.7	77.4

The Word2Vec embeddings based document vectors attained gender accuracies of 92.1 and 93.1 for SVM and RF classifiers respectively, age accuracies of 54.4 and 85.3 for SVM and RF classifiers respectively. The Doc2Vec embeddings based document vectors attained gender accuracies of 92.1 and 93.1 for SVM and RF classifiers respectively, age accuracies of 54.4 and 85.3 for SVM and RF classifiers respectively. The combination of Word2Vec with Random Forest classifier attained best accuracies for gender and age prediction when compared with Doc2Vec and support vector classifier accuracies.

6.3. Results of word embeddings based approach using similarity measures

The experimental results of proposed word embeddings based approach for author profiling using similarity measures are displayed in Table 4.

Table 4: The gender and age accuracies of proposed approach using similarity measures

Word Embedding Technique/ Profiles/ Similarity Measure	Gender Prediction Accuracy	Age Prediction Accuracy
	Cosine Similarity Measure	Cosine Similarity Measure
Word2Vec	94%	86.15%

The Word2Vec embeddings based document vectors attained accuracies of 94% and 86.15% for gender and age prediction respectively when cosine similarity measure is used for finding the similarity among documents.

Overall, the proposed approach based on similarity measures shows best performance for gender and age prediction when compared with performance of proposed approach based on machine learning algorithms.

VII. Conclusions and Future Scope

The author profiling technique is a text processing technique which is used for predicting the hidden information of authors by analysing their written texts. In this work, we proposed two approaches for author profiling based on machine learning algorithms and similarity measures. The proposed approach based on machine learning algorithms attained best accuracies of 93.1% and 85.3% for gender and age prediction respectively. The random forest classifier shows good performance when compared with support vector machine. The proposed approach based on similarity measures attained best accuracies of 94% and 85.3% for gender and age prediction respectively. The proposed approach based on similarity measures shows good performance in age and gender prediction than proposed approach based on machine learning algorithms.

In future work, we are planning to implement other word embedding techniques like FastText, Glove and BERT for enhancing the accuracy of gender and age prediction. We are also planning to implement these proposed approach for predicting nativity language and location of authors.

REFERENCES

[1] J. Schler, Moshe Koppel, S. Argamon and J. Pennebaker (2006), Effects of Age and Gender on Blogging, in Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, March 2006. Vol. 6, (2006), 199-205.

[2] M. SudheepElayidom , Chinchu Jose , AnittaPuthussery ,Neenu K Sasi "Text classification for authorship attribution analysis", Advanced Computing: An International Journal (ACIJ), Vol.4, No.5, September 2013.

[3] Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. J. Mach. Learn. Res. 8, 1261–1276 (Dec 2007).

[4] Koppel M. S. Argamon and A. Shimoni, Automatically categorizing written texts by author gender, Literary and Linguistic Computing, pages 401-412, 2003.

[5] Nerbonne, J., *The secret life of pronouns. What our words say about us*. 2013, ALLC.

[6] Christopher D. Manning, PrabhakarRaghavan, and HinrichSch • utze. Introduction to Information Retrieval. Cambridge University Press, 2008.

[7] 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).

[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] RishabhKatna, KashishKalsi, Srajika Gupta, Divakar Yadav, Arun Kumar Yadav, Machine learning based approaches for age and gender prediction from tweets, Multimedia Tools and ApplicationsVolume 81, Issue 19, 01 August 2022, pp 27799–27817, https://doi.org/10.1007/s11042-022-12920-1

[12] Lam, L., &Suen, C. Y. (1997). Application of majority voting to pattern recognition: An analysis of its behavior and performance. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 27 (5), 553–568.

[13] Breiman, L. (2001). Random forests. Machine Learning, 45(1),5–32.

[14] Cortes, C. &Vapnik, V. Machine Learning (1995) 20: 273.

[15] 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.

[16] MohebRamzyGirgis, Abdelmgeid Amin Aly& Fatima MohyEldinAzzam, "The Effect Of Similarity Measures On GeneticAlgorithm-Based Information Retrieval", International Journal of Computer Science Engineering and Information Technology Research, Vol. 4, Issue 5, pp. 91-100, Oct 2014.

[17] 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.

[18] Chennam Chandrika Surya, Karunakar K, Murali Mohan T, R Prasanthi Kumari, "Language Variety Prediction using Word Embeddings and Machine Leaning Algorithms", Journal For Research in Applied Science and Engineering Technology, https://doi.org/10.22214/ijraset.2022.48280.

[19] 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.