Journal of Information Systems Engineering and Management
2025, 10(11s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Novel Document Representation Method for Author Profiling using Auto-Encoders

Karunakar Kavuri[1], Kavitha M[2]

[1]*Research Scholar, Department of CSE, VelTech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Tamilnadu, India.*
karunakar.mtech@gmail.com
[2]*Professor, Department of CSE, VelTech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Tamilnadu, India.*
kavitha@veltech.edu.in

ARTICLE INFO

Received: 09 Nov 2024

Revised: 29 Dec 2024

Accepted: 10 Jan 2025

ABSTRACT

Author profiling is used for identifying certain demographic characteristics including age, gender, religion, language, nationality, and others of an author for a certain text. Author profiling has applications in different areas such as marketing, security, education, and forensics. Most of the research works concentrated on predicting the age and gender of an author by analyzing the writings of authors. The researchers started research on author profiling by using different types of stylistic features. Later, they realized that same set of stylistic features are used by the different classes of age and gender. Then, the research community observed that the writing styles of different authors can best differentiate by using the content-based features such as words that are utilized in the writing of a text. Some set of successful research works are used feature selection methods to identify the best relevant words from the datasets, and some other set of works are used term weight measures to denote the term importance within a document to represent the documents as vectors. The performance of author profiling approaches mainly depends on the type of information is used for representing the documents as vectors. In this work, we developed a novel document representation method by using different types of information in document representation. In the proposed method, the document representation considers three varieties of information such as the compressed feature representation of document identified by the auto-encoder, the contextualized information of word embeddings, and the importance of a word within a document to represent the documents as vectors. In this work, we conducted experiment on two standard datasets such as reviews dataset and Twitter dataset those are provided in PAN 2014 competition and PAN 2016 competitions respectively. In these datasets, the dataset pertaining to gender consists of two classes of documents such as female and male, and the dataset pertaining to age consists of five classes of documents such as 18-24, 25-34, 35-49, 50-64, and 65-xx. The proposed document representation method shows best performance for age and gender prediction on two datasets when compared with several popular methods of author profiling.

Keywords: GAN; Auto-encoders; BERT Embeddings; Document Representation; Author Profiling; Age Prediction; Gender Prediction.

I. INTRODUCTION

Author Profiling (AP) is a form of text classification technique that uses an analysis of the author style of writings to predict the profiling features of the writers, such as their age, location, gender, occupation, educational background, nativity language, and personality qualities [1]. In general, every person has their own writing style, which doesn't vary over the course of their career regardless of whether they are writing a post, document, blog, or a review. In the current information age, author profiling is a crucial approach that is utilized in a variety of applications, including forensic investigation, the educational sector, and marketing [1].

Social media websites plays a crucial role in our daily lives and are a source of text-based crimes such as public humiliation, defamation, fake profiles, extortion, and stalking. The study of writing styles, documents, signatures, and anonymous communications is known as forensics. In this regard, author profiling methods are helpful in forensic analysis and crime investigation to pinpoint the offender by examining the characteristics of text.

The key factor for influencing author profiling methodologies is the authors' writing style. To differentiate between the writing styles of male and female authors, the content-based features are more helpful. Words like pink, my husband, and boyfriend are more likely to be used in texts produced by female, while words like the World Cup and cricket are more likely to be used in texts written by male [2]. Men use more articles, prepositions, longer words, and numbers than women [3]. From the literature, the researchers learned about different age groups of authors and how their writing styles are reflected in the writing of a text [4, 5].

Most of the research works in author profiling used the content based features like words for differentiating the author's style of writings. Based on the analysis of various research works, the major issues that are influencing the performance of author profiling approaches are

I1. Imbalance of classes in datasets
I2. Best representation of words as word vectors
I3. Reducing the number of dimensions in word vector representation
I4. Identify the best set of features for document representation
I5. Utilization of contextualized information of words
I6. Consideration of importance of a word in a document
I7. Best combination of features for document vector representation
I8. Identification of best classification algorithms

In general, the textual information in a document is unstructured or semi-structured. The documents are denoted as feature vectors by using thousands of dimensions to train with machine learning methods. More number of dimensions in document representation causes more complex in execution of classification algorithm, inappropriate information was extracted for document representation, and poor performance of classification algorithm. So, there is a need of reducing the number of features in document vector representation for enhancing the performance and accuracy of text classification. The representation of document plays a crucial role to improve the performance of author profiles prediction. In this work, we proposed a new document representation method to address the above mentioned issues by using the concepts of Generative Adversarial Networks (GAN) and Auto-Encoders (AE). In the proposed approach, we used GAN for balancing the data in different classes of dataset. The auto-encoders are used for identification of best dimensions in word vector representation and recognition of best features from the document vector representation. The BERT model is used for representing the words as vectors. Every document is represented with three types of information such as the compressed representation of auto-encoder, the aggregated information of word vectors that are contained in that document, and the term significance within a document. The vectors of documents are trained with two classification algorithms such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost). The experiments are implemented on two author profiling datasets which are provided in PAN 2014 and PAN 2016 competitions.

This paper is structured in 7 sections. The section 2 explained about various latest research works proposed for author profiling. The section 3 presents the characteristics of two datasets that are considered in this experimental work. Section 4 explains the proposed document representation method and the concepts that are used in the proposed method. The experimentation results are presented and explained in section 5. The analysis of results is presented in section 6. The conclusion of this work and future enhancements to this work are explained in section 7.

II. LITERATURE SURVEY

Author profiling is a popular research field in which several computational methods were proposed for predicting the profiles of authors from their written texts. PAN competition started the author profiling task in 2013 and continued the task in further years. Initial times, researchers used stylistic features to differentiate the writing style of authors. Rishabh Katna et al., experimented [6] with different machine-learning methods and applied various natural language processing techniques such as Tokenization, lemmatization, char and word-N-Grams for author profiling. The proposed method accomplished gender prediction accuracies of 0.880, 0.632, 0.798, and 0.812 for SVM, DT, RF, and LR classifiers respectively, and attained age prediction accuracies of 81.0%, 53.7%, 68.1%, and 72.5% for SVM, DT, RF, and LR algorithms respectively. It is possible to estimate an author's age by looking at the texts they have written. Asogwa D.C et al., proposed [7] a model by using three varieties of features such as style-based, content-based and topic-based features to determine the age of an author from their written text. They implemented NB classifier for generating the trained model and attained an accuracy of 80% for age prediction.

Later, authors are used different types of features and trained with machine learning algorithms and combination of ML algorithms and deep learning techniques. Yutong Sun et al., focused [8] on the technique of author profiling published in the competition of FIRE ("Forum for Information Retrieval Evaluation") 2019. In order to detect the labels, they employed character and word-based TFIDF features and trained with the classifier of logistic regression. According to experimental results, the proposed system expresses best performance in case of age prediction and obtained an age prediction accuracy of 0.6250. Furthermore, they attained 0.9604 and 0.5111 for language variety and gender prediction respectively. The authors observed that the performance of system was greatly improved and prediction accuracy was increased with the combination of character and word based features.

Some researchers observed that the neural networks and ML algorithms shows good performance based on the dataset of author profiling. Daniel Dichiu et al., developed [9] a method for the PAN 2016 Author Profiling Task of gender and age prediction of users. They used neural networks and SVM classifiers on verbosity and TF-IDF features. The findings indicated that neural networks outperform SVM classifiers on datasets of Dutch and Spanish, whereas SVM classifiers perform better on English datasets. The proposed method yielded accuracy values of about 0.8 during the training's cross-validation phase, far below the test dataset's accuracy scores.

Researchers used word embedding techniques for representing words as features and observed that the word embedding techniques based word vectors attained good accuracies for author profiling. Roy Bayot et al., proposed [10] a method and compared with standard approaches that preprocess text, extract features, and employ those features in Support Vector Machines (SVMs) using cross-validation. The primary distinction is that the used features are taken from averages of word embeddings, particularly word2vec vectors. The proposed approach used the embeddings of Word2vec in combination with Support Vector Machines. They were able to obtain 68.2% and 44.8% for English gender and age classification respectively by using the PAN 2016 dataset. They were also able to obtain 67.1% and 51.3% for Spanish gender and age classification respectively. Lastly, they reported an accuracy of 71.9% for Dutch age classification. Roberto Lopez-Santillan et al., proposed [11] a method by combining the Centroids Method with Word Embeddings (WE) to generate Document Embeddings (DE) that outperform other approaches in author gender prediction over a dataset of text posts from Twitter. In particular, the proposed approach achieves 0.78% accuracy for English language users in the testing dataset, and an average accuracy score of 0.77 for users of Arabic, Spanish, and English languages.

Researchers used different varieties of deep learning techniques to predict the characteristics of authors. Roy et al., presented [12] a method for PAN 2018 author profiling. The method mostly uses LSTMs and word vectors for gender classification. Proposed method obtained an accuracy of 68.73% for Spanish, 77.16% for English, and 67.60% for Arabic gender classification using the PAN 2018 dataset. Rick Kosse et al., developed [13] a model by using a simple feed-forward neural network. They conducted experiment on PAN 2018 datasets and concentrated on textual data only. The proposed best-performing system used unigrams only as features, but neural networks have been used in conjunction with word embeddings in earlier works. Proposed model received scores of 0.792, 0.792, and 0.807 for Spanish, Arabic, and English respectively on the PAN 2018 test set. They determined that with an average score of 0.797, proposed model is fairly robust among all three languages.

Massive training data sets usually produce superior neural network performance. Initializing a network with pretrained layers is beneficial when there is limited access to training data. Pretraining recurrent networks for NLP tasks is challenging because networks are typically only given pretrained word embeddings for NLP tasks. Maximilian Bryan et al., presented [14] a siamese architecture on textual data for pretraining of recurrent networks. Sentence pairs must be mapped by the network onto a vector representation. They are evaluating using the PAN 2019 bots and gender profiling dataset. The pretrained Siamese architecture attained accuracies of 0.8597 and 0.7862 for bot detection on English and Spanish languages datasets respectively. For gender prediction, proposed work attained accuracies of 0.7494 and 0.7169 on English and Spanish languages datasets respectively.

Author profiling is the process of analyzing the written text to identify different attributes such as gender, age, personality etc., of an author. Cristian Onose et al., proposed [15] an approach to the PAN 2019 problem of bots and gender profiling. The 2019 PAN competition contains two challenges such as determining if the author is a bot or a human and the system must determine the gender of human authors. The proposed approach makes use of pretrained word embeddings for text representation along with a deep learning model built on Hierarchical Attention Networks (HAN). According to the official results, the model obtains an accuracy score of 0.8483 for Spanish and 0.8943 for English on the first challenge and the model achieves 0.6711 of accuracy for Spanish and 0.7485 of accuracy for English in the second challenge.

Andrea Cimino et al., involved [16] in the PAN@CLEF2019 shared task on bots and gender profiling for the English language. They developed three methods and tested based on three distinct classification algorithms. The first method uses an SVM classifier with manually constructed features that were extracted from a large corpus of language data. The second and third methods take advantage of the latest developments in natural language processing by utilizing word embeddings learned from Twitter and a Neural Network of Hierarchical GRU-LSTM, and lastly modifying the BERT system. Authors submitted the last run of the Hierarchical Neural Network model after an internal assessment, and it finished with a final accuracy of 0.9083 for the task of Bots Profiling and a score of 0.7898 for the task of Gender Profiling.

The goal of author profiling task is correlating the author's writing style with author's demographics. Roobaea Alroobaeaet al., developed [11] an approach to predict the gender and age from feeds of Twitter by implementing a DSS ("Decision Support System"). The DSS was more efficient for determining the gender and age of authors from their tweets. The proposed system adopted deep learning methods of Deep Neural Network (DNN), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) models, and machine learning techniques. They observed that the attained results outperform those attained in the CLEF 2019.

The author profiling techniques are used in various languages for predicting the author characteristics. Joao Pedro Moreira de Morais et al., proposed [18] a cascading approach and evaluated this approach by combining a classifier, a heuristic and weighted lexical approach for the problem of gender prediction in dataset of Portuguese language. Rahman, M.A. et al., explained [19] the process of developing a multilingual classification system for author profiling. The experiment carried out on Twitter corpus of several languages such as Spanish, Italian, Dutch, and English. The SVM classifier was used for developing language specific models to determine the gender and age of authors. The used scheme of 3-fold-cross-validation attained an overall highest gender prediction accuracy of 81.3% for Spanish language and highest age prediction accuracy of 81.3% for English language.

III. CHARACTERISTICS OF DATASET

In this work, the experiment was implemented on two different datasets such as PAN 2016 competition AP dataset of English language and PAN 2014 competition AP reviews dataset of English language for determining the gender and age of authors. The description about PAN 2014 competition dataset [20] is denoted in Table I.

TABLE I. DETAILS OF PAN 2014 COMPETITION REVIEWS DATASET

Classes	Sub-classes	Number of Authors	Total Number of Authors
Gender	*Female*	2080	4160
	Male	2080	
Age	*18-24*	360	4160
	25-34	1000	
	35-49	1000	
	50-64	1000	
	65-xx	800	

The description pertaining to the PAN 2016 competition author profiling dataset [21] is displayed in Table II.

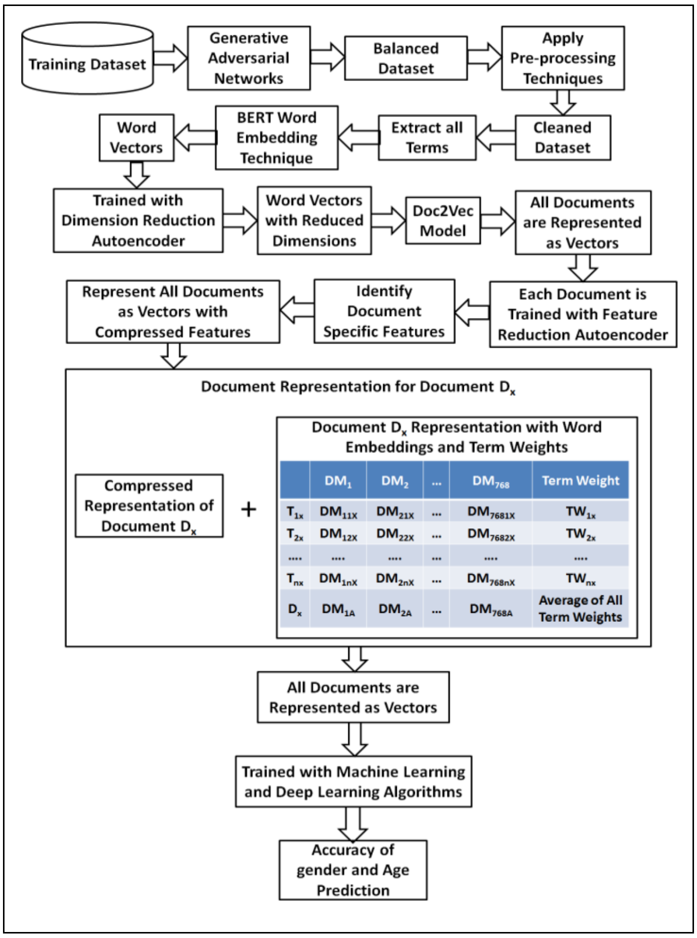
TABLE II. THE DESCRIPTION OF PAN 2016 COMPETITION AUTHOR PROFILING DATASET

Classes	Sub-classes	Number of Authors	Number of Tweets
Gender	*Male*	218	149059
	Female	218	113972
Age	*18-24*	28	363031

	25-34	140	
	35-49	182	
	50-64	80	
	65-xx	6	

IV. PROPOSED DOCUMENT REPRESENTATION METHOD

Fig. 1 shows the steps followed in proposed document representation method for author profiling.


Training Dataset
Generative Adversarial Networks
Balanced Dataset
Apply Pre-processing Techniques
Cleaned Dataset
Extract all Terms
BERT Word Embedding Technique
Word Vectors
Trained with Dimension Reduction Autoencoder
Word Vectors with Reduced Dimensions
Doc2Vec Model
All Documents are Represented as Vectors
Each Document is Trained with Feature Reduction Autoencoder
Identify Document Specific Features
Represent All Documents as Vectors with Compressed Features
Document Representation for Document D_x
Compressed Representation of Document D_x
+
Document D_x Representation with Word Embeddings and Term Weights
All Documents are Represented as Vectors
Trained with Machine Learning and Deep Learning Algorithms
Accuracy of gender and Age Prediction

	DM_1	DM_2	...	DM_{768}	Term Weight
T_{1x}	DM_{11X}	DM_{21X}	...	DM_{7681X}	TW_{1x}
T_{2x}	DM_{12X}	DM_{22X}	...	DM_{7682X}	TW_{2x}
....	....	....	...	....	....
T_{nx}	DM_{1nX}	DM_{2nX}	...	DM_{768nX}	TW_{nx}
D_x	DM_{1A}	DM_{2A}	...	DM_{768A}	Average of All Term Weights

Fig. 1. The Proposed Document Representation Method for Age and Gender Prediction

In this paper, we developed a new document representation technique for gender and age prediction by using Generative Adversarial Networks and Auto-encoders. In the proposed technique, collect the dataset for age and gender prediction. The classes in the dataset are balanced by using the technique of generative adversarial networks. Once the dataset is balanced, then apply appropriate techniques of pre-processing like tokenization, lowercase conversion, stop-words removal, and lemmatization to arrange the dataset for next-level analysis. Extract all informative terms from the cleaned dataset. All extracted terms are represented as word vectors by using BERT word embedding technique. We used small case BERT embedding technique to represent each term as 768 dimensional word embedding vector. We observed that all 768 dimensions are not most useful to represent the word vector. To reduce the number of dimensions or to identify best informative dimensions from these 768 dimensions, we proposed a Dimension Reduction Auto-Encoder (DRAE). The DRAE represent each word vector with compressed word vector representation. These compressed word vector representations are used to represent the documents as vectors by using Doc2Vec model. In Doc2Vec model, every document is denoted as a vector by aggregating the compressed word vectors of words that are contained in that document.

These document vectors are trained with proposed Feature Reduction Auto-Encoder (FRAE) to identify the best features from the vector representation or to identify the compressed representation of document. Each document is trained with a different FRAE to identify a different set of compressed features. These compressed features are utilized for denoting the document vector. Each document is represented with a different set of compressed representations.

Each document representation contains three parts of information. Part1 contains the compressed features identified through the Feature Reduction Auto-Encoder. Part2 contains the aggregation of word vectors of words that are contained in the document. Part3 contains the information of average of term weights of terms those are contained in that document. The term weights are calculated by using a term weight measure that was proposed in our previous work [22]. All documents are denoted as vectors by merging the three parts of information. These vectors are used to train various machine learning algorithms. The generated model of these algorithms produces the accuracy of proposed document representation method for age and gender prediction.

In Figure 1, D_x represents the x^{th} document, $\{T_{1x}, T_{2x},...., T_{nx}\}$ is a set of terms in the document D_x, $\{DM_1, DM_2, ..., DM_{768}\}$ is a set of dimensions in BERT word embedding vector, DM_{1nx} is dimension DM_1 value of n^{th} term T_n in document D_x, DM_{1A} is the aggregated value of all terms dimension 1 (DM_1) values, TW_{nx} is the weight of a n^{th} term T_n in document D_x. The next subsections explain the techniques used in the proposed method.

A. Generative Adversarial Networks (GANs)

In the training of deep learning networks, the amount of data is as significant as the quality because the training with smaller dataset generally degrades in performance results of neural network or introduce a problem of an over-fitting. It was a more difficult problem to learn from smaller datasets and becomes more challenging task when it is augmented with different complex problems such as imbalance of data, similarities among classes, and differences within a class.

The issue of data imbalance is generally affecting the classification performance of neural networks. Due to the difficulty with the data imbalance, the trained algorithm assigns more priority to the classes which contain the majority of samples, leading to biased classifiers. In this work, the GAN is used for enhancing the classification performance under the situations of data imbalance.

A family of neural networks recognized as Generative Adversarial Networks (GANs) [23] which can be separated into two hostile components such as the generator and the discriminator. These two components are competitively participates in the training. The aim of generator is resembles the original data by transforming noise, whereas the aim of discriminator is determining whether the data are original or generated by the generator component. The classification outcomes of the discriminator are subsequently sent to the training of the generator.

There are numerous GAN variations proposed to enhance sample generation and training stability. These variations include the conditional GANs (CGANs) [24], and the semi-supervised GANs (SS-GANs). In this study, dataset balance is accomplished using Vanilla Generative Adversarial Networks. Equation (1) gives the min-max game's objective function among the generator network and the discriminator network.

$$\min_{G} \max_{D} E_{x \square \mathrm{P_r}(x)}\left[\log D(x)\right] + E_{z \square P_z(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (1)$$

Where, x is original data sampled from the distribution of real data ($P_r(x)$). The generator G creates a synthetic image G(z), and $P_r(x)$, z is the vector of noise which is sampled from a uniform distribution ($P_z(z)$). GANs are able to produce realistic instances from a generative model and approximate the real input data distribution.

B. BERT Embeddings

A more recent development in natural language processing is deep contextual word embeddings which produces word embeddings at the time of pre-processing of data. ELMo and BERT are two common examples of these types of algorithms [25]. The Book Corpus, which has more than ten thousand books in a variety of genres, was used to train the BERT model. Deep contextual word embeddings outperform context-free counterparts (like fastText) in many tasks, but there is a significant trade-off because there is a need of more number of trainable parameters which results longer training durations.

Based on the layers quantity of transformer it comprises, BERT is presently developed in two models such as base-BERT and large-BERT. Base-BERT contains transformers of 12 with parameters of 110 million, whereas large-BERT contains transformers of 24 with parameters of 340 million. The Attention mechanism was implemented with the aid of the transformer layers. Each transformer layer in the base BERT and large BERT consists of 768 and 1024 hidden units respectively. Additionally, they each have 12 and 16 attention heads respectively. We explored with base BERT in this work. Base BERT model represents each word as a 768-dimensional word embedding vector.

C. Auto-Encoder (AE)

In machine learning algorithms based approaches, dimensionality reduction is one of the most significant techniques to improve the fastness of processing in case of data contains huge number of dimensions. Traditionally different feature selection algorithms are used for reducing the number of features from the dataset. In recent times, neural network based methods like auto-encoders are used for dimensionality reduction. Auto-encoders are applied on different varieties of data like images and texts.

First time, the auto-encoder was introduced in [26]. Successively, bourlard et al., expanded auto-encoders in detail in [27]. Auto-encoder is a kind of neural network which is used for generating a compressed representation of original data. Auto-encoder contains two sub-models such as encoder and decoder. The functionality of encoder is compression of the input and decoder functionality is reconstructing the input from the compressed version produced by the encoder. After process of auto-encoder training, save the encoder model and discard the decoder model. Then, the encoder model is used as a technique of data preparation for extracting features from the original data and these features are used to train various machine learning algorithms.

The architecture of an auto-encoder is purposefully restricting to a compressed layer at the model's midpoint, from which perform the input data reconstruction. In general, AE consists of at least three layers such as one input, one hidden, and one output layer. The purpose of the output layer is reconstruction of input layer data. After completion of training, the hidden layer is considered as a unit for dimensionality reduction of features. In this work, we used two different auto-encoders such as Dimension Reduction Auto-Encoder (DRAE) and Feature Reduction Auto-Encoder (FRAE) for different purposes.

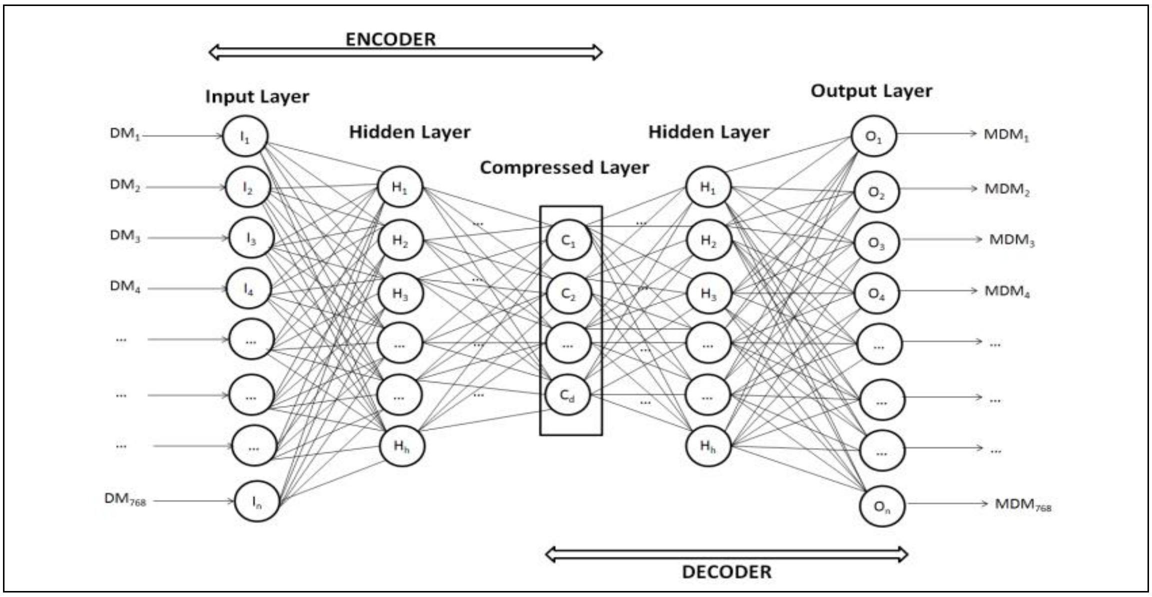


Fig. 2. Dimension Reduction Auto-Encoder (DRAE)

In Fig. 2, {DM_1, DM_2, ..., DM_{768}} is a set of dimension values of a particular BERT word embedding vector, {I_1, I_2, ..., I_n} is set of nodes in input layer, {H_1, H_2, ..., H_h} is set of neurons in Hidden layer, {C_1, C_2, ..., C_d} is set of compressed features identified by the auto-encoder, {O_1, O_2, ..., O_n} is a set of nodes in output layer, {MDM_1, MDM_2, ..., MDM_{768}} is set of reconstructed dimension values of a particular BERT word embedding vector.

Fig. 2 shows the structure of an auto-encoder for reducing the number of dimensions in word vector representation. This auto-encoder takes the input as the dimension values of a BERT word embedding vector. All the words embedding vectors are trained with this auto-encoder to reduce the number of dimensions of each word vector. The input layer contains 768 input units because each word is represented with 768 dimensional vectors. The DRAE is trained with different combinations of hidden layers, input layer and output layer.

Equation (2) is used to compute the minimization of reconstruction error in DRAE.

$$\arg\min\sum_{i=1}^{n}\left\|D_i - MD_i\right\|^2 \qquad (2)$$

Based on the reconstruction error values, we observed that the combination of input layer, compressed layer, output layer, and 0 hidden layer attained less reconstruction error.

The number of compressed features in the compressed layer is also one important task to identify the best patterns from the data. The DRAE is trained with different combinations of neurons in compressed layer and the compressed error of each combination is noted. According to reconstruction error values, we observed that the combination of 400 neurons in compressed layer attained less reconstruction error. Now, every word is represented with 400 dimensions. These 400 dimensional word vectors are passed to Doc2Vec model for representing each document as vector. These document vectors are trained with Feature Reduction Auto-Encoder (FRAE) to reduce the features count in the document vector representation.

Fig. 3 shows the structure of an auto-encoder for reducing the number of features in document vector representation. This auto-encoder takes the input as the document vector values of document vector representation. All the document vectors are trained with this feature reduction auto-encoder to reduce the number of features of each document vector. The input layer contains 400 input units because each document is represented with 400 features. The hidden layers count in the FRAE influences the efficiency of classification. The FRAE is trained with different combinations of hidden layers, input and output layers.

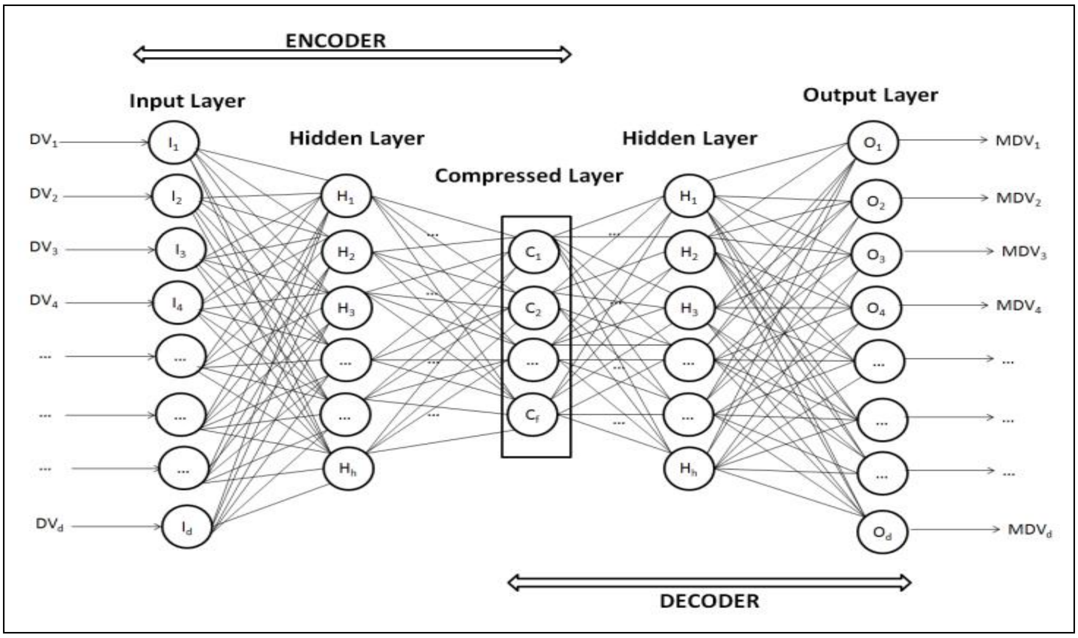


Fig. 3. Feature Reduction Auto-Encoder

In Fig. 3, {DV$_1$, DV$_2$, ..., DV$_d$} is a set of feature values in a particular document vector representation, {I$_1$, I$_2$, ..., I$_d$} is set of nodes in input layer, {H$_1$, H$_2$, ..., H$_h$} is set of neurons in Hidden layer, {C$_1$, C$_2$, ..., C$_f$} is set of compressed features identified by the feature reduction auto-encoder, {O$_1$, O$_2$, ..., O$_d$} is a set of nodes in output layer, {MDV$_1$, MDV$_2$, ..., MDV$_d$} is set of reconstructed document vector values of a particular document vector.

Equation (3) is used to compute the minimization of reconstruction error in FRAE.

$$\arg\min\sum_{i=1}^{n}\left\|DV_i - MDV_i\right\|^2 \qquad (3)$$

According to reconstruction error values, we observed that the combination of input layer, compressed layer, output layer, and 1 hidden layer attained less reconstruction error. The number of compressed features in the compressed layer is also one important task to identify the best patterns from the data. The FRAE is trained with different combinations of neurons in compressed layer and the compressed error of each combination is noted. We observed that the combination of 300 neurons in compressed layer attained less reconstruction error. Now, every document in the dataset is represented with 300 features identified in the compressed layer. These 300 feature document vectors are passed to different machine learning algorithms to predict the prediction accuracies of age and gender.

D. Term Weight Measure (TWM)

The word significance in a document is determined by the TWM. In this work, we used a TWM that was proposed in our previous work [22], which is represented in (4).

$$PTWM(T_i, D_k \in C_j) = \frac{TF\left(T_i, D_k\right)}{TNTD_k} * \frac{TF(T_i, D_k \in C_j)}{TF(T_i, D_k \notin C_j)} * \frac{A+D}{B+C} * \left(\frac{A}{A+B} - \frac{C}{C+D}\right) \qquad (4)$$

Where, A and C represent the count of documents in documents of class C_j and documents of other than class C_j respectively that contain the T_i term. The count of documents in documents of class C_j and documents of other than class C_j that do not contain the T_i term are B and D respectively. $TNTD_k$ is the total count of terms in D_k document, and $TF(T_i, D_k)$ is the count of T_i term in D_k document. The frequency of T_i in documents of class Cj is given by $TF(T_i, D_k \in C_j)$. $TF(T_i, D_k \in C_j)$ is the frequency of term T_i in documents that do not belong to class Cj.

E. Machine Learning (ML) Algorithms

In this work, we used two ML algorithms such as Random Forest [28] and XGBoost [29] for predicting the performance of gender and age prediction.

a) Random Forest (RF)

One of the non-parametric learning methods used in data mining is the Decision Tree (DT) classifier, which can be used with numerical, categorical or a combination of the two types of data. To map instances to the appropriate class labels, DT builds a classification model based on a tree structure. Each internal node in a DT denotes a feature that is used to take decision about the specified instance. The arcs are used to connect internal nodes to one another and internal nodes to leaves, which are referred to as feature outcomes. The class label is represented by each leaf in the tree. Based on the path that is satisfied or reached from the root to a leaf in the tree, the class label of a previously unknown sample is identified, and the corresponding leaf class label is applied to the new instance.

The primary step in the construction of a decision tree is which features are considered as root node and internal nodes in which path. Once features are decided for nodes, the next step is finding the number of splits for each feature in the node. The number of splits decides number of paths from that node. In general, top-down greedy strategy is followed in the development of decision trees. Various measures are used by different DT algorithms to find the important features for nodes in the decision tree, for instance, Chi-squared measure is used in CHAID, Gini Index is used in CART and information gain is used in C4.5.

Every DT model looks like a group of “if-then-else decision rules”, which made it simple to interpret and understand. Additionally, DT is capable of handling both numerical data-type and categorical data-type. The scalability of DT is good when the dataset contains large amount of data and it is also robust for noise data. However, the DT performance is not good when there are complex relations among features, because it maintains only one feature in every internal node. For the purpose of lowering variance, Breiman proposed [28] the classification method of RF by averaging the outcomes of different DTs. RF is an ensemble-based method for both classification and regression applications. It works by creating numerous diverse DTs on a subset of its features and the data points, reducing the likelihood that the data would be overfit.

b) Extreme Gradient Boosting (XGBoost) classifier

Due to their great performance in high accuracy and high speed, tree boosting algorithms have recently received increased attention from researchers. One of the most latest tree boosting techniques is XGBoost, which may be utilised by downloading the XGBoost library. It is popular, scalable, effective, powerful, and extremely robust. The XGBoost method [29] is a supervised learning algorithm which was proposed by Tianqi Chen et al. Additionally, XGBoost has recently won every machine learning competition, which encourages us to check into XGBoost. The foundation of XGBoost's operation is a gradient boosting method that uses a similar tree structure those other tree boosting techniques follows. The fundamental workings and characteristics of XGBoost is superior when compared with other tree boosting algorithms such as Adaboost and GBM (Gradient Boosting Machine). The condition is represented as the head node in the tree structure, and the process is divided into branches according to this condition. Depending on the tree length selected in the model, the tree descends farther. Depending on where the tree terminates, the node becomes a leaf node or a final decision node.

In a classification model, XGBoost uses labelled data as input for supervised learning, and then predicts the target by looking at a few parameters. The most significant tasks in this technique are extracting features and parameters from the data. One of the major benefits of the XGBoost method is the ability to manipulate the classifier and apply numerous parameters that are used to improve learnability or accuracy.

V. EXPERIMENTAL RESULTS

The performance of ML methods is denoted by using different evaluation measures such as recall, precision, F1-score and accuracy. Recall, Precision, Accuracy and F1-Score are represented in (5), (6), (7), and (8) respectively.

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \qquad (5)$$

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FP} \qquad (6)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad (7)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (8)$$

Where, TP and TN denotes the documents count in the positive class and negative class are assessed as Positive Class (PC) and Negative Class (NC) respectively, FP and FN denotes the documents count in the PC and NC are assessed as Negative Class, and PC respectively. .

A. *Experimental Results of Gender Prediction*

The experiments conducted on the dataset of PAN 2014 competition reviews and PAN 2016 competition datasets for determining the different performance measures of gender and age prediction.

a) Experiment results of gender prediction on dataset of PAN 2014 competition reviews

The Table III shows the performance of gender prediction on PAN 2014 competition reviews dataset when the model trained with two ML algorithms such as RF and XGBoost.

TABLE III. THE PERFORMANCE OF GENDER PREDICTION ON DATASET OF PAN 2014 COMPETITION 2014 REVIEWS

ML Techniques	Gender Prediction			
	Precision	***Recall***	***F1-Score***	***Accuracy***
RF	0.9601	0.9379	0.9489	0.9483
XGBoost	0.9678	0.9518	0.9597	0.9594

In Table III, the XGBoost machine learning algorithm attained best precision, recall, F1-score and accuracies of 0.9678, 0.9518, 0.9597, and 0.9594 respectively for gender prediction on PAN competition 2014 reviews dataset when compared with other techniques.

b) Experimental results of gender prediction on PAN 2016 competition Twitter dataset

The Table IV shows the performance of gender prediction on PAN 2016 competition Twitter dataset when the model trained with two different ML algorithms such as RF and XGBoost.

TABLE IV. THE PRECISION, RECALL, F1-SCORE, AND ACCURACIES OF GENDER PREDICTION ON PAN COMPETITION 2016 AUTHOR PROFILING TWITTER DATASET

ML Techniques	Gender Prediction			
	Precision	***Recall***	***F1-Score***	***Accuracy***
RF	0.9311	0.9062	0.9185	0.9174

XGBoost	0.9128	0.9431	0.9277	0.9289

In Table IV, the XGBoost machine learning algorithm attained best precision, recall, F1-score and accuracies of 0.9128, 0.9431, 0.9277, and 0.9289 for gender prediction on PAN competition 2016 Author Profiling Twitter dataset when compared with RF technique.

B. Experimental Results for Age Prediction

The experiments performed on two standard datasets such as PAN 2014 and PAN 2016 competition AP datasets for age prediction.

a) Experimental Results of age prediction on dataset of PAN 2014 competition reviews

The age profile dataset of PAN 2014 contains imbalance in number of documents in different classes of age profile. In this work, GAN network is used to balance the data in classes of dataset.

The Table V shows the performance of age prediction on PAN 2014 competition reviews dataset when the model trained with two ML algorithms such as XGBoost and RF.

TABLE V. THE PRECISION, RECALL, F1-SCORE, AND ACCURACIES OF AGE PREDICTION ON PAN 2014 REVIEWS DATASET

ML Algorithms	**Age Prediction**			
	Precision	***Recall***	***F1-Score***	***Accuracy***
RF	0.8824	0.8851	0.8837	0.8824
XGBoost	0.8948	0.8976	0.8961	0.8948

In Table V, the XGBoost machine learning algorithm attained best precision, recall, F1-score and accuracies of 0.8948, 0.8976, 0.8961, and 0.8948 respectively for age prediction on PAN competition 2014 reviews dataset when compared with RF technique.

b) Experimental results of Age prediction on PAN 2016 competition Twitter dataset

The age dataset of PAN 2016 competition contains more imbalances of data in different classes. GAN is used for balancing the data in dataset. The Table VI shows the performance of age prediction on PAN 2016 competition AP Twitter dataset when the model trained with two ML algorithms such as RF and XGBoost.

TABLE VI. THE PRECISION, RECALL, F1-SCORE, AND ACCURACIES OF AGE PREDICTION ON PAN 2016 COMPETITION TWITTER DATASET

ML Algorithms	**Age Prediction**			
	Precision	***Recall***	***F1-Score***	***Accuracy***
RF	0.8659	0.8746	0.8702	0.8659
XGBoost	0.8868	0.8935	0.8901	0.8868

In Table VI, the XGBoost machine learning algorithm attained best precision, recall, F1-score and accuracies of 0.8868, 0.8935, 0.8901, and 0.8868 for age prediction on PAN competition 2016 Author Profiling Twitter dataset when compared with other techniques.

VI. ANALYSIS OF EXPERIMENTAL RESULTS

The proposed document representation method shows best performance in terms of four evaluation measures such as precision, recall, F1-score and accuracy for gender and age prediction. The main reason of getting good results for age and gender prediction is consideration of important information for document vector representation. We developed the proposed method by addressing the issues specified in introduction of this work.

I1. Imbalance of classes in datasets

The imbalances of classes in datasets are solved in the proposed method by using the powerful data augmentation technique of GAN. The PAN 2014 age dataset contains less number of documents in 18-24 and 65-xx class. Other three classes contain equal number of documents of 1000. We used GAN to augment the number of documents to 1000 in 18-24 and 65-xx classes. The balancing of classes' data attained best results when compared with age prediction results of previous approaches without balancing. Likewise, the PAN 2016 age dataset contains more imbalances in classes' data. GAN is used to augment the number of author's documents in each class. The proposed method shows good performance for age prediction in PAN 2014 dataset when compared with the performance of PAN 2016 dataset, because more number of classes are imbalance in PAN 2016 dataset.

I2. Best representation of words as word vectors

After applying pre-processing techniques on balanced dataset, the informative words are extracted from the dataset. The representation of words in the form word vectors is required to represent the documents as vectors. Different word embedding techniques such as Word2Vec, Glove, fastText, and BERT are used for representing words as vectors. We identified that the performance of BERT word embeddings is good when compared with other word embedding techniques because BERT represent words as word vectors by considering the contextualized information of words. In this work, we used BERT model for representing words as word vectors.

I3. Reducing the number of dimensions in word vector representation

The number of dimensions is used for representing word vectors is one important issue in word embedding techniques. In this work, we used small-case BERT model for representing words as vectors. This version of BERT model represents each word with 768 dimensional vector representations. We observed that all 768 dimensions that are used to represent the word vector are not most important. In the proposed document representation method, we developed a dimension reduction auto-encoder to identify the most significant dimensions in the word vector representation. The DRAE identify compressed dimensions to represent the words as vectors. The compressed representation of words is more useful to improve the performance of proposed method.

I4. Identify the best set of features for document representation

Doc2Vec model is used for representing the documents as vectors by using the compressed word vectors. Identification of best features for document vector representation is one primary task in any task related to text classification using machine learning algorithms. In the proposed method, we developed Feature Reduction Auto-Encoder (FRAE) for reducing the number of features for document vector representation. FRAE takes the input as document vectors and generates the compressed representation of documents. The proposed document representation method used this compressed representation of document to represent the document vector.

I5. Utilization of contextualized information of words

The proposed document representation method used the information of document vectors that are generated through the BERT word embeddings. The BERT embeddings represent each word by considering the contextualized information of words. Each document is represented by aggregating the BERT word embeddings of words those are contained in that document. These document vectors are merged to the compressed document vectors generated through FRAE.

I6. Consideration of importance of a word in a document

The word vectors that are generated through BERT based on the importance of a word in a total dataset and the significance of word in other external datasets. The importance of a word within a document is also one important factor to consider in the representation of document. In the proposed document representation method, we considered the weight of term within a document.

I7. Best combination of features for document vector representation

The performance of machine learning algorithms mainly depends on the set of features those are used for document vector representation. the proposed method consider three varieties of features information such as the compressed set of features identified through FRAE, the document vector features generated through BERT embeddings, and the word weight within a document for the document vector representation. These three best combinations of information for representing documents as vectors are improving the performance of age and gender prediction.

I8. Identification of best classification algorithms

Identification of best algorithms for training is also one very important task in author profiling approaches. In this work, two best algorithms such as RF and XGBoost are considered based on their successful results in different text classification based tasks.

VII. CONCLUSION AND FUTURE SCOPE

Author profiling is the method of examining the textual data for extracting the information about various characteristics of the author. It was used in both commercial and social implications. In the past, many approaches were proposed to enhance the performance of author profiling. In this work, we proposed a new document representation method for denoting documents as vectors. The proposed method consider different varieties of information such as importance of term within document, contextualized information of terms and compressed representation of document generated by auto-encoders for representing document vectors. The experiments performed on two standard datasets such as PAN 2014 reviews dataset and PAN 2016 Twitter dataset. Two machine learning algorithms such as XGBoost and RF for generating the trained model and concentrated on two profiles such as age and gender prediction. The XGBoost classifier attained best results for prediction of age and gender when compared with the results of RF. XGBoost classifier attained best results on PAN 2014 dataset than the results of PAN 2016 dataset.

In future work, we are planning to merge appropriate stylistic features to existing document vector representation. We have also plan to implement our proposed method for predicting other demographic profiles of author.

VIII. REFERENCES

[1] Koppel M, S. Argamon and A. Shimoni, “Automatically categorizing written texts by author gender”, Literary and Linguistic Computing, Vol. 17, pp. 401-412, 2003.

[2] J. Schler, Moshe Koppel, S. Argamon and J. Pennebaker, “Effects of Age and Gender on Blogging”, in: Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, Vol. 6, pp. 199-205, 2006.

[3] Newman, M.L., Groom, C.J., Handelman, L.D. and Pennebaker,J.W., “Gender differences in language use: An analysis of 14,000 text samples”, Discourse Processes, Vol. 45, No. 3, ,pp. 211-236, 2008.

[4] Pennebaker, J.W., Francis, M.E. and Booth, R.J., “Linguistic inquiry and word count: Liwc 2001”, Mahway: Lawrence Erlbaum Associates, Vol. 71, No. 2001, pp. 2001-2009, 2001.

[5] Argamon S., Koppel M., Pennebaker, J.W. and Schler J., “Mining the blogosphere: Age, gender and the varieties of selfexpression”, First Monday, Vol. 12, No. 9, 2007.

[6] Rishabh Katna, Kashish Kalsi, 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

[7] Asogwa D.C, Anigbogu S.O, Anigbogu G.N, and Efozia F.N., “Development of A Machine Learning Algorithm to Predict Author’s Age from Text”, International Journal of Research - Granthaalayah, Vol. 7(10), pp. 380-389, 2020.

[8] Yutong Sun, Hui Ning, Kaisheng Chen, Leilei Kong, Yunpeng Yang, Jiexi Wang, Haoliang Qi, “Author Profiling in Arabic Tweets: An Approach based on Multi-Classification with Word and Character Features”, Journal of E – Technology, Vol. 11, No. 2, pp. 60-63, 2020.

[9] Daniel Dichiu, Irina Rancea, “ Using Machine Learning Algorithms for Author Profiling In Social Media”, Notebook for PAN at CLEF 2016

[10] S. Rao Polamuri, L. Nalla, A. D. Madhuri, S. Kalagara, B. Subrahmanyam and P. B. L. Aparna, "Analyse The Energy Consumption by Integrating the IOT and Pattern Recognition Technique," *2024 2nd International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India, 2024, pp. 607-610, doi: 10.1109/ICDT61202.2024.10489265.

[11] I. L. Manikyamba and S. R. Polamuri, "Spectrum Sensing-Optimized Data Transformation," *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)*, Bangalore, India, 2023, pp. 1-6, doi: 10.1109/ICCAMS60113.2023.10525989

[12] Roy Khristopher Bayot Teresa Goncalves, Multilingual Author Profiling using LSTMs Notebook for PAN at CLEF 2018

[13] Rick Kosse, Youri Schuur and Guido Cnossen, Mixing Traditional Methods with Neural Networks for Gender Prediction Notebook for PAN at CLEF 2018

[14] Maximilian Bryan, J. Nathanael Philipp, Unsupervised pretraining for text classification using siamese transfer learning Notebook for PAN at CLEF 2019

[15] Cristian Onose, Claudiu-Marcel Nedelcu, Dumitru-Clementin Cercel, and Stefan Trausan-Matu, “A Hierarchical Attention Network for Bots and Gender Profiling”, Notebook for PAN at CLEF 2019

[16] Andrea Cimino and Felice dell’Orletta, A Hierarchical Neural Network Approach for Bots and Gender Profiling Notebook for PAN at CLEF 2019

[17] Roobaea Alroobaea, Sali Alafif, Shomookh Alhomidi, “A Decision Support System for Detecting Age and Gender from Twitter Feeds based on a Comparative Experiments”, International Journal of Advanced Computer Science and Applications, Vol. 11, No. 12, pp. 370-376, 2020.

[18] João Pedro Moreira de Morais, Luiz Henrique de Campos Merschmann, “A Cascade Approach for Gender Prediction from Texts in Portuguese Language”, in: WebMedia '22: Proceedings of the Brazilian Symposium on Multimedia and the Web, pp. 142–149, 2022.

[19] Rahman, M.A., Akter, Y.A., “Multi-lingual Author Profiling: Predicting Gender and Age from Tweets!”, In: Image Processing and Capsule Network, pp. 505-513, 2021.

[20] https://pan.webis.de/clef14/pan14-web/author-profiling.html

[21] https://pan.webis.de/clef16/pan16-web/author-profiling.html

[22] K. Kavuri and M. Kavitha, “A Term Weight Measure based Approach for Author Profiling”, In: International Conf. on Electronic Systems and Intelligent Computing, Chennai, India, pp. 275-280, 2022.

[23] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al., “Generative adversarial nets,” in: Proc. of the 27th International Conf. on Neural Information Processing Systems, Vol. 2, pp. 2672–2680, 2014.

[24] Mirza, M., and Osindero, S., “Conditional generative adversarial nets”, arXiv Technical Report, 2014.

[25] J. Delvin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805, 2018.

[26] M. K. B, M. S. Kumar, F. D. Shadrach, S. R. Polamuri, P. R and V. N. Pudi, "A binary Bird Swarm Optimization technique for cloud computing task scheduling and load balancing," *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, Chennai, India, 2022, pp. 1-6, doi: 10.1109/ICSES55317.2022.9914085.

[27] H. Bourlard and Y. Kamp, “Auto-association by multilayer perceptrons and singular value decomposition”, Biological Cybernetics, Vol. 59, No. 4-5, pp. 291–294, 1988.

[28] L. Breiman, “Random forests”, Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[29] K. Nagamani, T. Benarji, S. A. Nabi, C. M. V. S. Akana, S. Rao Polamuri and M. Indrasenareddy, "Deep Learning based Porosity Inversion from Seismic Attributes," *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, Tirunelveli, India, 2024, pp. 1075-1079, doi: 10.1109/ICDICI62993.2024.10810831.