International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

A New Term Representation Method for Gender and Age Prediction

Upendar Para[1], Dr M S Patel[2]
[1]Assistant Professor, Department of Computer Science and Engineering,
Keshav Memorial Institute of Technology Hyderabad, Telangana,india
Research Scholar, Sapthagiri College of Engineering, Research Centre,under VTU University, Bengaluru
e-mail: upendar.para@gmail.com
[2]Professor & Head of the Department,Department of Computer Science and Engineering,
Amrutha Institute of Engineering and Management Sciences ,Bidadi ,Bangalore.
e-mail: msr_patel@yahoo.com
Corresponding Author Email: upendar.para@gmail.com

Abstract— Author Profiling is a kind of text classification method that is used for detecting the personality profiles such as age, gender, educational background, place of origin, personality traits, native language, etc., of authors by processing their written texts. Several applications like forensic analysis, security and marking are used the techniques of author profiling for finding the basic details of authors. The main problem in the domain of author profiling is preparation of suitable dataset for predicting the characteristics of authors. PAN is one organization conducting competitions on various types of shared tasks. In 2013, PAN organizers presented the task of author profiling in their series of competitions and continued this task in further years. They arranged different kinds of datasets in different varieties of languages. From 2013 onwards several researchers proposed solutions for author profiling to predict different personality features of authors by utilizing the datasets provided in PAN competitions. Researchers used different kinds of features like character based, lexical or word based, structural features, syntactic, content based, style based features for distinguishing the author's writing styles in their texts. Most of the researchers observed that the content based features like words or phrases those are used in the text are most useful for detecting the personality features of authors. In this work, the experiment conducted with the content based features like most important words or terms for predicting age group and gender from the PAN competition datasets. Two datasets such as PAN 2014 and 2016 author profiling datasets are used in this experiment. The documents of dataset are converted in to a vector representation which is a suitable format for giving training to machine learning algorithms. The term representation in a document vector plays a crucial role to improve the performance of gender and age group prediction.The Term Weight Measures (TWMs) are such techniques used for this purpose to represent the significance of a term value in document vector representation. In this work, we developed a new TWM for representing the term value in document vector representation. The proposed TWM's efficiency is compared with the efficiency of other existing TWMs. Two Machine Learning (ML) algorithms like SVM (Support Vector Machine) and RF (Random Forest) are considered in this experiment for estimating the accuracy of proposed approach. We recognized that the proposed TWM accomplished best accuracies for gender and age prediction in two PAN Datasets.

Keywords- Author Profiling, Age Group Prediction, Gender Prediction, Term Weight Measures

I. INTRODUCTION

In last decade of time, the social media environments play a significant role in our regular life where users are constantly interacting with Snapchat, Twitter, Facebook, etc., to share their opinions and experiences among other users on different topics. The availability of several datasets like blogs, public pages of Facebook and Twitter Tweets pertaining to data generated in social media platforms provides golden chances for social scientists to investigate about social and psychological questions at an extraordinary scale [1]. For example, social media content is used in different real time applications such as stock market prediction [2], measuring behavioural risk factors [3], national mood inferring throughout the day [4], health monitoring [5], box-office revenues forecasting for movies [5], disaster management [6], and Oil price prediction [7].

Alternatively, the free access of different social media environments allows any user of any age able to become a reader and an author without any formal restriction. Additionally, these platforms provided an ultimate environment to online predators for gaining access to sensitive information associated to users, which shows effect on the internet activities of several vulnerable groups like females, teenagers, kids, etc., at risk. Thus, automatic detection of personality features like gender and age groups of users from their conversations in social media would offers an advantage to prevent the crimes as well as various other activities such as content personalization, personalized advertisement, online tutoring, and plagiarism detection which identifies whether homework is

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

completed by student or other person. Author Profiling (AP) is one such technique for predicting author's personal characteristics like age and gender.

The methods for AP are classified into two classes such as Profile Based Approaches (PBA) and Instance Based Approaches (IBA). In IBA, individual document is considered for analysis. In PBA, combine the posts of single author into one document which is used for analysis. In twitter dataset analysis, the profile based approaches are used by the most of the researchers. Authorship Profiling is broadly used in several applications of text processing like education, security, analysis of forensics, and marketing [8]. In education point of view, the pupils are writing assignments and answers to questions. By analysing the writings of the pupils, the experts can identify the grammar they used, the complex words they used and presentation style of the content. Author profiling methods are used to know the exceptional talents among the pupils and educational background of the students. The terrorist groups are sending threatening posts or mails to the authorized agencies of government. Most of the time they are not specifying their correct details while sending the mails. In this context, AP techniques are helpful to know the details like location of a mail, nativity language of the author, gender, age of the authors who wrote the mail. In forensic analysis, the experts of forensic are analysing the documents, suicide notes and property wills etc. The AP techniques are used to know the basic details like age, gender, location etc. of the anonymous wills and documents. In the present digital world, people are selecting products or services based on the reviews given by the experienced customers. The reviews play a crucial role in the success of a product or service. After launching a product by the companies, they verified the reviews of the customers about their product. The reviews are helpful for the companies when the reviews are given by the customers with genuine details. If the reviews are anonymous they are not useful for the companies. To know the details of the customers who are not giving their details in the posting of reviews, the author profiling techniques are used.

The researchers identified many differences in the authors writing styles by analysing standard datasets of author profiling in their experiments. Some researchers [9, 10] found that female writings contain more prepositions usage when compared with writings of male. In the observation of [11], the writings of male authors contain words related to the topics of politics and technology whereas female author writings use more adverbs and adjectives and the words related to the topics of wedding and shopping. Koppel et. al., identified [9] that the men uses more number of determiners and quantifiers, whereas the more pronouns are used by woman in their writings. The authors also observed that the male authors argued about the topics related to politics, woman, sports and technology. In contrast to that female authors discussed the topics related to shopping, kitty parties, beauty and jewellery in their writings.

The contents of a feature and the features of a text play a crucial role in finding the author as a female or male. One researcher identified [12] that the writings of female contain the content related words like boyfriend, my husband and pink and also includes emotional words, friends, negations, words related to verbs and family. The male writings contain cricket and world cup also includes more statistics, prepositions and longer words. In the past some of the researches classified the authors based on their age groups by considering the writing styles of authors. These age groups categorized in to 3 to 4 categories. In the literature, some researchers considered the age groups of authors between 13 to 17 write about issues related to school and adolescence, the age group between 23 to 27 writes about heroes, heroines, college life, sex and premarital life, and the authors in the group of between 33 to 47 post their writing about their kids, post marriage life, social activities etc. J.W., pennebaker et. al. detected [13] that the usage of idioms, determiners, and prepositions will increase as age increase and also observed that older authors has tendency [14] to use longer words and longer posts and use more commas in their posts whereas the younger writers uses more articles, more pronouns and less nouns.

In this article, the experiment conducted with content related features like informative words. Select the informative words based on the occurrence count of a word in the total dataset. After extraction of terms for representing the documents as vectors, the value of a term in vector is determined by using TWMs. In this article, we proposed a new TWM for finding the value of a term specific to a document. The accuracy of TWM is compared with efficiencies of several well-known TWMs. The experimentation implemented with two ML techniques such as RF and SVM to build the model for classification.

This paper is ordered in 11 sections. The traditional works about author profiling are discussed in section 2. The description about two PAN Datasets are presented in section 3. The performance metrics for representing the accuracy of the proposed approach are explained in section 4. The ML algorithms are explained in section 5. TWMs based approach for AP is described in section 6. The existing TWMs are explained section 7. Section 8 discuss about proposed term weight measures and its analysis. The results of this experiment for the gender and age group prediction are discussed in section 9. The section 10 discuss about the experimental results. The section 11 concludes this paper with probable future plans.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

II. LITERATURE SURVEY OF WORKS RELATED TO AGE AND GENDER PREDICTION

The exponential evolution of social media utilization was encouraging some set of users to develop various methodologies for unknown writings, which direct to enhancement of suspicious and malicious activities. This anonymity creates difficulties in identification of suspected author. AP handles with author characteristics through some primary attributes like age, gender, variety of region, language, personality, and so on. The significant task of AP is gender identification of a suspect document's author. The user's linguistic profile helps more for determining their demographic features. Users are regularly using various social media environments such as Instagram, Facebook, Twitter, etc., for sharing their activities in day-to-day life. Additionally, some users post text messages along with images on various social media environments, thus, the utilization of multi-modal information is more common in these days. Chanchal Suman et al., proposed [15] an effective neural framework for solving the multimodal problem of gender detection from multimodal data of Twitter automatically. The proposed framework used the popular BERT base for the tweet's text part to learn the encoded representation, and EfficientNet is utilized which was introduced recently for mining image features from images part of dataset. Finally, they applied fusion strategy based on direct product for fusing the representations of image and text. Further, a fully connected layer was used for gender detection of a user. PAN 2018 dataset of author profiling task was used for estimating the efficiency of proposed framework. The proposed framework attained accuracies of89.53%, 86.22%, and 82.05% for multimodal setting, pure-text, and pure-image respectively. This framework also outperformed the previous popular research works in all cases.

Author profiling is a technique of extracting information pertaining to different personality characteristics of the author by analysing their textual format of content. It has both social and commercial implications. Several researchers proposed approaches in past for enhancing the accuracy of author's information extraction. Rishabh Katna et al., developed [16] a technique for author profiling by applying ML and NLP ("Natural Language Processing") techniques. In the proposed approach, the NLP techniques of lemmatization, Tokenization, and word and character n-grams were used in collaboration with ML techniques such as SVM, Decision Tree (DT), RF, and Logistic Regression (LR). The proposed approach attained best accuracies of 88.0%, 63.2%, 79.8%, and 81.2% for gender prediction and accuracies of 81.0%, 53.7%,68.1%, and 72.5% for prediction of age when the experiment performed with SVM, DT, RF and LR classifiers respectively.They observed that the SVM shows best performance for age group and gender prediction than other classifiers.

Presently, the research community has engaged with the author profiling techniques due to its encouraging uses in identification of fake accounts in social networks, forensic, marketing, and security. PAN competition organized competitions on different varieties of shared tasks by releasing several relevant benchmark datasets in different languages. Ameer Iqraa et al., developed [17] an approach for identification of traits like gender and age group of author from same genre of author profiles. In their proposed approach, they experimented with various kinds of features such as grouping of character n-grams, grouping of word n-grams, part-of-speech tags based traditional n-grams, and part-of-speech tags based syntactic n-grams. They tried experiment with various classification algorithms for different sizes of profiles. They considered character tri-grams and word unigrams as baseline approaches for their experiment. The proposed approach accomplished good accuracies of 0.734 and 0.496 for predicting gender and age group respectively by implementing the word n-grams combinations of dissimilar sizes. They observed that the experiment results of proposed approach indicate that the word n-grams combinations achieved best results on various benchmark datasets.

The demographic features of customers like age group and gender play a primary role in the time of data driven solution, which allows companies to improve their services and offers for attracting the right customers in right place and right time. In marketing domain, companies are trying to target GSM's (Global System for Mobile-communications) real user not the owner of line. Ibrahim Mousa Al-Zuabi et al., proposed [18] a method to predict the users age group and gender based on their contract, services and behaviour information. The authors analysed the behaviour of telecom customer by considering the data source as billing information, customer relationship management (CRM) and call detail records (CDRs). They implemented various kinds of machine learning algorithms to obtain more accurate information pertaining to demographic attributes of customers for providing marketing campaigns. The proposed model is developed by using a reliable dataset of 18000 users that was delivered by Telecom Company of SyriaTel which is used for training and test purposes. The major contribution of the proposed work is enhancement of the accuracy in terms of gender age prediction by using data of mobile phones. The model was implemented by using technology of big data and attained accuracies of 65.5% and 85.6% for age and gender prediction respectively.

Erhan Sezerer et al., proposed [19] a RNNwA model (RNN model with Attention) for predicting the gender of users in twitter by considering twitter tweets. The proposed model was enhanced by concatenating features of n-gram with the user's learned neural representation. They used LSA technique for reducing the count of n-gram features. The proposed model

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

was validated on three different languages such as Arabic, Spanish, and English. The enhanced kind of the proposed n-gram + RNNwA model shows competitive results on Arabic and Spanish and shows best performance on English dataset.

Piot–Perez-Abadin, P. et al., observed [20] that the utilization of language from both semantic and psycho-linguistic features was more interesting for detecting different aspects like origin of user, age and gender. A good combination of features set is very important for decision-making software systems, classification, and performance of retrieval. The authors addressed the classification of authors based on gender is a part of automatic profiling task.They displayed the performance of existing models for gender classification based on baselines and external corpus for automatic profiling.They analysed in deeper level about the impact of the linguistic features in the prediction of the classification models accuracy. The authors represented both feature set used for models of gender classification in social networks with the performance of accuracy above existing baselines after the analysis.

Janneke van de Loo et al., explored [21] the abilities of text based gender and age detection by analysing social media harmful content. Particularly, the authors concentrated on the use-case of determining sexual predators who are trying to groom children in online by providing false gender and age information in the profiles of users. The experiment conducted with a dataset of nearly 380000 posts of chatting from social network classification of gender and age. They compared and evaluated the trained binary age classifiers for separating older and younger authors based on different boundaries of age and observed that the macro-averaged F-scores are increased when the boundary of age was raised. Additionally, they proved that the applicable performance levels of use-case was attained for the adults versus minors classification, thereby providing a valuable component as a monitoring tool for cyber-security in moderators of social network.

Seifeddine Mechti et al., presented [22] a novel method for finding the author profiles of an anonymous text in English language. The major goal of AP is finding the demographic profiles like education level, region, gender and age, psychological profiles like mental health, personality of the authors by analysing their text, particularly authors who generated the textual content in social networks. Authors selected different machine learning algorithms for attaining good classification. Authors initially started experiment with Bayesian networks and identified that naïve Bayesian classifiers are not attained good results. Later, they proposed a method by considering advanced Bayesian networks for prediction of age group to solve the specified detailed problem. The researchers attained excellent results by experimenting on PAN 2013 author profiling corpus. They observed that their proposed method attained comparable results than the results of best approaches in that competition.

The author profiling in forensics plays a vital role in denoting probable demographic profiles of suspects. Among different automated approaches proposed recently for AP, Transfer learning techniques shows best performances than various popular techniques in NLP. Most of the traditional techniques proposed for AP largely depend on feature engineering which shows a significant difference for each developed model, whereas the transfer learning techniques generally needs a pre-processed text to feed into the model. Esam Alzahrani et al., reviewed [23] several existing works those are proposed for author profiling and identified the most appropriate pre-processing techniques that are associated with gender profiling of author. They considered different variations of possible pre-processing methods in their experiments involving five techniques and BERT model is used to estimate the effect of each technique for prediction of gender. They used a transformer library of Hugging face for implementing the code for every case of pre-processing. In their five experiments, they identified that the BERT attained good accuracy for the author's gender profile prediction when the methods of pre-processing was not applied. The best case of proposed method attained an accuracy of 86.67% for author's gender prediction.

Analysis of texts in order to identify the profiles of an anonymous author by considering their used words in the text is called AP. For instance, author profiling is used in analysis of forensic text for detecting the stated profile of authors matched with their writings and determining the fraudulent behaviour. Danique Sabel observed that [24] most of the traditional works proposed for AP was used content-based and stylistic features with different machine learning algorithms for detecting the age, gender, and native language of author. According to latest research works, the word knowledge differs greatly between various classes of gender like male and female. They conducted experiment with features that are created by using dataset of word knowledge with reaction times and prevalence scores. Furthermore, they used machine learning algorithms such as SVM, RF, DT, and KNN for finding the gender of author. The results of experiment showed that these features that are based on knowledge of words have potential for contributing to tasks of author profiling.

The content in social media platforms represents a primary source for behavioural analysis of the aging population. Abhinay Pandya et al., addressed [25] the age prediction problem from Twitter dataset, where the issue of prediction is considered as a task of classification. For this purpose, they developed an innovative model by using Convolutional Neural Networks. To this end, they relied on metadata specific to social media and language related features. Particularly, authors introduced two features hash-tags occurring in tweets and the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

content of URLs which are not addressed in previous literature. They also employed distributed representations of phrases and words that are present in tweets, URLs and hash-tags, pretrained on suitable dataset in order to utilize their semantic information for age prediction. The results of experiment showed that the CNN model attained an enhancement of micro-averaged F1 score up to 6.6% for dataset of English2, 9.8% for dataset English1, and 12.3% for Dutch dataset when compared with other baseline models.

The main aim of author profiling is correlating the style of writing with demographics of author. Roobaea Alroobaea et al., presented [26] an approach by developing a DSS ("Decision Support System") for identifying gender and age from Twitter tweets. The developed system implemented by using Machine Learning (ML) algorithms and Deep Learning (DL) algorithms for differentiating the classes of gender and age. The experimental results showed that every algorithm attained dissimilar results for gender and age prediction based on power points of every algorithm and the architecture of model. The proposed model adopted the DL models of LSTM and CNN techniques. The authors observed that the proposed DSS is more accurate for detecting the age and gender of authors from authors written texts.

III. DATASET CHARACTERISTICS

In this research work, the experiments carried out with two datasets such as PAN competition 2014 English reviews dataset and PAN competition 2016 English author profiling dataset for predicting the age group and gender of authors. The PAN organization conducted competitions on different tasks from last two decades. In 2013, PAN organizers started the competitions on the task of author profiling. From 2013 onwards, organizers provided different varieties of datasets in different languages for predicting different kinds of author profiles.

The organizers of PAN 2014 competition provided four varieties of datasets such as Reviews, Social Media, Blogs, Twitter in two languages such as English and Spanish, but, Reviews dataset provided in English language only [27]. Two author profiles such as age and gender information is specified in the dataset. In this work, we considered PAN 2014 Competition dataset of Reviews as first dataset. The reviews dataset contains hotel reviews which are collected from tripadvisor.com for a certain period. The reviews dataset contains 4160 reviews of different hotels. Two sub profiles like male and female are specified for gender profile and five sub profiles like 18-24, 25-34, 35-49, 50-64 and 65-xx for age profile. The characteristics of reviews dataset are specified in Table I.

TABLE I. PAN 2014 REVIEWS DATASET DETAILS OF GENDER AND AGE PROFILES

	Gender		Age				
	Male	Female	18-24	25-34	35-49	50-64	65-xx
Number of Documents	2080	2080	360	1000	1000	1000	800
Total number of documents	4160		4160				

We used PAN competition 2016 author profiling dataset is considered as second dataset for experimentation. The PAN 2016 competition dataset contains three languages such as Dutch, English and Spanish Twitter Tweets datasets for predicting two profiles such as gender and age [28]. The same sub profiles of gender and age are used in PAN 2016 competition also. In this dataset also, the gender dataset is balanced but the age dataset is not balanced. The details about the PAN 2016 competition English dataset for author profiling task are displayed in Table II.

TABLE II. THE CHARACTERISTICS OF PAN 2016 ENGLISH TRAINING DATASET

Profile	Class	Number of Documents	Number of Tweets
Gender	Male	218	149059
	Female	218	113972
Age Group	18-24	28	363031
	25-34	140	
	35-49	182	
	50-64	80	
	65-xx	6	

The training dataset contains the xml files of authors which contains the tweets of authors. These training dataset files are pre-processed by deleting html tags and xml tags etc., to prepare the author files with only plain text. The pre-processed files are used for this experimentation.

IV. PERFORMANCE REPRESENTATION MEASURES

Most of the research works used traditional performance evaluation measures like precision, recall, accuracy and F1-measure for evaluating the effectiveness of author profiling approaches. The Table III displays the confusion matrix for a profile P.

TABLE III. CONFUSION MATRIX

		Predicted Class Labels	
		Positive	Negative
Actual Class Labels	Positive	TP	FP
	Negative	FN	TN

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

In this paper, the experiment performed for prediction of two profiles like age and gender. In case of gender, the profile contains two classes like female and male. In case of age, the profile contains 5 classes such as 18-24, 25-34, 35-49, 50-64, 65-xx. In gender profile prediction, the parameters in the confusion matrix table are described like this. The positive class denotes male class and negative class denotes female class. Now, the TP (True Positives) is count of documents under male class predicted as male class, FP (False Positives) is count of documents under male class predicted as female class, FN (False Negatives) is count of documents under female class predicted as male class, and TN (True Negatives) is count of documents under female class predicted as female class by the machine learning algorithm.

In age profile prediction, if positive class is considered as 18-24 class then all other classes such as 25-34, 35-49, 50-64, 65-xx become negative class. TP is count of documents under 18-24 class predicted as 18-24 class, FP is count of documents under 18-24 class is predicted as any class of 25-34, 35-49, 50-64 or 65-xx, FN is the documents under 25-34, 35-49, 50-64 or 65-xx classes are predicted as 18-24 class, TN is documents under 25-34, 35-49, 50-64 or 65-xx classes are predicted as 25-34, 35-49, 50-64 or 65-xx classes.

The precision is described as ratio between counts of documents under positive class is detected as positive and count of positive class of documents. In other words, precision represents the count of relevant documents retrieved and are actually correct. The Equation (1) is used for determining the precision value.

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

The recall is determined as counts of documents under positive class are predicted as positive divided by count of documents is detected as positive class. In other words, Recall represents the count of relevant documents extracted from the correct set of documents. The Equation (2) is used for computing the recall measure.

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

The accuracy measure is determined as the ratio among count of documents correctly predicted their positive class or negative class and total count of documents considered for validation. The Equation (3) is used for determining the accuracy measure.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

F1-measure is computed by using the two measures like recall and precision. It is the harmonic mean of two measures like recall and precision. Equation (4) is used for computing the F1-measure.

$$F_1 = \frac{2 \times \Pr ecision \times \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \quad (4)$$

In this research work, the accuracy is used as a measure for evaluating the performance of proposed approach. In the context of author profiling, the accuracy is the fraction of authors test documents are predicted their author details like gender and age correctly.

V. MACHINE LEARNING ALGORITHMS

The ML algorithms are used for evaluating the efficiency of the proposed approach. The ML algorithms do not understand the textual documents directly. It understands the vector representations of documents. In proposed approaches, different authors used different type of terminology for converting the documents into vector representations. Once the documents are converted as vectors, now the ML algorithm train these document vectors and prepare the classification model by understanding the hidden patterns in the document vectors. The ML algorithms used the cross-validation approach to evaluate the proposed approach. In cross-validation approach, the training documents are separated into two parts, one part is for train the algorithm and the other part is for validating the algorithm. In this paper, RF and SVM are used for assessing the efficiency of our propose approach.

A. Random Forest (RF)

The RF algorithm decides the efficiency of proposed approach based on the collection of decision tress [29]. Decision Tree (DT) is one of the more widely used algorithms because of its ease of understanding and robustness to noisy data. DTs are used for evaluating both classification and regression problems. In decision tree, nodes are represented with attributes and edges represent the condition on attribute. In general, number of edges from a node equal to number of different values of attribute. In binary classification, DT used the binary splitting in every node recursively in order to develop classification tree. Finding the attribute at a certain point of node is one important task in development of decision tree. Gain ratio, Information gain, Gini index measures are used for determining the relevant attributes. Random forest is developed by using DTs. RF is an ensemble method, where n number of decision trees is constructed by using bagging ("bootstrap aggregating") technique. Bagging technique recursively selects training set samples randomly with replacement and fits these samples to n number of trees. After training of all training set samples with different trees, the class label of unknown sample is predicted by considering the majority votes of n number of decision trees.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

B. Support Vector Machine (SVM)

The SVM classifier was proposed in the work of Cortes & Vapnik, 1995 [30]. The classification process of SVM contains two steps. In the first step, identify a hyperplane that is used for separating training data space into two subspaces. Every subspace is denoted with a dissimilar class. In the second step, maximize the margin among the support vectors (also called as borderline vectors) of both classes and hyperplane. To compute the margin among the subspaces, define a hyperplane for every class. These two hyper-planes also denoted as support vectors that are parallel to separating hyperplane and at least one sample vector of every class in training dataset need to lies on support vector. SVM assign one class to input vectors that are lying on one side of hyperplane and assign other class that are lying on another side of hyperplane. SVM is most appropriate for problems of binary classification. In general, most of the problems of text classification are multiclass classification problems, where the dataset contains more than two class labels. SVM used the non-linear kernel functions such as sigmoid, polynomial, and RBF (Radial Basis Function) to solve complex problems by transforming multiple classes data into linearly separable. Kernel functions maps the higher dimensional training dataset vectors into a lower dimensional space for defining a hyperplane. Several researchers observed that the SVM shows best performance in text classification [31].

VI. PROPOSED APPROACH

In this paper, we proposed a TWM based approach for AP. The Figure 1 represents the steps used in the proposed approach. In this approach, first, we apply different pre-processing methods like tokenization, punctuations removal, stop words exclusion and stemming for eliminating the irrelevant data from the dataset. After cleaning the dataset, we extract important words/terms from the dataset based on the frequencies of words. Words or terms are used as features in this experiment. The identified features are used for representing the documents as vectors. The TWMs are used for representing the term value in the document vector representation. The final document vectors are passed to ML techniques for training. ML techniques generate model for classification internally and estimate the accuracy of age and gender prediction.

VII. EXISTING TERM WEIGHT METHODS

In text classification, term weighting play an important role. The computed weight of a term directly affects the significance of a term in total dataset to allow ML algorithm for attaining best classification result. The TWMs determine the weights of terms by considering the term's importance in the document. The term weighting is classified into two types such as STW (Supervised Term Weighting) and UTW (Unsupervised Term Weighting). STW measure used class information of documents for determining the weight of term, whereas UTW does not used class information to compute term weight. Most of the research works succeeded in obtaining good results when STW measures are used because STW measures consider previous information of predefined categories and characteristics of dataset [32]. In this paper, we performed experiment with one UTW measure and seven STW measures for determining the weight of terms.

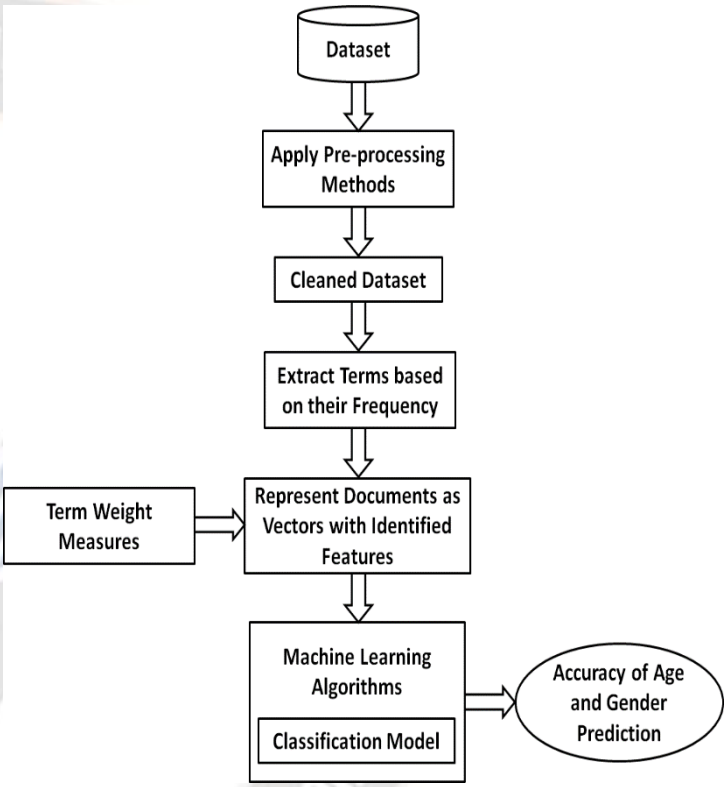

Figure 1. The proposed Approach for Gender and Age Prediction

A. TW-1 - TFIDF ("Term Frequency and Inverse Document Frequency")

TFIDF [33] is generally used as term weighting measure in different research domains such as text mining, information retrieval, and user modelling. It determines the word's importance in a document of a dataset. According to TFIDF, terms that are presented in less number of documents from the total dataset are more useful to differentiate the class of documents [34]. This measure was first introduced in the research of information retrieval for the intention of retrieval of suitable results for the given query. This measure is a combination of two parts such as TF and IDF. The TFIDF of a term Ti specific to document D_k is calculated by using Equation (5).

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

$$TFIDF\left(t_i, D_k\right) = TF\left(t_i, D_k\right) \times \log\left(\frac{N}{DF_i}\right) \quad (5)$$

Where, $TF(t_i, D_k)$ is how many times term t_i present in k^{th} document, N is documents count in dataset, DF_i is count of documents in dataset that contain i^{th} term t_i.

B. TW-2 - TFRF ("Term Frequency and Relevance Frequency")

TFRF is a STW measure which states that the terms those are presented in more documents of positive class having more weight than the terms that are present in more negative class documents [35]. Equation (6) is used for finding the i^{th} term t_i weight in k^{th} document D_k using TFRF.

$$TFRF\left(t_i, D_k\right) = TF\left(t_i, D_k\right) \times \log\left(2 + \frac{A}{C}\right) \quad (6)$$

Where, A is count of documents those contain i^{th} term t_i in documents of positive class and C is count of documents those contain i^{th} term T_i in documents of negative class.

C. TW-3 - TFProb

TFProb measure is developed based on the probability distributions of terms in the dataset. This measure states that the terms those are spreading in more documents under positive class and fewer documents under negative class having more importance than other terms [36]. Equation (7) is used for finding the i^{th} term t_i weight in k^{th} document D_k using TFProb.

$$TF\ \mathrm{Prob}\left(t_i, D_k\right) = TF\left(t_i, D_k\right) \times \log\left(1 + \frac{A}{B}\frac{A}{C}\right) \quad (7)$$

Where, B is is documents count that are not containing T_i term in documents of positive class, A is total documents count that contain i^{th} term T_i in documents of positive class, and C is total documents count that consists of T_i term in documents of negative class.

D. TW-4 - TFCRF (Term Frequency-Category Relevance Frequency)

Y. Liu et al., developed [36] TFCRF measure which is a STW measure. TFCRF measure improves results by using odds of negative and positive class probabilities. This measure is used for enhancing the performance of probability-based term weighting. The Equation (8) is used for calculating the weight of term using TFCRF measure.

$$TFCRF\left(t_i, D_k\right) = ntf \times \log\left(1 + \frac{A-B}{A+B}\right) \quad (8)$$

Where, A is count of documents of positive class contain term t_i, B is count of documents under negative class contain term t_i, ntf is occurrence count of t_i term in D_k document. A and B counts are mutually exclusive. A+B represents the total count of documents in both negative and positive classes contain the term t_i.

E. TW-5 - TF-ICF-Based

TF-ICF [37] is generally used as a STW measure in text mining and information retrieval. The TF-ICF measure states that the terms those are distributed in less number of classes are having highest discriminative power, which means that if a term occurs in only one category of documents then that term having highest weight than the terms are occurred in more than one category of documents. The TF-ICF-Based [38] is a combination of three parts such as TF and ICF (Inverse Class Frequency) and RF (Relevance Frequency). ICF-Based part of measure finds the weight of terms by combining ICF and RF. ICF part determines the term ti distribution between categories. The RF part estimates the term ti distribution among negative and positive category. Equation (9) is used for determining the weight of term using TF-ICF-Based measure.

$$TF-ICF-Based\left(t_i, D_k\right) = TF\left(t_i, D_k\right) \times \log_2\left(2 + \frac{d\left(t_i, c_j\right)}{\max\left(1, d\left(t_i, \overline{c_j}\right)\right)} \times \frac{q}{c\left(t_i\right)}\right) \quad (9)$$

Where, $d(t_i, c_j)$ is count of class c_j documents contain the term t_i, $d(t_i, \overline{c_j})$ is count of other than class c_j documents contain the term t_i, q is classes count in dataset, $c(t_i)$ is count of different classes contain term t_i.

F. TW-6 - TF-BDC (Term Frequency - Balanced Distributional Concentration)

Tao Wang et al., developed TF-BDC measure based on the idea of entropy [39]. The considerations of set of terms that are specific to a category are more needful for discriminating various categories and these set of terms have less entropy value with respect to these specific categories. TF-BDC measure is developed based on the exploration of relationship among term's distinguishing power and its entropy value specific to a set of categories. TF-BDC determines the distinguishing power of a term by considering its global distributional concentration in all categories of a dataset. The Equation (10) denotes the TF-BDC weight of t_i term in document D_k.

$$TF-BDC\left(t_i, D_k\right) = TF\left(t_i, D_k\right) \times \left(1 + \frac{1}{\log_2 q} \times \sum_{j=1}^{q}\left(\frac{\frac{d\left(t_i, c_j\right)}{d\left(t_i\right)}}{\sum_{j=1}^{q}\frac{d\left(t_i, c_j\right)}{d\left(t_i\right)}} \log_2\left(\frac{\frac{d\left(t_i, c_j\right)}{d\left(t_i\right)}}{\sum_{j=1}^{q}\frac{d\left(t_i, c_j\right)}{d\left(t_i\right)}}\right)\right)\right) \quad (10)$$

Where, d(ti, cj) is documents count of class c_j those contain the t_i term, $d(t_i)$ is count of documents contain term t_i, q specifies count of classes in dataset.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
Article Received: 26 February 2023 Revised: 29 March 2023 Accepted: 15 April 2023

G. TW-7 - TF-IDF-ICSDF ("Term Frequency – Inverse Document Frequency – Inverse Class Space Density Frequency")

TF-IDF-ICSDF is developed by revising ICF measure by implementing a novel concept of inverse class space density frequency. This measure offers a positive discrimination on frequent and infrequent terms. According to TF-IDF-ICSDF, the terms that are appeared in fewer classes and that are presented in less number of total dataset documents are having more weight [40]. Equation (11) is used for finding the i^{th} term T_i weight in k^{th} document D_k using TF-IDF-ICSDF.

$$TF-IDF-ICSDF(t_i, D_k) = TF(t_i, D_k) \times \left(1+\log\left(\frac{N}{DF_i}\right)\right) \times \left(1+\log\left(\frac{m}{\sum_{j=1}^{m}\left(\frac{n_{cj}(t_i)}{N_{cj}}\right)}\right)\right) \quad (11)$$

Here, N is count of documents in total dataset, m is count of classes in dataset, $TF(t_i, D_k)$ is how many times term t_i present in k^{th} document D_k, DF_i is how many documents in total dataset contains i^{th} term t_i, N_{cj} is total count of documents in class c_j, $n_{cj}(t_i)$ is total count of documents that contain i^{th} term t_i in class c_j.

H. TW-8 - TF-IGM ("Term Frequency – Inverse Gravity Moment")

TF-IGM measure mainly focused on the rankings information of all classes in the dataset [41]. The ranks are assigned to the classes by focusing on the terms spreading information in multiple classes. The highest rank is assigned to a class when the term is appeared in more documents of that class. TF-IGM measure primarily developed for the purpose of multi-label classification. Equation (12) is used for finding the i^{th} term T_i weight in k^{th} document D_k using TF-IGM measure.

$$TF-IGM(t_i, D_k) = TF(t_i, D_k) \times \left(\frac{f_{i1}}{1+\lambda \times \sum_{j=1}^{m} f_{ij} \times j}\right) \quad (12)$$

Here, f_{i1} is total count of documents those belongs to highest ranked class contain i^{th} term T_i, m is different classes count in the dataset, λ is an adjustable coefficient (λ values are varying from 5 to 9). Most of the research works considered λ value as 7 in their experiments and set it as default value for λ. f_{ij} is total documents count in c_j class those contain the T_i term.

VIII. PROPOSED TERM WEIGHT METHOD (TW– 9)

In this paper, we developed a new TWM based on the distribution of term in different places such as within a document, within total dataset, within documents under positive class and within documents under negative class. The proposed TWM is represented in Equation (13).

$$TW(t_i, D_k) = TWSD * (TWSDS + TWSPC - TWSNC) \quad (13)$$

In Equation (13), $TW(t_i, D_k)$ is the term weight of term t_i in document D_k, TWSD is Term Weight Specific to Document, TWSPC is Term Weight Specific to documents under Positive Class, and TWSNC is Term Weight Specific to Negative Class of documents.

In proposed TWM, the weight of a term is computed as "the weight of a term specific to a document that belongs to positive class has good weight in positive class of documents and total set of document in the dataset. These two weights are summed and subtract the weight of term specific to negative class of documents. The result of this computed value is multiplied with the weight of a term specific to a document."

TWSD defines the strength of a term within a document which is calculated by using Equation (14).

$$TWSD(t_i, D_k) = \frac{TF(t_i, D_k)}{|D_k|} \quad (14)$$

In Equation (14), $TF(t_i, D_k)$ is the weight of term t_i in document D_k, $|D_k|$ is the count of terms in document D_k. In general, the weight of term in a document is determined based on the count of occurrences of a term in a document. The count of a term occurrence mainly depends on the length of a document or count of words in a document. In other words, the document length is proportional to the occurrence count of a term. In Equation (14), the weight of term is normalized by dividing the occurrence count of a term in document with the length of a document.

TWSPC defines the strength of a term in positive class of documents which is represented in Equation (15).

$$TWSPC(t_i, D_k \in C_j) = \frac{DF(t_i, C_j) \times DF(\bar{t_i}, \bar{C_j})}{DF(\bar{t_i}, C_j) \times DF(t_i, \bar{C_j})} \quad (15)$$

In Equation (15), $DF(t_i, C_j)$ is the count of documents in class C_j those contain term t_i, $DF(\bar{t_i}, \bar{C_j})$ is the count of documents in other than class C_j those doesn't contain term t_i, $DF(\bar{t_i}, C_j)$ is the count of documents in class C_j those doesn't contain term t_i, $DF(t_i, \bar{C_j})$ is the count of documents in other than class C_j those contain term t_i. The term t_i is a good representative for a class C_j when the count of documents in positive class contain the term t_i is more than the count of documents in negative class contain the term t_i. In other words, the t_i term strength is more in positive class is more when the $DF(t_i, C_j)$ value is more when compared with $DF(t_i, \bar{C_j})$. The $DF(t_i, \bar{C_j})$ count is less means the count of $DF(\bar{t_i}, \bar{C_j})$ is more. In Equation (15), we

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
Article Received: 26 February 2023 Revised: 29 March 2023 Accepted: 15 April 2023

represented higher values in the numerator and lesser values are represented in denominator to represent the weight of a term specific to positive class.

TWSNC defines the strength of a term in negative class of documents which is represented in Equation (16).

$$TWSNC\left(t_i, D_k \in C_j\right) = \frac{\sum_{k=1, C_k \neq C_j}^{m} \frac{DF\left(t_i, C_k\right)}{DC_k}}{\frac{DF\left(t_i, C_j\right)}{DC_j}} \quad (16)$$

In Equation (16), $DF(t_i, D_k)$ is count of positive class documents contain the term t_i, DC_j is total count of documents in class Cj. The numerator of the equation indicates the ratio among the sum of counts of different negative class documents those contain term t_i and total count of documents in different negative class documents. The denominator indicates the ratio among the count of positive class documents contain term t_i and total count of documents in positive class. If the term is good representative for positive class, then the numerator value is less when compared with denominator which results the value of Equation (16) is less.

TWSDS defines the strength of a term in total dataset of documents which is represented inn Equation (17).

$$TWSDS\left(t_i, D_k\right) = \log_2\left(2 + \frac{m}{CF_i} \times \frac{N}{DF_i}\right) \quad (17)$$

In Equation (17), the m is count of total classes in dataset, CF_i is count of classes contains the term, N is count of total documents in the dataset, DF_i is count of total documents contain the term t_i. This factor says that terms that are occurred in less number of documents in total dataset and terms that are occurred fewer classes are having good discriminative power to differentiate different documents.

A. Analysis of Proposed TWM

In this paper, we take an example for analysing the efficiency of our proposed TWM. We considered 100 documents under positive class and 100 documents under negative class. Table IV displays the term T1 distribution in negative and positive classes.

TABLE IV. THE TERM T1 DISTRIBUTION IN DATASET

Positive Class	Negative Class
80	40
20	60

In Table IV, the T1 term occurred in 80 documents of positive class and 20 documents of negative class.
The TWSPC(T1) = (80 * 60) / (40 * 20) = 4800/800 = 6
The TWSPC determines good value for term T1 because the T1 term appeared in more documents of positive class when compared with negative class documents.

Table V displays the distribution of T2 term under positive and negative classes.

TABLE V. THE T2 TERM DISTRIBUTION IN DATASET

Positive Class	Negative Class
30	80
70	20

In Table V, the term T2 appeared in 30 documents of positive class and 80 documents of negative class.
The TWSPC(T2) = (30 * 20) / (80 * 70) = 600/5600 = 0.107.

The TWSPC determines less value for term T2 because the term T2 appeared in less documents of positive class when compared with documents under negative class.
TWSNC(T1) = 40/80 = 0.5
The TWSNC value of T1 is less because the T1 occurred in less documents of negative class than documents under positive class.
TWSNC(T2) = 80/30 = 2.667
The TWSNC value of T1 is good because the T1 appeared in more documents of negative class than documents under positive class.

IX. EXPERIMENTAL RESULTS OF AGE AND GENDER PREDICTION

In this article, the experiment implemented with content related features of words for predicting age and gender of authors. The important words are identified based on the occurrence count of words. In this experiment, most frequent 10000 words are used in the experiment. In most of the cases, it was identified that the accuracies are diminished when experiment conducted with more than 10000 terms. The documents are denoted as vectors with identified features. The TWMs are used for determining the term importance in the document vector representation. Two PAN competition datasets are used in this experiment and two machine learning algorithms like RF and SVM are used for accuracy prediction of age and gender.

A. Experiment Results of Gender Prediction on PAN 2014 Dataset

The Table VI shows the accuracies of proposed approach with different term weight measures on PAN 2014 Dataset for gender prediction when SVM is used as machine learning algorithm.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

TABLE VI. ACCURACIES OF SUPPORT VECTOR MACHINE FOR GENDER PREDICTION ON PAN 2014 DATASET

TWM's / No. of Terms	TW-1	TW-2	TW-3	TW-4	TW-5	TW-6	TW-7	TW-8	TW-9
2000	70.41	71.11	73.34	75.21	75.19	77.34	78.53	79.35	82.26
4000	70.89	72.26	73.97	75.57	76.27	77.82	79.38	80.15	83.14
6000	71.38	72.87	74.58	75.83	77.56	78.59	79.81	80.87	84.56
8000	72.57	73.57	75.43	77.65	76.78	79.15	80.45	82.86	83.97
10000	71.68	74.96	76.19	76.49	78.91	80.08	81.66	81.23	85.32

In Table VI, the proposed approach with proposed TWM accomplished an accuracy of 85.32% for gender prediction when experiment conducted with 10000 terms for document vector representation. It was observed that the proposed TWM accomplished best accuracies for prediction of gender when compared with specified TWMs. It was also identified that the accuracies are increased when count of terms are increased for document vector representation in most of the cases.

The Table VII shows the accuracies of proposed approach with different term weight measures on PAN 2014 Dataset for prediction of gender when RF is used as machine learning algorithm.

TABLE VII. ACCURACIES OF RANDOM FOREST FOR GENDER PREDICTION ON PAN 2014 DATASET

TWM's / No. of Terms	TW-1	TW-2	TW-3	TW-4	TW-5	TW-6	TW-7	TW-8	TW-9
2000	72.39	74.11	75.06	77.14	78.64	80.28	80.15	81.39	84.42
4000	73.41	74.86	75.52	78.29	79.27	80.73	81.23	81.98	85.29
6000	73.87	75.25	77.49	79.37	79.69	81.41	81.81	83.24	85.76
8000	74.24	75.79	76.27	80.64	80.52	82.62	82.71	82.15	87.49
10000	75.08	76.45	78.01	79.83	81.49	81.94	83.42	84.65	86.51

In Table VII, the proposed approach with proposed TWM attained an accuracy of 87.49% for gender prediction when experiment implemented with 8000 terms for vector representation of document. It was recognized that the proposed TWM accomplished best accuracies for prediction of gender when compared with specified TWMs. It was also identified that the accuracies are increased when count of terms are increased for document vector representation in most of the cases.

B. Experiment Results of Gender Prediction on PAN 2016 Dataset

The Table VIII shows the accuracies of proposed approach with different term weight measures on PAN 2016 Dataset for gender prediction when SVM classifier is used as a machine learning algorithm.

In Table VIII, the proposed approach with proposed TWM accomplished an accuracy of 86.23% for prediction of prediction when experiment executed with 10000 terms for vector representation of document. It was detected that the proposed TWM accomplished best accuracies for prediction of gender when compared with specified TWMs. It was also identified that the accuracies are increased when count of terms are increased for document vector representation in most of the cases.

TABLE VIII. ACCURACIES OF SUPPORT VECTOR MACHINE FOR GENDER PREDICTION ON PAN 2016 DATASET

TWM's / No. of Terms	TW-1	TW-2	TW-3	TW-4	TW-5	TW-6	TW-7	TW-8	TW-9
2000	72.18	74.13	76.14	76.11	77.07	79.14	78.15	80.27	83.32
4000	73.13	74.57	76.32	76.64	78.17	79.48	79.25	81.42	84.24
6000	73.72	75.11	77.54	77.34	79.28	80.19	79.81	81.85	85.28
8000	75.93	76.82	77.03	78.69	80.43	80.78	82.79	82.38	84.87
10000	74.37	77.45	78.95	79.66	79.57	81.44	81.19	83.65	86.23

The Table IX shows the accuracies of proposed approach with different term weight measures on PAN 2016 Dataset for prediction of gender when RF is used as machine learning algorithm.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

TABLE IX. ACCURACIES OF RANDOM FOREST FOR GENDER PREDICTION ON PAN 2016 DATASET

TWM's / No. of Terms	TW-1	TW-2	TW-3	TW-4	TW-5	TW-6	TW-7	TW-8	TW-9
2000	77.45	79.23	78.56	80.13	81.46	82.31	83.21	84.26	86.34
4000	78.56	79.87	79.15	81.27	81.89	83.12	84.14	85.28	86.87
6000	79.23	80.14	81.56	81.69	82.27	83.73	85.38	86.31	87.16
8000	80.29	80.95	80.24	82.38	84.97	84.39	84.79	86.83	88.74
10000	79.76	81.39	82.74	83.67	83.58	85.41	86.27	87.46	87.93

In Table IX, the proposed approach with proposed TWM attained an accuracy of 88.74% for gender prediction when experiment executed with 10000 terms for vector representation of document. It was detected that the proposed TWM accomplished best accuracies for prediction of gender when compared with specified TWMs. It was also identified that the accuracies are increased when count of terms are increased for document vector representation in most of the cases.

C. Experiment Results of Age Prediction on PAN 2014 Dataset

The Table X shows the accuracies of proposed approach with different term weight measures on PAN 2014 Dataset for age prediction when SVM classifier is used as machine learning algorithm.

TABLE X. ACCURACIES OF SUPPORT VECTOR MACHINE FOR AGE PREDICTION ON PAN 2014 DATASET

TWM's / No. of Terms	TW-1	TW-2	TW-3	TW-4	TW-5	TW-6	TW-7	TW-8	TW-9
2000	68.24	69.36	70.47	71.14	73.27	72.39	73.17	75.21	77.42
4000	68.76	70.13	70.89	72.39	73.81	73.12	73.72	75.76	77.89
6000	69.35	71.42	71.36	73.24	74.27	75.48	74.36	76.42	79.65
8000	70.57	70.79	72.28	74.79	74.69	74.37	76.76	77.63	78.32
10000	71.34	72.61	73.56	73.86	75.06	76.08	75.49	78.26	80.71

In Table X, the proposed approach with proposed TWM accomplished an accuracy of 80.71% for age prediction when experiment executed with 10000 terms for vector representation of document. It was detected that the proposed TWM accomplished best accuracies for prediction of age when compared with specified TWMs. It was also identified that the accuracies of age prediction are increased when count of terms are enhanced for vector representation of document in most of the cases.

The Table XI shows the accuracies of proposed approach with different term weight measures on PAN 2014 Dataset for age prediction when RF classification algorithm is used as machine learning algorithm.

In Table XI, the proposed approach with proposed TWM accomplished an accuracy of 82.59% for age prediction when experiment executed with 10000 terms for vector representation of document. It was detected that the proposed TWM accomplished best accuracies for prediction of age when compared with specified TWMs. It was also identified that the accuracies of age prediction are increased when count of terms are enhanced for vector representation of document in most of the cases.

TABLE XI. ACCURACIES OF RANDOM FOREST FOR AGE PREDICTION ON PAN 2014 DATASET

TWM's / No. of Terms	TW-1	TW-2	TW-3	TW-4	TW-5	TW-6	TW-7	TW-8	TW-9
2000	71.26	71.38	72.41	72.52	75.23	75.24	77.18	77.49	79.16
4000	71.96	72.23	72.89	73.19	75.72	75.83	77.83	78.52	79.79
6000	72.37	72.84	73.26	75.34	76.28	76.36	78.37	79.87	80.31
8000	72.79	74.67	74.38	74.27	76.92	78.14	78.68	79.31	82.59
10000	73.24	73.47	75.84	76.79	77.88	77.48	79.51	80.92	81.42

D. Experiment Results of Age Prediction on PAN 2016 Dataset

The Table XII shows the accuracies of proposed approach with different term weight measures on PAN 2016 Dataset for age prediction when SVM classifier is used as machine learning algorithm.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

TABLE XII. ACCURACIES OF SUPPORT VECTOR MACHINE FOR AGE PREDICTION ON PAN 2016 DATASET

TWM's / No. of Terms	TW-1	TW-2	TW-3	TW-4	TW-5	TW-6	TW-7	TW-8	TW-9
2000	63.32	65.13	65.36	68.54	69.62	70.26	71.19	72.32	75.17
4000	64.28	65.82	66.24	69.27	70.61	70.94	72.42	72.96	75.82
6000	64.76	67.53	67.37	69.82	71.29	72.62	73.34	73.27	76.91
8000	65.31	66.64	68.19	71.91	71.83	71.45	73.86	75.77	76.26
10000	66.85	68.48	69.50	70.48	72.94	73.38	74.15	74.41	77.68

In Table XII, the proposed approach with proposed TWM accomplished an accuracy of 77.68% for age prediction when experiment implemented with 10000 terms for vector representation of document. It was detected that the proposed TWM accomplished best accuracies for prediction of age when compared with specified TWMs. It was also identified that the accuracies of age prediction are increased when count of terms are enhanced for vector representation of document in most of the cases.

The Table XIII shows the accuracies of proposed approach with different term weight measures on PAN 2016 Dataset for age prediction when RF classifier is used as machine learning algorithm.

TABLE XIII. ACCURACIES OF RANDOM FOREST FOR AGE PREDICTION ON PAN 2016 DATASET

TWM's / No. of Terms	TW-1	TW-2	TW-3	TW-4	TW-5	TW-6	TW-7	TW-8	TW-9
2000	68.25	70.36	70.41	71.35	72.52	74.28	74.39	75.17	76.67
4000	69.56	70.84	71.36	71.89	73.17	74.82	74.82	75.72	77.26
6000	70.33	71.32	71.82	73.62	73.76	75.18	75.28	76.81	77.73
8000	70.79	72.58	72.57	72.53	74.49	76.59	76.46	76.24	79.42
10000	71.15	71.91	73.44	74.79	75.82	75.73	77.21	77.89	78.35

In Table XIII, the proposed approach with proposed TWM accomplished an accuracy of 80.71% for age prediction when experiment executed with 10000 terms for vector representation of document. It was observed that the proposed TWM accomplished best accuracies for prediction of age when compared with specified TWMs. It was also identified that the accuracies of age prediction are increased when count of terms are enhanced for vector representation of document in most of the cases.

Overall, the RF classifier shows best performance for age and gender prediction than the performance of SVM classifier.

X. DISCUSSION OF RESULTS

The proposed TWM accomplished best accuracies for age and gender prediction when compared with the other TWMs. The TWMs are proposed by different researchers based on the information of the way the terms are distributed in documents of dataset. Most of the TWMs like TF-RF and TF-PROB developed based on the term distribution in positive class of documents and negative class of documents. TFIDF assign importance to terms that are distributed in less number of documents in the dataset. TF-ICF measure determines the importance based on the number of classes contain the term.

In this article, we proposed a new term weight measure based on the consideration of all possible information to find the significance of a term. The proposed TWM considers four concepts such as term importance in a document, term importance in a dataset, term importance in positive class of documents and term importance in a negative class of documents, where it is possible to enhance the importance of a term. We observed that the run time of proposed term weight measure is little bit more when compared with the run time of other TWM, but the proposed TWM attained good accuracy.

The TWMs attained good accuracies for age and prediction on PAN 2014 dataset when compared with the accuracies of PAN 2016 dataset. We observed that the documents are distributed in a balanced way in PAN 2014 dataset when compared with PAN 2016 dataset. This is the main reason to obtain less accuracy on PAN 2016 dataset.

XI. CONCLUSIONS AND FUTURE SCOPE

The AP is a method of predicting information about authors like gender and age by analyzing their written texts. Several applications used these methods for estimating the basic information about the authors. In this work, we proposed an approach for gender and age prediction. In this proposed approach, the content based features of terms are used for vector representation of documents. We proposed a TWM to represent the importance of a term in document vector representation. The performance of proposed TWM is compared with the performance of different existing TWMs. The proposed approach with proposed TWM attained accuracies of 85.32% and 86.23% for gender prediction on PAN 2014 dataset and PAN 2016 datasets respectively when SVM was used as a machine learning algorithm. The proposed approach with proposed TWM accomplished accuracies of 87.49% and 88.74% for gender prediction on PAN 2014 dataset and PAN 2016 datasets respectively when RF was used as a machine learning algorithm. The proposed approach with proposed TWM accomplished accuracies of 80.71% and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

77.68% for age prediction on PAN 2014 dataset and PAN 2016 datasets respectively when SVM was used as a machine learning algorithm. The proposed approach with proposed TWM accomplished accuracies of 82.59% and 80.71% for age prediction on PAN 2014 dataset and PAN 2016 datasets respectively when RF was used as a machine learning algorithm.

In future work, we are planning to experiment with different word embedding techniques like Word2Vec, FastText, GloVe and BERT for representing words as word embedding vectors. We are also planning to experiment with advanced deep learning models to extract contextualized information from text.

REFERENCES

[1] D. Lazer, A.S. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al., Life in the network: the coming age of computational social science, Science (New York, NY) 323 (5915) (2009) 721.

[2] E. Bothos, D. Apostolou, G. Mentzas, Using social media to predict future events with agent-based markets, IEEE Intell. Syst. (1) (2010).

[3] M.J. Paul, M. Dredze, You are what you tweet: Analyzing twitter for public health, ICWSM 20 (2011) 265–272.

[4] A. Mislove, Pulse of the nation: Us mood throughout the day inferred from twitter, 2010, http://www.ccs.neu.edu/home/amislove/twittermood/.

[5] S. Asur, B.A. Huberman, Predicting the future with social media, in: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 01, IEEE Computer Society, 2010, pp. 492–499.

[6] A. Dittrich, C. Lucas, A step towards real-time detection and localization of disaster events based on tweets, in: Proceedings of the 10th International ISCRAM Conference, 2013.

[7] M. Oussalah, A. Zaidi, Forecasting weekly crude oil using twitter sentiment of us foreign policy and oil companies data, in: 2018 IEEE International Conference on Information Reuse and Integration, IRI, IEEE, 2018, pp. 201–208.

[8] Argamon, S., Koppel, M., Fine, J., Shimoni, A. R.: Gender, genre, and writing style in formal written texts. TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-, 23(3), 321-346 (2003).

[9] Koppel, M., Argamon, S., Shimoni, A. R.: Automatically categorizing written texts by author gender. Literary and linguistic computing, 17(4), 401-412 (2002).

[10] Schler, J., Koppel, M., Argamon, S., Pennebaker, J. W.: Effects of age and gender on blogging. In AAAI spring symposium: Computational approaches to analyzing weblogs, Vol. 6, pp. 199-205 (2006).

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

[12] 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, (2008), 211-236.

[13] 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, (2001), 2001-2009.

[14] 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).

[15] Chanchal Suman, Anugunj Naman, Sriparna Saha, Pushpak Bhattacharyya, "A Multimodal Author Profiling System for Tweets ",IEEE Transactions on Computational Social Systems, Volume: 8 Issue: 6, July 2021, PP. 1407 – 1416

[16] 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 81Issue 19, Aug 2022, pp 27799–27817

[17] Ameer, Iqraa, Sidorov, Grigoria, Nawab, Rao Muhammad Adeelb, Author profiling for age and gender using combinations of features of various types, Journal of Intelligent & Fuzzy Systems, vol. 36, no. 5, pp. 4833-4843, 2019

[18] Ibrahim Mousa Al-Zuabi , Assef Jafar and Kadan Aljoumaa, "Predicting customer's gender and age depending on mobile phone data", Journal of Big Data (2019) 6:18, pp. 1 – 16, https://doi.org/10.1186/s40537-019-0180-9

[19] Erhan Sezerer, Ozan Polatbilek, Selma Tekir, "Gender Prediction from Tweets: Improving Neural Representations with Hand-Crafted Features", arXiv:1908.09919v2 [cs.CL] 6 Sep 2019

[20] Piot–Perez-Abadin, P., Martin–Rodilla, P. and Parapar, J. Experimental Analysis of the Relevance of Features and Effects on Gender Classification Models for Social Media Author Profiling. In Proceedings of the 16th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2021), pages 103-113, DOI: 10.5220/0010431901030113, ISBN: 978-989-758-508-1

[21] Janneke van de Loo and Guy De Pauw and Walter Daelemans, "Text-Based Age and Gender Prediction for Online Safety Monitoring", International Journal of Cyber-Security and Digital Forensics (IJCSDF) 5(1): 46-60, The Society of Digital Information and Wireless Communications, 2016

[22] Seifeddine Mechti, Maher Jaoua, Rim Faiz, Heni Bouhamed and Lamia Hadrich Belguith, " Author Profiling: Age Prediction Based on Advanced Bayesian Networks", Research in Computing Science 110 (2016), pp. 129–137

[23] Esam Alzahrani and Leon Jololian, " How Different Text-Preprocessing Techniques using The Bert Model Affect the Gender Profiling of Authors", CS & IT - CSCP 2021, 2021, pp. 01-08

[24] Danique Sabel, " Gender Prediction Based on Word Knowledge using Machine Learning Techniques", Thesis submitted for Department of Cognitive Science & Artificial

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6633
***Article Received**: 26 February 2023 **Revised**: 29 March 2023 **Accepted**: 15 April 2023*

Intelligence, Tilburg, the Netherlands, January 2019, pp. 01-21

[25] Abhinay Pandya, Mourad Oussalah, Paola Monachesi, Panos Kostakos, "On the use of distributed semantics of tweet metadata for user age prediction", Future Generation Computer Systems 102 (2020) 437–452

[26] Roobaea Alroobaea, Sali Alafif, Shomookh Alhomidi, Ahad Aldahass, Reem Hamed, Rehab Mulla, Bedour Alotaibi, "A Decision Support System for Detecting Age and Gender from Twitter Feeds based on a Comparative Experiments", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 12, 2020, pp. 370-376

[27] 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).

[28] 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.

[29] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1,pp. 5–32, 2001.

[30] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

[31] Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. European conference on machine learning, 137-142.

[32] F. Carvalho, G. P. Guedes, TF-IDFC-RF: A Novel Supervised Term Weighting Scheme, https://arxiv.org/abs/2003.07193, 2020.

[33] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., 1986.

[34] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

[35] M. Lan, C. L. Tan, J. Su, Y. Lu, Supervised and traditional term weighting methods for automatic text categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 4, pp. 721-735, April, 2009.

[36] Liu, Y., Loh, H. T., & Sun, A. (2009). Imbalanced text classification: A term weighting approach. Expert Systems with Applications, 36 (1), 690–701. http://doi.org/10.1016/j.eswa.2007.10.042

[37] V. Lertnattee, T. Theeramunkong, Analysis of inverse class frequency in centroid-based text classification, IEEE International Symposium on Communications and Information Technology (ISCIT) 2004, Sapporo, Japan, 2004, pp. 1171-1176.

[38] D. Wang, H. Zhang, Inverse-Category-Frequency Based Supervised Term Weighting Schemes for Text Categorization, Journal of Information Science and Engineering, Vol. 29, No. 2, pp. 209-225, March, 2013.

[39] Tao Wang, Yi Cai, Ho-fung Leung, Zhiwei Cai and Huaqing Min, "Entropy-based Term Weighting Schemes for Text Categorization in VSM", 2015 IEEE 27th International Conference on Tools with Artificial Intelligence, 2015, pp 325-332.

[40] Ren, F., & Sohrab, M. G. (2013). Class-indexing-based term weighting for automatic text classification. Information Sciences, 236 , 109–125. http://doi.org/10.1016/j.ins.2013.02.029.

[41] K. Chen, Z. Zhang, J. Long, H. Zhang, Turning from tf-idf to tf-igm for term weighting in text classification, Expert Systems with Applications 66(2016) 1339-1351.