A Term Weight Measure based Approach for Author Profiling

Karunakar Kavuri
Research Scholar, Department of CSE
VelTech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology
Avadi, Chennai, India
karunakar.mtech@gmail.com

M Kavitha
Professor, Department of CSE
VelTech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology
Avadi, Chennai, India
kavitha@veltech.edu.in

***Abstract*— Author Profiling (AP) is a task of determining the demographic features like Age, Gender, Nativity language, location, Personality traits etc. about the author of a document. AP is used in different applications such as marketing linguistic profile, security and forensic science. The researchers proposed different types of solutions to author profiling based on stylistic features, content based features and deep learning techniques. The content based features proved their significance to improve the performance of author profiles prediction. Several approaches faced a problem of high dimensionality of features when experimented with content based features. The researchers used feature selection algorithms to identify the important features for experimentation. In this work, a Term Weight Measure (TWM) based approach is proposed for author profiling problem. In this approach, the important features are identified by using Feature Selection (FS) algorithm. After features are identified, the next important task is representation of document with identified features. The documents are represented as vectors and computation of each feature value in the vector representation is another important research task. TWMs are used to determine the importance of a feature in the vector representation. In the proposed approach, we proposed a new TWM based on the way the terms are distributed in corpus of documents. The proposed term weight measure performance is compared with different existing TWMs. The PAN competition 2014 reviews dataset is used for age and gender prediction of the author. Two Machine Learning (ML) algorithms such as Random Forest (RF) and Support Vector Machine (SVM) are used to evaluate the proposed term weight measure based approach. The experimental results attained in this work for age and gender prediction are good when compared with several popular solutions to author profiling.**

Keywords*— *Author Profiling, Age Prediction, Gender Prediction, Feature Selection Algorithm, Term Weight Measure, Machine Learning Algorithms

I. INTRODUCTION

The information in the form of text is exponentially enhancing in the internet through different types of websites like blogs, forums, reviews and other social media websites. The crimes also enhancing with the increment of textual information like harassing messages, threatening mails and fake profiles. Knowing the author details about textual information becomes an important task to reduce the crimes in the internet. Authorship Analysis is one research area to extract the information of the author from their written texts. Authorship analysis was classified into three classes such as Authorship Verification (AV), Authorship Attribution (AA) and Authorship Profiling [AP] [1].

The AV method verifies whether the new text is written by a suspected author or not by examining different texts of suspected author. AA determines the author of a new text by examining multiple authors' text [2]. The Authorship Profiling task (AP) is to extract demographic aspects like gender, age, location, occupation, socio-economic level and native language, level of well-being, personality traits or educational background of a person from their texts [3]. At the beginning of the AP task, formal texts such as newspapers, books or magazines were analyzed to determine the aspects of their authors. In recent years, researchers concentrated on determining the profile of people through their social network accounts. Authorship Profiling is used in various real-time applications like marketing, forensics, security etc.

The use of author profiling was changed from only be used in the forensic investigation and internet security, also to be applied in targeted marketing, advertisement and educational domain [3]. Companies interested in obtaining the knowledge of what describe the people that like or dislike their product. In marketing domain, the expert's interest is to know the identity and demographic characteristics of the various users, with the intention of directing the advertising for exploiting the product in a better way. Authorship Profiling helped in security to narrow the potential number of suspects, or even help rule out potential suspects. In social media websites, most of the people are hiding their original details in their profiles and perform illegal or deceptive acts such as sexual harassment and extortion through textual messages. AP techniques are helpful to determine the basic information of perpetrator. The Forensic linguists make use of linguistic knowledge and authorship profiling techniques to study texts and determining the type of bad behaviors. In educational domain, AP techniques are helpful to determine the exceptional students by examining their styles of writings.

Most of the research works exploited various types of Content based Features (CFs) to distinguish the authors' style of writing. CFs are based on the content bearing terms used by the authors in their text. The corpus contains huge number of content bearing terms, but few terms are more informative terms which are helpful for discriminating the authors. Most of the researchers used feature selection algorithms to determine the best informative features. These

978-1-6654-8386-5/22/$31.00 ©2022 IEEE

identified features are used for vector representation of documents and these vectors are forwarded to Machine Learning (ML) algorithms to generate the classification model. This model is used to identify the gender and age class label of a new document. In this work, a TWM based approach is proposed for AP. In this approach, the experiment started with FS Algorithm to identify the most relevant features. Once features are extracted, the next step is representing the document with identified features as vectors. The ML algorithms understand only document representation only. In the vector representation of features, the value of a feature influences the performance improvement of proposed approaches. TWMs are used for this purpose to compute the value of a feature in document vector representation. In this proposed approach, a new TWM is proposed by considering the term distributions in different classes of documents in dataset. The gender and age profiles are considered in this work for prediction. The ML algorithms such as RF classifier and SVM are used for generating classification model. This model predicts the accuracy of age and gender prediction. In this work, PAN competition 2014 Reviews dataset is used for age and gender prediction.

This paper is planned in 7 sections. The survey on AP techniques is explained in section 2. The dataset properties are presented in section 3. The section 4 discusses the proposed approach for AP and explained the different TWMs and FSA. The experimental results of proposed approach are analysed in section 5. The section 6 discusses the results of gender and prediction. The section 7 concludes this work with future plans.

II. Review on Existing Works of Author Profiling

Author profiling is a technique of predicting traits of one or many authors automatically from their text. Author profiling groups documents of authors based on their content, semantic tags used, topics discussed and similarity of documents. Author Profiling also group the authors based on their writing styles in social media circle. Automatic detection of author profiles from the text has various applications in harassment cases detection, forensic analysis and marketing. The benchmark dataset is required to analyze and developing suitable solutions for author profiling. In recent years, different types of standard measures were generated for evaluating different genres such as blogs, social media, tweets and hotel reviews [4].

The exponential growth of textual content in social media creates highly undesirable problems like propagation of offensive and abusive language in the internet. The author profiling approaches are used in detection of hate speech messages. The existing research suggests that the hateful messages which are propagated by the users, form communities around them and share a set of common stereotypes. The existing popular approaches for detection of hate speech messages majorly depend on sematic and lexical cues of text. Pushkar Mishra et al., proposed [5] a novel approach for hate speech detection which includes the profiling features of Twitter users based on community. They experimented on a dataset that contains 16000 tweets and observed that the proposed approach performance is significantly higher than existing popular works in hate speech detection.

Chiyu Zhang et al., Proposed [6] models for identifying the gender, language variety and age from social media text in the shared task of AP and deception detection in Arabic language. The model is developed by using pre-trained BERT (Bidirectional Encoders Representation from Transformers). They obtained accuracies of 81.67%, 54.72% and 93.75% for gender, age and language variety prediction respectively. Zhang et al., developed [7] different models for predicting age, gender and language variety in the Arabic AP and deception detection shared task. The dataset contains Tweets in Arabic language that was splitted into a training data of 225,000 tweets and a test data of 720,000 tweets. They used multi-lingual BERT-based model with 768 hidden units, 12 layers, 12 attention heads and 110,000,000 parameters. The proposed BERT-based model attained accuracies of 81.67% and 54.72% for gender and age prediction respectively.

In recent past years, the bots which are accounts in social media that were operated automatically, gained substantial importance in worldwide. Some of them are used bots for malicious activities like spreading of disinformation or swaying political elections. Detection of social bots becomes an emerging research area in recent times. Inna Vogel et al., proposed [8] a system by using character n-grams, word unigrams and word bigrams as features. Linear SVM was used to train the system. The proposed model obtained an overall accuracy of 0.91, 0.92 for bot detection in Spanish and English datasets respectively. The model achieved 0.78 and 0.82 accuracies for gender prediction in Spanish and English datasets respectively.

Kowsari et al., experimented [9] on a dataset of Twitter messages in English language for gender prediction of an author. The training dataset of Twitter contains messages written by 1800 females and 1800 males. The test dataset contains messages written by 1200 females and 1200 males. They applied different models like CNN and Random Multi-model DL and using TFIDF and Glove features and the final decision was taken by using a scheme of majority vote. The proposed models attained an accuracy of 0.8633 and F1-score of 0.8583 for gender classification.

Author profiling techniques also used to detect the gender information from images also. Moniek Nieuwenhuis et al., developed [10] a system to participate in AP shared task of PAN 2018. The task is predicting the gender of an author from text and images in Arabic, English and Spanish datasets. They participated in all subtasks. The final system submitted in the competition used Logistic Regression as classifier, word and character n-grams as textual features and proportion, presence and number of faces to detect face emotions as well as selfies as image-based features. The experiment also conducted with word embeddings and observed that the performance of a system affected negatively. The experimental results show that the performance was improved slightly for Spanish and Arabic datasets when text-based features are added to image-based features. They obtained highest accuracies of 80.3% for Spanish dataset using only text based features, 78.7% for Arabic dataset using both image and text based features, 81.2% for English dataset using only text based features on test dataset of PAN 2018.

III. DATASET CHARACTERISTICS

In this work, the experiment conducted on the reviews dataset which is collected from PAN 2014 competition reviews dataset [11]. The dataset contains 4160 documents and two profiles such as gender and age about an author. The gender profile contains two classes such as male and female. Both male and female classes contain 2080 documents each. The age profile contains five classes such as 18-24, 25-34, 35-49, 50-64 and 65+. The characteristics pertaining to dataset is displayed in Table 1. The dataset is balanced in case of gender profile and unbalanced in case of age profile.

TABLE 1. THE REVIEWS DATASET CHARACTERISTICS

Classes / Profiles		Number of Reviews
Gender	***Male***	2080
	Female	2080
Age	***18-24***	360
	25-34	1000
	35-49	1000
	50-64	1000
	65+	800

In this work, two ML algorithms such as SVM and RF are used for evaluating the proposed approach. Random Forest (RF) Classifier is a classifier that utilizes both bagging of features and random selection which intern uses ensemble learning technique. Random selection is a process of developing a set of decision trees. From the training data, a decision tree is constructed with replacement is called bagging. Each decision tree acts as base classifier in finding the information about the class label of a new document. Later, a voting process conducted which is part of ensembling. In SVM classifiers, identify the support vectors among the set of instances in the dataset. These support vectors are used to predict the author profiles of unknown document.

The outcomes of machine learning algorithms are represented by using different performance measures such as recall, precision, F1-Score and accuracy for evaluation. In this work, the proposed approach results are displayed in terms of accuracy measure. Accuracy is number of author documents are correctly predicted their gender and age from a set of author documents considered for experimentation.

IV. TERM WEIGHT MEASURES BASED APPROACH FOR AUTHOR PROFILING

In this work, a TWM based approach is proposed for age and gender prediction. The procedure of proposed approach is shown in Fig. 1. In this approach, first, the pre-processing techniques such as stopwords elimination and stemming are applied on the training dataset to remove unwanted data from the dataset. The stopwords are words like articles, prepositions, determiners, and pronouns etc., which are not having any distinguishing power to discriminate the authors' style of writing. The stemming is performed to transform the words into their root form to decrease the unique words count in the dataset. After cleaning the dataset, the next important step is identification of features from the dataset.

A feature is a property of a document which is used to differentiate the given document. Majority of the entities and documents have many features. FSAs are used to identify the relevant features and for eliminating the redundant features. In this approach, RDC FSA is used to find the important features.

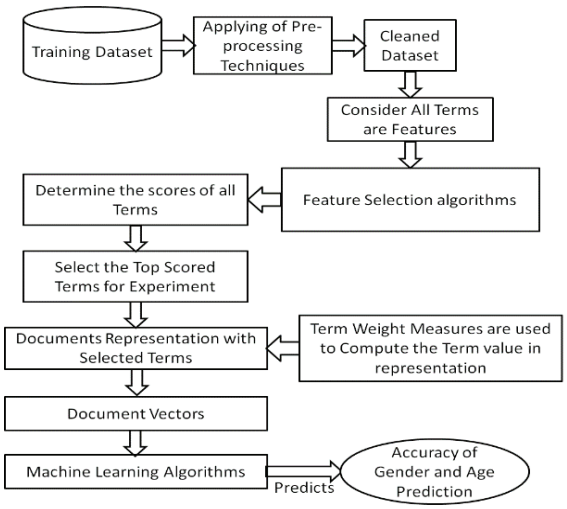

Fig. 1. The Proposed TWMs based Approach

After computing the scores of features by the feature selection algorithms, the top scored features are considered for experiment. In this work, the content based features of terms that are used by the authors in their writings are considered as features. The documents are represented as vectors by using the top ranked terms which are identified in the previous step. In the vector representation, the term value is computed by using TWM. In this work, a new TWM is proposed for computing the term weight. The document vectors are given to MLAs for training. The trained model is used to predict the accuracy of gender and age prediction. The accuracy of gender and age prediction primarily depends on the TWM and FSA that are used in the experimentation.

A. Relative Discriminative Criterion (RDC)

The RDC measure [12] is is represented in Equation (1).

$$RDC(T_i, C_j) = \left(\frac{|DF_{pos}(T_i) - DF_{reg}(T_i)|}{\min(DF_{pos}(T_i), DF_{reg}(T_i)) \times TF(T_i, C_j)} \right) \quad (1)$$

Where, $DF_{pos}(T_i)$ and $DF_{neg}(T_i)$ are the number of positive and negative class of documents contain the term T_i respectively, $TF(T_i, C_j)$ is the occurrence count of term T_i in all the class C_j documents.

B. Existing Term Weight Measures

The researchers proposed various TWMs in different text classification based research domains. In this work, various TWMs are used for computing the term weight in vector representation of a document.

1) TFIDF (Term Frequency and Inverse Document Frequency)

TFIDF measure is a popular measure which is proposed to determine the term importance in a document [13]. TFIDF is

successfully used in text classification domain to calculate the term weight in a document. Equation (2) is used to compute the TFIDF of a term T_i in document D_k.

$$TFIDF(T_i, D_k) = TF(T_i, D_k) * \log\left(\frac{N}{1+DF_i}\right) \quad (2)$$

Where, $TF(T_i, D_k)$ is the occurrence count of T_i in document D_k, N is documents count in dataset and DF_i is count of documents which contain term T_i at least one time.

2) TFIEF (Term Frequency and Inverse Exponential Frequency)

Equation (3) represents the TFIEF measure [14]. This measure uses the same information used in TFIDF measure.

$$TFIEF(T_i, D_k) = TF(T_i, D_k) \times e^{-\frac{DF_i}{N}} \quad (3)$$

3) TFRF (Term Frequency and Relevance Frequency)

Equation (4) is used to determine the TFRF [15] of a term T_i in document D_k.

$$TFRF(T_i, D_k) = TF(T_i, D_k) \times \log\left(2+\frac{A}{C}\right) \quad (4)$$

Where, A and C are count of positive and negative class documents contain term T_i respectively.

4) TF-Prob

Equation (5) represents the TF-Prob measure [16].

$$TF - \Pr ob(T_i, D_k) = TF(T_i, D_k) \times \log\left(1+\frac{A}{B}\frac{A}{C}\right) \quad (5)$$

Where, B is count of documents in positive class that doesn't contain term T_i.

5) TF-MI (Term Frequency – Mutual Information)

Mutual Information is one feature selection technique which calculates the term importance in a class of documents by considering the information of the way term was distribute id positive and negative class of documents. Deng et al., presented [17] a term weight measure based on mutual information. The term T_i weight in a document is computed by using TF-MI Equation (6).

$$TF - MI(T_i, D_k) = TF(T_i, D_k) * \log\left(\frac{AN}{(A+B)\times(A+C)}\right) \quad (6)$$

6) WLLR (Weighted Log Likelihood Ratio)

The WLLR term weight technique proved its efficiency in several text classification approaches [17]. The WLLR weight of term T_i in a document D_k that belongs to either positive or negative class C is determined by using Equation (7).

$$WLLR(T_i, D_k \in C) = \frac{A}{(A+B)} \log\left(\frac{A*(N-(A+B))}{C*(A+B)}\right) \quad (7)$$

7) TF-IDF-ICSDF (Term Frequency – Inverse Document Frequency – Inverse Class Space Density Frequency)

The IDF measure says that the terms which are discussed in less documents attained good weight. Like IDF measure, ICF (Inverse Class Frequency) measure says that the terms which are discussed less number of classes attained good weight. The TF-IDF-ICSDF measure was developed in the works of [18] by combining TF, IDF and ICSDF factors. The ICSDF is a variant of ICF, which gives the average number of classes that contain the given term. The ICSDF is determined by aggregating the probabilities of documents count in individual classes. The Equation (9) is used to calculate the term T_i weight using TF-IDF-ICSDF measure.

$$TF - IDF - ICSDF(T_i, D_k) = TF(T_i, D_k) \times \left(\log\left(\frac{N}{DF(T_i)}\right)\right) \times \left(\log\left(\frac{m}{\sum_{j=1}^{m}\left(\frac{n_{cj}(T_i)}{N_{cj}}\right)}\right)\right)$$

Where, m is classes count, $n_{cj}(T_i)$ is count of documents in class C_j that contain term T_i, N_{cj} is total documents count in class C_j.

8) Proposed Term Weight Measure (PTWM)

The TWM determines the term importance in a document. The Equation (10) used to calculate the proposed term weight measure.

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

Where, A and C are number of documents that contain the term T_i in class C_j documents and other than class C_j documents respectively. B and D are number of documents that doesn't contain the term T_i in class C_j documents and other than class C_j documents respectively. $TF(T_i, D_k)$ is frequency of T_i in document D_k, $TNTD_k$ is total number of terms in document D_k. $TF(T_i, D_k \in C_j)$ is frequency of Ti in a class C_j of documents. $TF(T_i, D_k \notin C_j)$ is frequency of T_i in other than class C_j of documents.

The proposed term weight measure considers four different factors of information to determine the weight of the terms. The first factor says that the terms having more frequency in a document have more weight. The frequency of a term is normalized by dividing the term frequency with number of terms in a document.

The second factor is a ratio of number of times term T_i occurred in all documents of C_j class and all documents of other than C_j class. This factor gives more weight to the terms that are occurred more number of times in interested class and less number of times other than interested class.

The third factor becomes high when the A+D value is higher and B+C value is lower. The A+D is higher when the term T_i occurred more documents of C_j class and less number of documents of other than C_j class. B+C value is lower when the term T_i occurred more documents of C_j class and less number of documents of other than C_j class.

The fourth factor is the difference among number of documents of Cj class contains the term Ti and number of documents of other than Cj class contain the term Ti. A+B is total number of documents in Cj class and C+D is total documents count in other than Cj class. This factor says that when the term occurred in more Cj class documents than

other than Cj class documents then the difference is more and it implies the weight of Ti is more.

V. EXPERIMENTAL RESULTS

In this work, the experiment conducted with the top scored terms as features. The experiment started with top scored 2000 terms and incremented by 2000 terms in every next iteration. The experiment stopped at top scored 10000 terms. It was observed that the accuracy was dropped after experimenting with top scored 10000 terms. Two ML algorithms are used for producing the classification model. This model predicts the accuracies of age and gender prediction. The PAN 2014 competition reviews dataset was considered for experimentation. The gender dataset contains two classes such as male and female, age dataset contains five classes such as 18-24, 25-34, 35-49, 50-64 and 65+. Table 2 shows the SVM classifier accuracies of gender prediction when experiment conducted with different number of terms and various term weight measures.

TABLE 2. GENDER PREDICTION RESULTS OF SVM CLASSIFIER ON REVIEWS DATASET

Features / TWM's	TF-IDF	TFIEF	TFRF	TF-Prob	TF-MI	WLLR	TF-IDF-ICSDF	PSTWM
2000	0.6294	0.6323	0.7025	0.7136	0.7411	0.7444	0.7361	0.8053
4000	0.6461	0.6520	0.7285	0.7315	0.7532	0.7635	0.7015	0.8215
6000	0.6515	0.6723	0.7335	0.7412	0.7674	0.7797	0.7564	0.8397
8000	0.6665	0.6848	0.7535	0.7613	0.7735	0.7944	0.6414	0.8419
10000	0.6848	0.6951	0.7620	0.7787	0.7974	0.8079	0.8361	0.8537

In Table 2, the proposed TWM attained best accuracy of 85.37% for gender prediction when compared with other TWMs. It was recognized that the gender prediction accuracy was enhanced when the terms count is increased for document vector representation. Table 3 shows the SVM classifier accuracies of age prediction when experiment conducted with different number of terms and various term weight measures.

TABLE 3. AGE PREDICTION RESULTS OF SVM CLASSIFIER ON REVIEWS DATASET

Features / TWM's	TF-IDF	TFIEF	TFRF	TF-Prob	TF-MI	WLLR	TF-IDF-ICSDF	PSTWM
2000	0.6061	0.6192	0.6661	0.6797	0.6961	0.7274	0.7305	0.7401
4000	0.6117	0.6215	0.6717	0.6990	0.7031	0.7361	0.7425	0.7589
6000	0.6292	0.6423	0.6992	0.7072	0.7282	0.7461	0.7515	0.7612
8000	0.6453	0.6515	0.7053	0.7101	0.7334	0.7534	0.7691	0.7823
10000	0.6515	0.6725	0.7215	0.7321	0.7521	0.7667	0.7865	0.7963

In Table 3, the proposed TWM attained best accuracy of 79.63% for prediction of age when compared with other TWMs. It was recognized that the age prediction accuracy was enhanced when the terms count is increased for document vector representation. Table 4 shows the RF classifier accuracies of gender prediction when experiment conducted with different number of terms and various term weight measures.

TABLE 4. GENDER PREDICTION RESULTS OF RF CLASSIFIER ON REVIEWS DATASET

Features / TWM's	TF-IDF	TFIEF	TFRF	TF-Prob	TF-MI	WLLR	TF-IDF-ICSDF	PSTWM
2000	0.6502	0.6663	0.7392	0.7465	0.7592	0.7961	0.8097	0.8323
4000	0.6734	0.6779	0.7457	0.7619	0.7705	0.8011	0.8172	0.8422
6000	0.6792	0.6809	0.7607	0.7761	0.7862	0.8116	0.8365	0.8476
8000	0.6845	0.7015	0.7745	0.7903	0.8093	0.8197	0.8493	0.8635
10000	0.7061	0.7197	0.7814	0.8001	0.8141	0.8323	0.8528	0.8796

In Table 4, the proposed TWM attained best accuracy of 87.96% for prediction of gender when compared with other TWMs. It was recognized that the gender prediction accuracy was enhanced when the terms count is increased for document vector representation. Table 5 shows the RF classifier accuracies of age prediction when experiment conducted with different number of terms and various term weight measures.

In Table 5, the proposed TWM attained best accuracy of 82.58% for age prediction when compared with other TWMs. It was recognized that the age prediction accuracy was enhanced when the terms count is increased for document vector representation

TABLE 5. AGE PREDICTION RESULTS OF RF CLASSIFIER ON REVIEWS DATASET

Features / TWM's	TF-IDF	TFIEF	TFRF	TF-Prob	TF-MI	WLLR	TF-IDF-ICSDF	PSTWM
2000	0.6161	0.6455	0.6965	0.7102	0.7171	0.7400	0.7681	0.7816
4000	0.6220	0.6524	0.7066	0.7223	0.7380	0.7512	0.7793	0.7917
6000	0.6466	0.6793	0.7145	0.7448	0.7490	0.7689	0.7809	0.8092
8000	0.6592	0.6823	0.7338	0.7510	0.7529	0.7823	0.8074	0.8153
10000	0.6776	0.6961	0.7467	0.7653	0.7790	0.7951	0.8108	0.8258

VI. DISCUSSION ON RESULTS

In this work the experiment carried out with two machine learning algorithms such as SVM and RF for gender and age prediction. Different term weight measures are used to determine the importance of term in a document. The Fig. 2 shows the accuracies of SVM and RF classifiers for gender prediction.

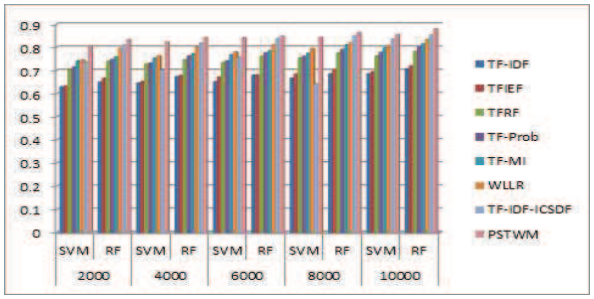

Fig. 2. The Accuracies Gender prediction

In Fig. 2, it was observed that the proposed term weight measure attained best accuracy for gender prediction when experimented with top scored 10000 terms. The RF classifier obtained good accuracies for gender prediction than the accuracies of SVM classifier. The Fig. 3 shows the accuracies of SVM and RF classifiers for gender prediction.

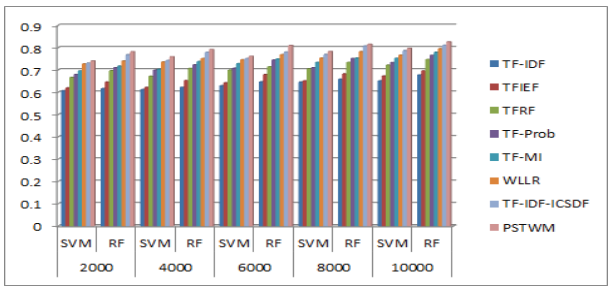

Fig. 3. The Accuracies of Age Prediction

In Fig. 3, it was observed that the proposed term weight measure attained best accuracy for age prediction when experimented with top scored 10000 terms as features. The RF classifier obtained good accuracies for age prediction than the accuracies of SVM classifier.

VII. Conclusions and Future Scope

Author profiling is defined as the task of identifying one or more attributes such as gender, age, personality traits of an author based on how they write. In this work, the age and gender was predicted from the dataset of PAN competition 2014 reviews. The term weight measure based approach was proposed to predict the author attributes of age and gender. In this approach, a new TWM was proposed to determine the weight of a term in the document vector representation. The content based features of terms that are selected by the feature selection algorithm are used as features to represent the document vectors. In vector representation, the term value is computed by using term weight measures. The experiment conducted with different TWMs and proposed TWM. Two ML algorithms such as SVM and RF are used to develop the classification model. The RF classifier performance is good than SVM classification algorithm. The proposed term weight measures attained best accuracies of 0.8796 and 0.8258 for gender and age prediction respectively than the accuracies of other TWMs when RF classifier is used. The results are good when compared with other approaches in AP for predicting gender and age of author.

In future work, we are planning to implement a new document representation technique with a combination of new feature selection technique and new term weight measure. We are also planning to implement deep learning techniques to predict the accuracies of gender and age prediction.

References

[1] Koppel M, Argamon S, Shimoni A. Automatically categorizing written texts by author gender. Literary and Linguistic Computing; 2003. p. 401–12.
[2] E. Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. JASIST.
[3] Raghunadha Reddy, T., Vishnu Vardhan, B., Vijayapal Reddy, P., A survey on author profiling techniques. Int. J. Appl. Eng. Res. 11(5), 3092–3102 (2016).
[4] 4 Muhammad Waqas Anjum Ch, Waqas Arshad Cheema, "A Study of Content Based Methods for Author Profiling in Multiple Genres ", International Journal of Scientific & Engineering Research Volume 9, Issue 9, September-2018, PP. 322-327
[5] 5 Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, Ekaterina Shutova, "Author Profiling for Hate Speech Detection", arXiv:1902.06734v1 [cs.CL] 14 Feb 2019
[6] 6 Chiyu Zhang and Muhammad Abdul-Mageed, "BERT-Based Arabic Social Media Author Profiling", arXiv:1909.04181v3 [cs.CL] 31 Oct 2019
[7] 7 Zhang, C., & Abdul-Mageed, M. (2019). BERT-Based Arabic Social Media Author Profiling. In: Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. CEUR-WS.org, Kolkata, India.
[8] 8 Inna Vogel and Peter Jiang, "Bot and Gender Identification in Twitter usingWord and Character N-Grams ", CLEF 2019, 9-12 September 2019, Lugano, Switzerland.
[9] 9 Kowsari, K., Heidarysafa, M., Odukoya, T., Potter, P., Barnes, L. E., & Brown, D. E. (2020). Gender detection on social networks using ensemble deep learning. arXiv preprint arXiv:2004.06518.
[10] 10 Moniek Nieuwenhuis and Jeroen Wilkens, "Twitter Text and Image Gender Classification with a Logistic Regression N-gram Model", Notebook for PAN at CLEF 2018.
[11] 11 https://pan.webis.de/clef14/pan14-web/author-profiling.html
[12] 12 Rehman, A. , Javed, K. , Babri, H. A. , & Saeed, M. (2015). Relative discrimination crite- rion–A novel feature ranking method for text data. Expert Systems with Applica- tions, 42 , 3670–3681.
[13] 13 Debole, F., & Sebastiani, F. (2003). Supervised term weighting for automated text categorization. In Proceedings of the 2003 ACM symposium on applied computing (pp. 784–788). Melbourne, Florida, USA.
[14] 14 Z. Tang, W.Q. Li, Y. Li, An improved term weighting scheme for text classification, Concurr. Comput.: Pract. Exper. 32 (2020) e5604.
[15] 15 M. Lan, C. Tan, J. Su, Y. Lu, Supervised and traditional term weighting methods for automatic text categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (4) (2009) 721- 735.
[16] 16 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
[17] 17 Z. Deng, K. Luo, H.Yu, A study of supervised term weighting scheme for sentiment analysis, Expert Syst. Appl. 41(2014) 3506–3513.
[18] 18 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.