JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2377 - 2391
https://publishoa.com
ISSN: 1309-3452

A Term Weight Measures based Approach for Celebrity Profiling

Siva Nagi Reddy Kalli[1], B. Narendra Kumar[2] S. Jagadeesh[3]

[13]Professor, Department of Electronics and Communication and Engineering, Sridevi Women's Engineering College, Hyderabad, Telangana, India,
[2]Professor, Department of Information and Technology, Sridevi Women's Engineering College, Hyderabad, Telangana, India,
[1]sivanagireddykalli@gmail.com,[2]narendra5346@gmail.com, [3]jaga.ssjec@gmail.com

ABSTRACT

Celebrity Profiling is a type of text classification problem which is used for predicting the profiling features like birth-year, gender, fame and occupation of celebrity authors by analysing their writing styles. PAN competition introduced this task in 2019 competition. They provided a corpus for celebrity profiling task and the corpus contains four characteristics like gender, birth-year, fame and occupation of celebrity authors. In order to differentiate the authors writing style the researchers extracted a different types of features such as style based, content based, lexical, character based, syntactic and structural features in the approaches of celebrity profiling. The researchers found that content based features play a crucial role when contrasted with other features in the identification of the author. In this work, the content based features are used in the experiment of celebrity profiling. The frequencies of terms in the total corpus are considered to recognize the important features for the experiment. The most frequent terms are used as features for representing the document vectors. The term value in the vector representation plays a vital role to enhance the performance of celebrity profiling. The Term Weight Measures (TWMs) are used for this purpose to compute the importance of a term in a document. In existing literature, various TWMs are proposed by the researchers in various research domains. In this paper, a term weight measures based approach is proposed for celebrity profiling. In this approach, a new TWM is proposed and compared the performance of proposed term weight measure with existing term weight measures. We observed from the results of celebrity profiling the proposed term weight measure attained best accuracies for profiles prediction than other term weight measures. Three Machine Learning (ML) algorithms such as Naïve Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF) are used for evaluating the performance of proposed approach. The PAN 2019 competition celebrity profiling corpus is used in this work. We considered gender, fame and occupation profiles prediction.

Keywords: Celebrity Profiling, Gender Prediction, Fame Prediction, Occupation Prediction, PAN 2019 Competition

1.INTRODUCTION

In last decade, most of the researchers are concentrated more on extracting the information of the author by analysing their written texts. The PAN is one competition organizing competitions every year on different tasks like author profiling, authorship attribution, authorship verification, author clustering, etc. In 2019, PAN organizers conducted a competition on celebrity profiling task [1]. Celebrity profiling is used for predicting the profiling characteristics like birth-year, gender, degree of fame and occupation of celebrities. PAN competition provided a corpus for celebrity profiling task and this dataset contains 48335 tweets of celebrity users written in 50 different languages. The training data contains 33836 users tweets and remaining users tweets are considered as test data. The corpus contains four profiles such as gender, birth year, fame and occupation information about celebrities. The gender trait has three sub profiles such as male, female and non-binary. The degree of fame has three sub profiles such as rising, star and super star. The occupation has eight sub profiles such as creator, performer, sports, professional, science, manager, religious and politics. The birth-year has age ranges from 1940 to 2011.

In general, the celebrities are posting their opinions about societal issues and post their personal photos in the social websites and twitter. Celebrity profiling is one interesting research area to know the characteristics of the celebrities like gender, degree of fame, occupation and birth year by analysing their written text. To know the demographic features of

the authors, the celebrity profiling used the information of changes in the writing styles of the celebrities. The fans of the celebrities are very much interested to know the demographic characteristics of the celebrities. Celebrity profiling is used in different applications like marketing, forensic analysis etc. In marketing, the celebrities are giving campaigning to the products of different companies and giving their opinions of products in the form of text in social media websites like discussion forums, blogs and twitter. The people may not aware of all the celebrities in the social media. Based on the popularity of the celebrity, the people choose the product to buy. In this context, Celebrity Profiling is used to know the details of the celebrities by analysing their written text. In forensic analysis, the cybercrime related cases like identity theft, sexual harassment messages and threatening messages are analysed by using Celebrity Profiling to detect the basic details of the perpetrator.

Generally, every author follows a specific writing style while they are writing text in blogs, forums, reviews and social media. In general, the authors never change their writing style in their life time. Stylometry is one research area to find the differences in the author style of writing. The researchers started finding the diverse types of stylistic features to distinguish the author style of writing. These stylistic features are used by various researchers in the area of author profiling for predicting features of the author by analysing their written text. The researchers analysed various datasets and expressed different types of differences in their writing style. One researcher [2] identified that the female writings contains more expressions when compared with male. The female used more number of positive and negative words in their writings. They also observed that the male narrate the new stories focused on what happened in the stories whereas the female explains how they felt.

The plethora of information decimation in the internet profoundly increases the need to develop various methods to meet the document category. The categorization or classification of documents generally aims to assign a label to documents from a set of predefined candidate class labels. The training set concentrates on the predefined labels and the testing set of document is to be assigned with label which is closely associated with the pre-defined label. Celebrity profiling is nothing but the prediction of a class label (fame, gender, birthyear and occupation) of a given document. The researchers experimented with various techniques to represent the documents. In general, the vector space model is one such method for representing the documents as vectors.

In this paper, the experiment performed with content based features of informative words. The informative words are selected based on the frequency of a word in the entire corpus. After identification of terms for representing the document vectors, the term value in vector representation is computed by using term weight measures. In this paper, a new TWM is proposed for determining the weight of a term in a document. The efficiency of proposed TWM is compared with various existing TWMs. The experimentation performed with three ML algorithms such as NB, SVM and RF to generate the classification model.

This paper is organized in 8 sections. The existing works related to celebrity profiling are described in section 2. Section 3 presents the corpus characteristics. The evaluation measures for finding the efficiency of the proposed approach are discussed in section 4. Term weight measures based approach for celebrity profiling are described in section 5. The experiment results of the gender, fame and occupation prediction are presented in section 6. The discussion of experimental results of fame, gender and occupation are presented in section 7. The section 8 concludes this work with possible future directions.

2.SURVEY ON EXISTING APPROACHES OF CELEBRITY PROFILING

In celebrity profiling, some researchers extracted stylistic features from the dataset to distinguish the authors style of writings. In [3], researchers presented a new strategy to characterize the profiles of the celebrities from twitter tweets. The proposed system follows a set of steps such as pre-processing, standardization and transformation, features extraction, configuration of classifiers and testing. In the preprocessing step, they combined tweets of one particular user into one document, substitute all the hashtags with label_hashtag, URLs with label_url, mentions with label_mention and emojis with label_emoji. A set of socio-linguistic features are generated from the tweets which served as an input to different classifiers. They extracted 18 features to represent the document vectors. The experiment conducted with different classifiers such as Complement NB, Gaussian NB, Multinomial NB, Logistic Regression (LR) and RF for predicting the accuracy of Celebrity Profiling. The Multinomial NB achieved an accuracy of 0.567 for occupation prediction. The Logistic Classifier achieved an accuracy of 0.65, 0.88 and 0.387 for frame, gender and birthyear

prediction respectively. The authors observed that Multinomial NB, Logistic Regression obtained good accuracies for predicting the traits of celebrities among all classifiers and their approach obtained 2nd rank in the competition.

In [4], authors extracted a set of features which includes supervised cross entropy, supervised lexicon extraction features, KL divergence measure and cross entropy measure, Corpus statistic features such as IR features (TF-IDF and IDF), supervised corpus statistics including gender score measure, stylistic and corpus statistic features, lexical, Bayes score. They observed that unsupervised corpus statistics were not good predictors compared to the supervised corpus statistics which are providing better accuracy for gender prediction and also identified that the stylistic and lexical features are more suitable for age prediction. The authors in [5] developed a model for the Celebrity Profiling task which is organized by PAN 2019 competition. They applied different pre-processing techniques such as removal of stopwords, punctuation marks, alphanumeric words, numbers, links / URLs, all escape characters, @, hashtag (#), brackets, spaces from the tweets to retrieve good set of words. The authors used word distance features as input to different classifiers for generating the models to predict the different profiles such as gender, fame, birthyear and occupation of Celebrity Profiling. Six different ML algorithms such as Decision Tree (DT), Gaussian Naive Bayes, LR, K- Neighbours, RF and SVC are used in the experimentation. The sklearn library was used for implementing machine learning algorithms. 80% of corpus was used to train the model and 20% of corpus was used to evaluate the model. A separate model was prepared for each language. Totally they build 200 (4 * 50) models, 50 models for each trait of a celebrity.

Matej Martinc et al., used [6] simple n-grams as features and logistic regression classifier for PAN 2019 celebrity profiling task. In their work, they are predicting the fame, gender, birthyear and occupation of celebrities in twitter. Different preprocessing techniques such as removal of punctuation marks, removal of stopwords, replace all hashtags with #HASHTAG, replace all mentions with @MENTION and replace all URLs with HTTPURL are applied in their experiment. They considered first 100 tweets of the celebrities to prepare the document if they have more than 100 tweets also and concatenated the tweets of one celebrity into one document. The authors believed that 100 tweets are sufficient for predicting author profiles and observed that the procedure of reduction of tweets decreases the space and time complexity. Three varieties of n-grams features such as word unigrams, suffix character tetragrams and word bound character tetragrams are extracted from the dataset and these features are normalized with MinMaxScaler from Scikit-learn library. The experiment performed with different classifiers such as SVM with RBF kernel, Random Forest, LR, Gradient Boosting and Linear SVM and found that Logistic Regression classifier obtained good accuracies for profiles prediction compared with other classifiers. In their observation, their approach achieved best performance in predicting the gender, worst performance for predicting birthyear and also observed that their system felt hard for predicting fame and occupation. The authors generated four classification models for four profiling traits and obtained 3rd rank in the competition.

In [7], they implemented a TF-IDF approach based on character n-grams and word bigrams for Celebrity Profiling competition conducted by PAN CLEF 2019. Different preprocessing techniques like removal of retweets, removal of special symbols other than @, #, digits and letters, substituting hyperlinks with <url>, substituting user tags with <user>, replacing multiple continuous white spaces with single white space are applied on the dataset. The experiment performed with top TF-IDF scored 10000 character n-grams (where n = 3, 4) for vector representations of tweets. The combination of SVM and LR are used for each trait prediction. The authors observed that the results of word bigrams are good when compared with the results of the character n-grams. In order to prevent over fitting problem, Linear SVM and logistic regression are replaced with multilayer perceptron.

The researchers in [8] developed a method by using TFIDF measure for extracting features and random forest classifier for generating the model. They focused mainly on pre-processing techniques wherein they implemented different text normalization methods such as URL replacing, lemmatization and emoji transformation. In their solution, they used a strategy of 10-fold cross validation for testing. They removed all handles from the twitter dataset and reduced the dimensionality of words by squeezing the multiple occurrences of same letters in a word. The URL's are replaced with url token, Unicode emojis are replaced with their corresponding descriptions which helps us for better understanding, converted all the text into lowercase and remove the accent and stop words. They removed overlapping samples by using the SMOTE sampling technique along with Tomek links. To overcome the class imbalance problem, they experimented with synthetic oversampling techniques. They experimented with word n-grams (n range is from 1 to 7) and reduced the features count from 3000 to 300. The authors observed that their approach is not achieved good results and also observed that because of more feature usage their approach consume more memory and processing time.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2377 - 2391
https://publishoa.com
ISSN: 1309-3452

In [9], they implemented a transfer learning based system which evaluated with four classifiers for predicting the traits of the authors like gender, fame, birthyear and occupation, one classifier for each trait. The classifiers are trained based on tweet-wise. Google BERT and ULMFiT are two popular approaches for transfer learning. Pelzer used ULMFiT in this experiment by considering the hardware requirements of the two approaches. ULMFiT is a pre-trained model for English which works based on Wikipedia. All four classifiers trained based on one-cycle-policy, which is recommended by ULMFiT. They obtained accuracies of 0.39 for fame, 0.51 for occupation, 0.68 for gender and 0.32 for birthyear.

3.CORPUS CHARACTERISTICS

The PAN competition creates an arena of identifying the researchers to participate in various text mining areas. The organizers of PAN competitions will select research area initially and prepare the training and testing data sets and made it available to the participants. Initially only a training data set is given. After a model is built, the test sample corpus is available for the testing. The participants build their models and then test samples are applied on the model and the result were notified to the PAN organizers, there by the best result based model will be selected for the award of the prize. The organizers will give various inherent features and the training and testing data.

In this paper, the corpus was taken from PAN 2019 competition Celebrity Profiling track [1]. The training data of the corpus contains English tweets with the author details of fame, gender and occupation. The corpus consists user profiles of 48835 users tweets with an average of 2181 tweets per user. The corpus characteristics are displayed in Table 1.

Table 1. Corpus Properties

Profile Name	Sub Profile	Number of Tweets
Degree of Fame	Star	25230
	Rising	1490
	Superstar	7116
Gender	Male	24221
	Female	9583
	Nonbinary	32
Occupation	Creator	5475
	Manager	768
	Performer	9899
	Politics	2835
	Professional	525
	Science	818
	Sport	13481
	Religious	35

The corpus was not balanced. In case of fame, huge proportion of user profiles is stars, whereas the frequency of superstars and rising is very low when compared with star. Similarly, more than 50% profiles are of male celebrities, whereas, only 32 users belongs to nonbinary. In this work, the nonbinary sub profile of gender is not considered in the experiment. Same is the case with occupation, where there are sufficient instances of sports, performer and creator, whereas remaining categories are in minority.

4. PERFORMANCE MEASURES

The conventional performance measures like precision, recall, accuracy and F1-measure was used for the evaluating the effectiveness of celebrity profiling approaches. The contingency table for a category Ci is represented in the table 2 where YES or NO represents the binary decision of existence and nonexistence for each document under the category Ci. Table 2 shows the contingency table for class Ci.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2377 - 2391
https://publishoa.com
ISSN: 1309-3452

Table 2. Contingency table

Category Ci		Original labels of documents	
		YES	NO
Label by System	YES	TPi (A)	FPi (B)
	NO	FNi (C)	TNi (D)

Where, TPi (A) isTrue Positives which represent for a given YES category Ci the no of documents predicted as YES, FPi (B) is False Positives which represent for a given YES category Ci the no of documents predicted as NO, TNi (C) is True Negatives which represent for a given NO category Ci the no of documents predicted as YES, FNi (D) is False negatives which represent for a given NO category Ci the no of documents predicted as NO. The information retrieval based performance measure Precision is represented in equation (1) and recall is represented in equation (2).

$$\Pr ecision = \frac{TP_i}{TP_i + FP_i} \tag{1}$$

$$\mathrm{Re}\, call = \frac{TP_i}{TP_i + FN_i} \tag{2}$$

Recall represents the no. of relevant documents retrieved from the correct set of documents and precision represents the no. of relevant documents retrieved and are actually correct. The Accuracy is represented in equation (3) and error is defined in equation (4).

$$Accuracy = \frac{TP_i + TN_i}{TP_i + FP_i + FN_i + TN_i} \tag{3}$$

$$Error = \frac{FP_i + FN_i}{TP_i + FP_i + FN_i + TN_i} \tag{4}$$

Another information retrieval based performance measure F1-measure is also used in authorship attribution technique. F1-measure is shown in equation (5).

$$F_1 = \frac{2 \times \Pr ecision \times \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \tag{5}$$

In this work, we used accuracy measure to test the accuracy of the classifiers. In the context of celebrity profiling, the accuracy is the fraction of test documents that are correctly predicted their author details like gender, fame and occupation.

5.TERM WEIGHT MEASURES BASED APPROACH FOR CELEBRITY PROFILING

In this paper, a term weight measures based approach is proposed for celebrity profiling. The Fig. 1 shows the procedure followed in proposed approach.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2377 - 2391
https://publishoa.com
ISSN: 1309-3452

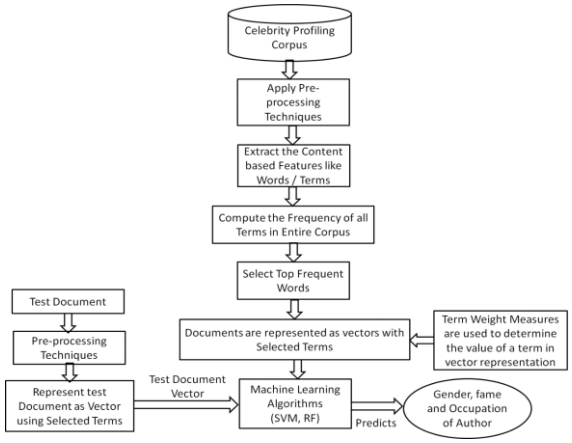

Fig. 1. Term Weight Measures based Approach for Celebrity Profiling

In this approach, first step is cleaning the corpus of celebrity profiling. Different pre-processing techniques such as lowercase conversion, punctuation removal, tokenization, stop-words elimination and stemming are implemented on the dataset to remove uninteresting data from the corpus. After cleaning the dataset, the next step is extracting all the terms and calculates the each term frequency in the entire corpus. Select the terms as features that are having more frequency in the dataset as features. The features are used for document vector representation. The value of a term in vector representation is determined by using term weight measures. The vectors are passed to ML algorithms to generate the model. This model is used for predicting the fame, gender and occupation of a test document.

In this paper, we proposed a new TWM to determine the term value in the vector representation. The next sections explain the existing TWMs and proposed TWM.

5.1 Existing Term Weight Measures

The TWMs determine the weight of a term in a document. Researchers proposed various TWMs in different research domains by considering the information of the way the term is distributed in a document, distributed in positive and negative class of documents.

5.1.1. Term Frequency (TF)

The TF measure a simple and popular term weight measure. The TF measure gives more weight to the terms that are occurred more times in a document. The length of a document influences the TF value of a term [10]. Equation (6) is used to normalize the TF value for a term irrespective of document length.

$$TF\left(T_i, D_k\right) = \frac{TF_{ik}}{\sqrt{\sum_{i=1}^{n} TF_{ik}^{\,2}}} \tag{6}$$

Where, TF_{ik} is the count T_i term in a document D_k, n is total terms count considered for experiment.

5.1.2. Term Frequency-Inverse Document Frequency (TFIDF)

The IDF measure allots more weight to the terms that occurred at least one time in less number of documents [10]. TFIDF is used in several research domains. Equation (7) is used to calculate the TFIDF of a term in a document

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2377 - 2391
https://publishoa.com
ISSN: 1309-3452

irrespective of length of a document.

$$TFIDF\left(T_i, D_k\right) = \frac{TF_{ik} \times \log\left(\frac{N}{DF_i}\right)}{\sqrt{\sum_{i=1}^{n}\left(TF_{ik} \times \log\left(\frac{N}{DF_i}\right)\right)^2}} \quad (7)$$

Where, N is documents count in corpus and DF_i is documents count in dataset that contain term T_i.

5.1.3. Term Frequency-Relevance Frequency (TF-RF) Measure

TF-RF measure computes the weight of a term based on its Relevance Frequency (RF), which is the ratio among the term frequency in the positive class of documents (A) and the term frequency in the negative class of documents (B) [11]. TF-RF measure assigned more weight to the terms which are specific to a class because A >> B. The basic plan of TF-RF measure is the terms which are occurred in more positive class of documents when compared with negative class of documents were more useful to select positive text from a negative text. Equation (8) is used to compute the TFRF value of a term.

$$TFRF\left(T_i, D_k\right) = TF_{ik} \ * \ \log\left(2 + \frac{A}{MAX(1, B)}\right) \quad (8)$$

Some researchers observed that the efficiency of TF-RF measure is better than most traditional and STW measures. However, in multi class text classification problems the TF-RF measure allocates weight to the terms by considering one class as positive class and groups all other classes into a single negative class.

5.1.4. TF-PROB Term Weight Measure

TF-PROB measure is based on the probability information of terms. The TF-PROB measure is a combination of A/B and A/C [12]. Equation (9) is used to compute TF-PROB of a term.

$$TF-PROB\left(T_i, D_k\right) = TF_{ik} \times \log\left(1 + \frac{A}{B} * \frac{A}{C}\right) \quad (9)$$

TF-RF measure includes only the term's inter-class distribution and is represented by A and B. But TF-PROB measure includes the intra-class distribution and inter-class distribution of a given term, represented by A and C. The reason for introducing the intra-class distribution in TF-PROB measure is that the terms which were appeared in most of documents in a positive class i.e., A >> C obtained good weight to represent the positive class.

5.1.5. TF-IDF-ICSDF

The TFIDF is popularly used in various text classification problems like authorship analysis, sentiment analysis, etc., to calculate weight of a term. The TFIDF allocates higher weight to terms that are discussed in fewer documents. Several researchers proposed TWMs based on the TFIDF measure by replacing IDF with other weight metric factor like TFRF (Term Frequency and Relevance Frequency), TFCHI2 (Term Frequency and Chi square), TFIG (Term Frequency and Information gain) etc. Some researchers enhanced the TFIDF with additional weight factor like TF-IDF-ICF, TF-IDF-ICSDF etc. to increase the performance of term weight process. In this paper, TF-IDF-ICSDF measure is used to calculate the term weight.

Ren and Sohrab's developed [13] a TF-IDF-ICSDF measure by adding a collection frequency factor of ICSDF (Inverse Class Space Density Frequency) to the TF-IDF measure. ICSDF consider the term distributions across inter

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2377 - 2391
https://publishoa.com
ISSN: 1309-3452

class documents while computing the weight values of terms. Equation (10) is used to calculate the term T_i weight by using TF-IDF-ICSDF measure.

$$TF-IDF-ICSDF\left(T_i,D_k\right)=TF_{ik}\times\left(1+\log\left(\frac{|D|}{DF_i}\right)\right)\times\left(1+\log\left(\frac{|C|}{\sum_{j=1}^{C}\frac{DF_{ij}}{D_j}}\right)\right) \quad (10)$$

Where, TF_{ik} is the count of term T_i in document D_k, |D| is documents count in dataset, DF_i is documents count which contain term T_i in dataset, |C| is classes count, DF_{ij} is count of class C_j documents contain term T_i, D_j is count of documents if class C_j.

5.2. Proposed Term Weight Measure (PTWM)

In this paper, a new TWM is proposed to determine the weight of a term in a document. The Equation (11) is used to compute the weight of a term in a document using proposed term weight measure.

$$PTWM\left(T_i,D_k\right)=\frac{TF\left(T_i,D_k\right)}{TC_k}\times\log\left(\frac{|D|}{A+C}\right)*\log\left(\frac{A*D}{B*C}\right)*\log\left(\frac{|C|}{cf}\right) \quad (11)$$

Where, TC_k is count of terms in a document D_k, |D| is count of documents in corpus, |C| is count of classes in a profile.The proposed term weight measure is a combination of four factors. First factor determines the frequency of a term and it is normalized with total terms in a document. Second factor gives more importance to the terms that are discussed in fewer documents in the corpus. Third factor assigns more weight to the terms which are occurred in more documents of positive class than negative class documents. The fourth factor gives more weight to the terms that are discussed in fewer classes.

6. EXPERIMENTAL RESULTS

In this work, the experiment was conducted for predicting the profiles like fame, gender and occupation of celebrity authors. Two classification algorithms such as SVM and RF are used for evaluating the performance of our proposed term weight measures based approach for celebrity profiling. Next sections explain the experimental results of different profiles prediction.

6.1. Experimental results of Gender Prediction

The Table 3 shows the prediction accuracies of gender when experiment conducted with various number of terms and NB classifier.

Table 3. The accuracies of Gender Prediction when NB classifier is used

TWMs / Frequent Terms	TF	TF-IDF	TF-RF	TF-PROB	TF-IDF-ICSDF	PTWM
1000	54.86	56.49	60.73	62.23	62.98	65.74
2000	55.58	57.04	61.28	62.98	63.43	66.75
3000	55.92	57.72	61.63	63.48	64.25	67.39
4000	56.31	58.29	62.25	63.98	64.68	67.69

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2377 - 2391
https://publishoa.com
ISSN: 1309-3452

5000	56.61	58.34	62.46	64.28	65.01	67.94
6000	56.97	59.28	62.94	64.43	65.56	68.42
7000	57.71	59.37	63.16	65.18	66.38	68.72
8000	58.94	60.85	64.14	65.31	66.93	70.18

In Table 3, the proposed TWM attained best accuracy of 70.18% for gender prediction than other TWMs. The Table 4 shows the prediction accuracies of gender when experiment conducted with various number of terms and SVM classifier.

Table 4. The accuracies of Gender Prediction when SVM classifier is used

TWMs / Frequent Terms	TF	TF-IDF	TF-RF	TF-PROB	TF-IDF-ICSDF	PTWM
1000	57.23	59.16	64.43	66.10	67.14	70.21
2000	57.95	59.71	65.01	66.85	67.59	71.17
3000	58.27	60.39	65.36	67.35	68.41	71.86
4000	58.68	60.87	65.98	67.85	68.84	72.16
5000	58.98	61.01	66.19	68.15	69.17	72.41
6000	59.34	61.95	66.67	68.30	69.72	72.89
7000	60.07	62.04	66.89	68.97	70.54	73.19
8000	61.31	63.52	67.87	69.18	71.06	74.65

In Table 4, the proposed TWM attained best accuracy of 74.65% for gender prediction than other TWMs. The Table 5 shows the prediction accuracies of gender when experiment conducted with different number of terms and RF classifier.

Table 5. The Gender Prediction accuracies when RF classifier is used

TWMs / Frequent Terms	TF	TF-IDF	TF-RF	TF-PROB	TF-IDF-ICSDF	PTWM
1000	60.30	63.21	67.01	68.10	69.14	74.21
2000	60.45	63.53	67.43	68.85	69.59	74.86
3000	61.23	63.89	68.36	69.15	70.41	75.17
4000	61.87	64.08	68.98	69.47	70.87	75.76
5000	62.20	64.91	69.19	69.86	70.17	75.91
6000	62.50	65.19	69.67	70.30	70.72	76.23

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2377 - 2391
https://publishoa.com
ISSN: 1309-3452

7000	62.79	65.76	69.85	70.97	71.54	76.59
8000	63.24	66.85	70.52	71.89	72.45	77.68

In Table 5, the proposed TWM attained best accuracy of 77.68% for gender prediction when experiment conducted with most frequent 8000 terms and the performance of proposed TWM is good when compared with other TWMs.

6.2. Experimental results of Fame Prediction

The Table 6 displays the accuracies of fame prediction when experiment conducted with different number of frequent terms and NB classifier.

Table 6. The accuracies of Fame Prediction when NB classifier is used

Term Weight Measures / Frequent Terms	TF	TF-IDF	TF-RF	TF-PROB	TF-IDF-ICSDF	PTWM
1000	64.77	67.91	70.66	72.37	73.04	77.25
2000	65.12	68.34	71.08	73.36	74.05	77.46
3000	66.01	68.65	71.46	73.67	74.52	77.86
4000	66.22	69.22	72.23	74.02	75.11	78.14
5000	66.63	69.54	72.55	74.32	75.45	78.78
6000	66.69	69.93	72.91	74.7	75.78	79.21
7000	67.38	70.32	73.42	75.76	76.21	80.12
8000	67.86	71.77	75.14	76.27	77.43	80.94

In Table 6, the proposed TWM attained best accuracy of 80.94% for fame prediction when experiment conducted with most frequent 8000 terms and the performance of proposed term weight measure is good than other TWMS. The Table 7 displays the accuracies of fame prediction when experiment conducted with different number of frequent terms and SVM classifier.

Table 7. The accuracies of Fame Prediction when SVM classifier is used

Term Weight Measures / Frequent Terms	TF	TF-IDF	TF-RF	TF-PROB	TF-IDF-ICSDF	PTWM
1000	68.14	70.82	74.39	76.17	77.20	81.67
2000	68.47	71.21	74.81	77.23	78.21	81.93
3000	69.38	71.56	75.19	77.54	78.68	82.27

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2377 - 2391
https://publishoa.com
ISSN: 1309-3452

4000	69.59	72.13	75.73	77.89	79.27	82.61
5000	69.97	72.45	76.28	78.19	79.61	83.25
6000	70.06	72.84	76.64	78.57	79.94	83.68
7000	70.75	73.23	77.15	79.63	80.37	84.59
8000	71.23	74.61	78.87	80.14	81.59	85.41

In Table 7, the proposed TWM attained best accuracy of 85.41% for fame prediction when experiment conducted with most frequent 8000 terms and the performance of proposed term weight measure is good than other TWMS. The Table 8 shows the accuracies of fame prediction when experiment conducted with different number of terms and RF classifier.

Table 8. The accuracies of Fame Prediction when RF classifier is used

TWMs / Frequent Terms	TF	TF-IDF	TF-RF	TF-PROB	TF-IDF-ICSDF	PTWM
1000	70.61	73.15	76.80	77.16	79.89	83.82
2000	71.08	73.36	77.15	77.71	80.32	84.16
3000	71.46	74.23	78.05	78.39	81.33	84.87
4000	71.77	74.76	79.01	78.87	81.85	85.04
5000	72.13	74.91	79.25	79.01	82.44	85.53
6000	72.62	75.94	79.85	80.95	83.31	85.94
7000	72.87	76.38	80.27	81.41	83.87	86.25
8000	73.42	77.49	81.21	82.47	84.86	87.14

In Table 8, the PTWM attained best accuracy of 87.14% for fame prediction when experiment conducted with most frequent 8000 terms and the performance of PTWM is good when compared with other TWMs.

6.3. Experimental results of Occupation Prediction

The Table 9 displays the accuracies of occupation prediction when experiment conducted with different number of frequent terms and NB classifier.

Table 9. The accuracies of Occupation Prediction when NB classifier is used

Term Weight Measures / Frequent Terms	TF	TF-IDF	TF-RF	TF-PROB	TF-IDF-ICSDF	PTWM
1000	65.82	72.3	74.83	75.49	76.79	78.24

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2377 - 2391
https://publishoa.com
ISSN: 1309-3452

2000	66.01	72.46	74.92	75.91	77.42	78.9
3000	67.05	73.1	75.94	76.49	77.5	79.26
4000	67.48	74.03	76.18	77.54	78.43	79.95
5000	68.14	74.48	76.57	77.92	79.17	80.81
6000	68.7	74.56	76.92	78.3	79.45	81.17
7000	69.22	75	77.95	78.93	80.49	81.93
8000	70.16	75.51	78.95	79.56	81.11	82.88

In Table 9, the proposed TWM attained best accuracy of 82.88% for fame prediction when experiment conducted with most frequent 8000 terms and the performance of proposed term weight measure is good than other TWMS. The Table 10 displays the accuracies of occupation prediction when experiment conducted with different number of frequent terms and SVM classifier.

Table 10. The accuracies of Occupation Prediction when SVM classifier is used

Term Weight Measures / Frequent Terms	TF	TF-IDF	TF-RF	TF-PROB	TF-IDF-ICSDF	PTWM
1000	69.19	75.21	78.56	79.36	80.95	82.71
2000	69.38	75.37	78.65	79.78	81.58	83.37
3000	70.42	76.01	79.67	80.36	81.66	83.73
4000	70.85	76.94	79.91	81.41	82.59	84.42
5000	71.51	77.39	80.30	81.79	83.33	85.28
6000	72.07	77.47	80.65	82.17	83.61	85.64
7000	72.59	77.91	81.68	82.80	84.65	86.40
8000	73.53	78.42	82.68	83.43	85.27	87.35

In Table 10, the PTWM attained best accuracy of 87.35% for occupation prediction when experiment conducted with most frequent 8000 terms and the performance of proposed term weight measure is good than other TWMs. The Table 11 shows the accuracies of occupation prediction when experiment conducted with different number of terms and RF classifier.

Table 11. The accuracies of Occupation Prediction when RF classifier is used

TWMs / Frequent Terms	TF	TF-IDF	TF-RF	TF-PROB	TF-IDF-ICSDF	PTWM

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2377 - 2391
https://publishoa.com
ISSN: 1309-3452

1000	71.29	75.39	79.70	80.07	82.30	85.07
2000	71.84	75.81	80.14	80.34	82.94	85.34
3000	72.73	76.26	80.36	81.24	83.36	86.24
4000	73.18	76.83	80.79	81.72	83.79	86.82
5000	73.56	77.51	81.85	82.17	84.85	87.27
6000	74.37	77.80	82.20	82.64	85.20	87.64
7000	64.89	78.39	82.94	83.25	85.74	88.25
8000	75.43	79.28	83.96	84.49	86.27	89.72

In Table 11, the PTWM attained best accuracy of 89.72% for occupation prediction when experiment conducted with most frequent 8000 terms and the performance of PTWM is good when compared with other TWMs.

7. DISCUSSION OF RESULTS

Fig. 2 displays the accuracies of gender prediction when experiment conducted with different TWMs and different classifiers.

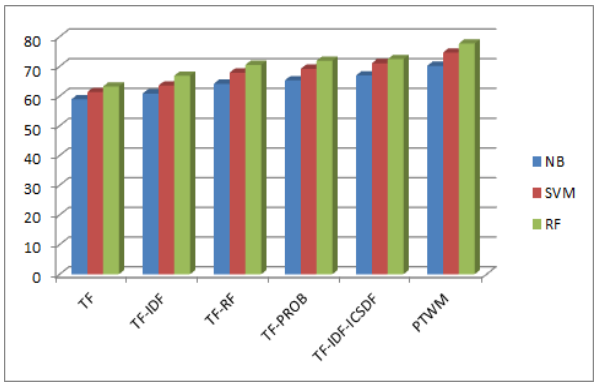

Fig. 2. The Gender Prediction Accuracies for various classifiers

In Fig. 2, it was observed that the PTWM attained best accuracy of 77.68% for gender prediction when RF classifier is used. The RF classifier shows best performance than the performance of NB and SVM classifiers. Fig. 3 shows the accuracies of fame prediction when experiment conducted with different TWMs and different classifiers.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 2377 - 2391
https://publishoa.com
ISSN: 1309-3452

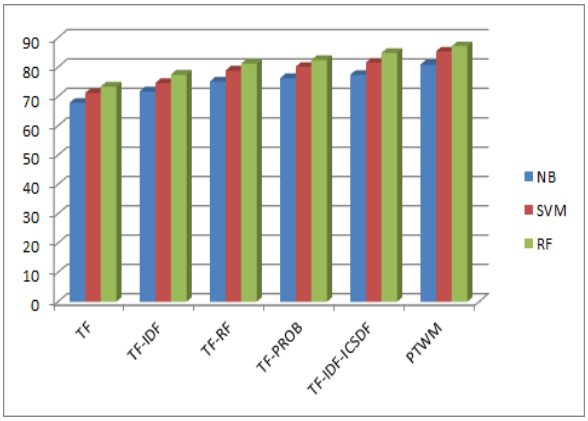

Fig. 3. The Accuracies of Fame Prediction for various classifiers

In Fig. 3, it was identified that the PTWM attained best accuracy of 87.14% for fame prediction when RF classifier is used. The RF classifier shows best performance than the performance of NB and SVM classifiers. Fig. 4 displays the accuracies of occupation prediction when experiment conducted with different TWMs and different classifiers.

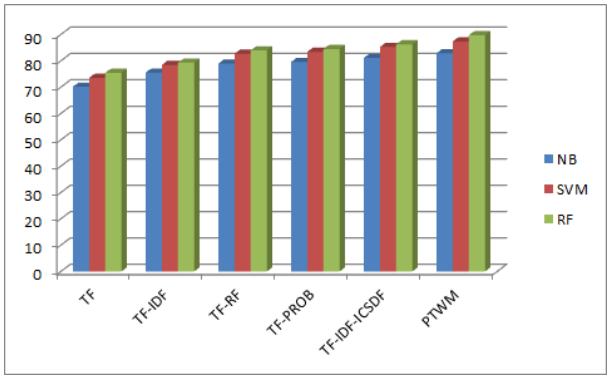

Fig. 4. The Accuracies of Occupation Prediction for various classifiers

In Fig. 4, it was identified that the PTWM attained best accuracy of 89.72% for occupation prediction when RF classifier is used. The RF classifier shows best performance than the performance of NB and SVM classifiers.

8. CONCLUSION AND FUTURE SCOPE

The celebrity profiling is a technique of predicting the characteristics like gender, age, fame and occupation of celebrity authors by analysing their texts. In this work, a TWMs based approach was proposed for celebrity profiling. In this approach, a new TWM was proposed and compared the performance of PTWM with existing term weight measures. Three celebrity characteristics such as gender, fame and occupation are considered in this experiment. Three ML algorithms such as RF, SVM and RF are used to evaluate the efficiency of proposed approach and predicting the performance of gender, fame and occupation. The proposed term weight measure attained best accuracy of 77.68%, 87.14% and 89.72% for gender, fame and occupation prediction respectively when RF classifier is used.

In future work, we are planning to implement a BERT based deep learning technique to reduce the feature engineering process as well as to improve the accuracies of celebrity profiling.

REFERENCES

1. https://pan.webis.de/clef19/pan19-web/celebrity-profiling.html

2. J. Schler, Moshe Koppel, S. Argamon and J. Pennebaker (2006), Effects of Age and Gender on Blogging, in Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, March 2006. Vol. 6, (2006), 199-205.
3. Luis Gabriel Moreno-Sandoval, Edwin Puertas, Flor Miriam Plaza-del-Arco, Alexandra Pomares-Quimbaya, Jorge Andres Alvarado-Valencia, and L.Alfonso Ureña-L\`opez. Celebrity Profiling on Twitter using Sociolinguistic Features—Notebook for PAN at CLEF 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.
4. Maria De-Arteaga, Sergio Jimenez, George Duenas, Sergio Mancera, and Julia Baquero. Author Profiling Using Corpus Statistics, Lexicons and Stylistic Features—Notebook for PAN at CLEF 2013.
5. Muhammad Usman Asif, Naeem Shahzad, Zeeshan Ramzan, and Fahad Najib. Word Distance Approach for Celebrity profiling—Notebook for PAN at CLEF 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.
6. Matej Martinc, Bla\vz \vSkrlj, and Senja Pollak. Who is hot and who is not? Profiling celebs on Twitter—Notebook for PAN at CLEF 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.
7. Victor Radivchev, Alex Nikolov, and Alexandrina Lambova. Celebrity Profiling using TF-IDF, Logistic Regression, and SVM—Notebook for PAN at CLEF 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.
8. Juraj Petrik and Daniela Chuda. Twitter feeds profiling with TF-IDF—Notebook for PAN at CLEF 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.
9. Björn Pelzer. Celebrity Profiling with Transfer Learning—Notebook for PAN at CLEF 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.
10. G. Salton, A. Wong, C. Yang, A vector space model for automatic indexing, Communications of the ACM 18 (11) (1975) 613- 620.
11. 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.
12. 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
13. Ren, F., & Sohrab, M. G. (2013). Class-indexing-based term weighting for automatic text classification. Information Sciences, 236, 109–125.
14. Yogesh Hole et al 2019 J. Phys.: Conf. Ser. 1362 012121