JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Word Embedding Techniques based Approach for Celebrity Profiling

[1]Kottu Divya Jyothi, [2]Karunakar Kavuri, [3]Dr. P Pandarinath

[1]PG-Student, [2]Associate Professor, [3] Professor& Principal
[1, 2, 3] Department of Computer Science and Engineering,
[1, 2, 3] Swarnandhra Institute of Engineering and Technology, Narasapur, India.

Abstract: The celebrity profiling is a variety of author profiling, which is introduced to detect the profiles like gender, birthyear, fame, occupation, etc., of authors by analysing the textual content of authors. Most of celebrities are using different social media platforms to convey their interests and their plans to keep in touch with their followers and well-wishers. Some intruders are trying to imitate the celebrities and post false information in the social media platforms. In recent times, experts of authorship analysis concentrated on whether the text is posted by the celebrities or not and knowing the profiling characteristics of celebrity authors. In this aspect, PAN competition introduced the task of celebrity profiling in 2019 to predict the gender, birthyear, fame and occupation of celebrity authors based on analysis of their text. Several researchers participated in this competition and submitted their works by using different stylistic features, machine learning algorithms and deep learning algorithms. In this work, we proposed word embedding techniques based approach for celebrity profiling. In the proposed approach, we used two word embedding techniques such as Word2Vec and FastText for representing words as vectors. The most important part in success of text classification approaches is representation of text most effectively. The word embedding techniques are efficiently representing the words vectors. These word vectors are used in representation of a document. Every document is represented as a vector by aggregating the word vectors of words in that document. The document vectors are trained with the three machine learning algorithms such as Naïve Bayes Multinomial, Support Vector Machine, and Random Forest. In this work, we concentrated on the fame and occupation of celebrity authors and accuracy measure is used to present the results of proposed approach. The proposed approach with combination of Random Forest and Word2Vec attained best accuracies for Fame and Occupation prediction.

Index Terms **- Celebrity Profiling, Fame detection, Occupation Detection, Word Embeddings, Machine Learning Algorithms**

I. INTRODUCTION

In general, most of the celebrities are posting their opinions about societal issues and post their personal photos in the social media platforms like twitter and Instagram. The fans of the celebrities are very much interested to know the demographic characteristics of the celebrities. Celebrity Profiling is one interesting research area to know the characteristics of the celebrities like gender, degree of fame, occupation and birthyear by analysing their written text. The task of Celebrity Profiling was organized by PAN competition in the year 2019 [1]. This task is introduced for predicting the profiling traits such as gender, birthyear, degree of fame, and occupation of celebrity authors. The gender trait has three sub profiles such as male, female and nonbinary. The birtyear has age ranges from 1940 to 2011. 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 dataset provided in the competition contains 48335 tweets of celebrity users written in 50 different languages. The training data contains 33836 celebrities' tweets and remaining celebrity tweets are considered as test data.

To know the demographic characteristics of the authors, the celebrity profiling is used the changes in the celebrities writing styles. In 2013, PAN competition organizers introduced a new research area named Author Profiling in 2013 [2]. Author Profiling is a technique to predict gender, age, location, educational background and nativity language of the authors by analysing their writing skills. Celebrity Profiling is a variety of Author Profiling. 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.

Celebrity Profiling is used in various 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.

Every author follows certain 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 writing styles of the authors. The researchers started finding the different types of stylistic features like character based, word based, structural,

syntactic and semantic features to differentiate the writing styles of the authors. These stylistic features are used by the most of the researchers in the area of author profiling to predict the gender, age group, location, educational background and nativity language of the author by analysing their written text. The researchers analysed various datasets and expressed different types of differences in their writing style. Most of the researchers related to domain of celebrity profiling used stylistic features to differentiate the writing styles of celebrities.

In this work, we used content based features like the informative words in the text. In traditional approaches the informative words are used directly to represent the document vectors. In these approaches, the importance of a word is not properly considered for representing the documents. After introduction of word embedding techniques, the word representation considers the contextualized information to represent the words as vectors. We developed an approach for celebrity profiling by using the word vectors generated by the word embedding techniques. Two word embedding techniques such as Word2Vec and FastText are used for representing words as vectors. The documents are represented as vectors by aggregating word vectors of words that are contained in that document. These document vectors are trained with three machine learning algorithms such as Naïve Bayes Multinomial, Support Vector Machine, and Random Forest for predicting the fame and occupation of celebrity authors.

This paper is planned in 6 sections. Section 2 discusses about different existing works proposed for the domain of celebrity profiling. The dataset characteristics are presented and explained in section 3. The proposed approach with the components used in the proposed approach like word embedding techniques and machine learning algorithms are described in section 4. The empirical evaluations of proposed approach for fame and occupation prediction are discussed in section 5. The section 6 concludes this paper with future enhancements to improve accuracy of celebrity profiling.

II. Existing Solutions for Celebrity Profiling

Most of the research works were used a set of stylistic features to differentiate the writing styles of the authors in celebrity profiling and authorship profiling. Maria De-Arteaga et al. [3] extracted a set of features which includes supervised lexicon extraction features, supervised cross entropy, KL divergence measure and cross entropy measure, supervised corpus statistics including gender score measure, Corpus statistic features such as IR features (TF-IDF and IDF), stylistic and corpus statistic features, lexical, bayes score and detected that unsupervised corpus statistics were not good predictors compared to the supervised corpus statistics which are providing better accuracy for gender prediction and also observed that for age prediction the lexical and stylistic features are more suitable.

In [4], they presented a new strategy to characterize the profiles of the celebrities from twitter tweets. A set of socio-linguistic features are generated from the tweets which served as an input to different classifiers. The proposed system follows a set of steps such as pre-processing, standardization and transformation, features extraction, configuration of classifiers and testing. The experiment conducted with different classifiers such as Logistic Regression, Complement NB, Gaussian NB, Multinomial NB, and Random Forest for predicting the accuracy of Celebrity Profiling. The authors observed that Multinomial NB, Logistic Regression obtained good accuracies for predicting the traits of celebrities among all classifiers. The Logistic Classifier achieved an accuracy of 0.65, 0.88 and 0.387 for frame, gender and birthyear prediction respectively. The Multinomial NB achieved an accuracy of 0.567 for occupation prediction. They extracted 18 features to represent the document vectors. The experiment performed with bag of words model with most frequent words at least six times occurrence in the corpus. In the pre-processing 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. They experimented with different types of stylistic features and their approach obtained 2nd rank in the competition.

In [5], the authors proposed a method which uses TFIDF measure for extracting features and random forest classifier is used to generate the model. They focused mainly on pre-processing techniques wherein they implemented different text normalization methods such as URL replacing, lemmatization and emoji transformation. To overcome the class imbalance problem, they experimented with synthetic oversampling techniques. They experimented on the corpus of celebrity profiling task introduced by PAN 2019 competition. In their solution, they used 10-fold cross validation as a strategy 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 Synth CSOB etic Minority Oversampling Technique (SMOTE) along with Tomek links. They experimented with word n-grams (n range from 1 to 7) and reduced the number of features in the range from 3000 to 30000. 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.

In [6], researchers developed a model for the task of Celebrity Profiling which is organized by PAN 2019 competition. The corpus of Celebrity Profiling task contains tweets of 33,836 celebrities with 50 different languages. The task is to predict the sub profiles of fame (rising, star, and superstar), sub profile of gender (male, female, and nonbinary), sub profile of occupation (performer, sports, creator, manager, professional, science, religious, politics) and birthyear (1940-2011) of celebrity. 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 machine learning algorithms such as Decision Tree, Gaussian Naive Bayes, Logistic Regression, K- Neighbours, Random Forest 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. For example a model for gender prediction for English language considers English tweets for generating the model. This model is used to predict the gender of a new tweet which is in English language only. They applied different preprocessing techniques such as removal of stopwords, punctuation marks, alphanumeric words, numbers, links / URLs, all escape characters, @, hashtag (#), brackets, and spaces from the tweets to retrieve good set of words.

In [7], authors used simple n-grams as features and logistic regression classifier. They experimented on the corpus provided for PAN 2019 celebrity profiling task. In their work, they are predicting the fame, gender, birthyear and occupation of celebrities in twitter. 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. 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 to predict the profiles of the authors and observed that the procedure of reduction of tweets decreases the space and time complexity. 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. 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, Logistic Regression, Gradient Boosting and Linear SVM and found that Logistic Regression classifier obtained good accuracies for profiles prediction compared with other classifiers.

In [8], authors implemented TF-IDF approach based on character n-grams and word bigrams for Celebrity Profiling competition conducted by PAN CLEF 2019. The task is finding the author traits of fame, gender, birthyear and occupation for a given tweet. The dataset contains 3 sub classes for fame and gender, eight sub classes for occupation and birthyear ranges from 1940 to 2012. 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. They identified top 10000 word bigrams based on TF-IDF for representing vectors of tweets. The combination of SVM and logistic regression are used for each trait prediction. The experiment performed with top TF-IDF scored 10000 character n-grams (where n = 3, 4) for vector representations of tweets. 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.

In [9], researchers 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.

III. Dataset Description

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. In this work, 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. In this work, we concentrated on prediction of fame and occupation of authors. The dataset characteristics are displayed in Table 1.

Table 1: Properties of Dataset

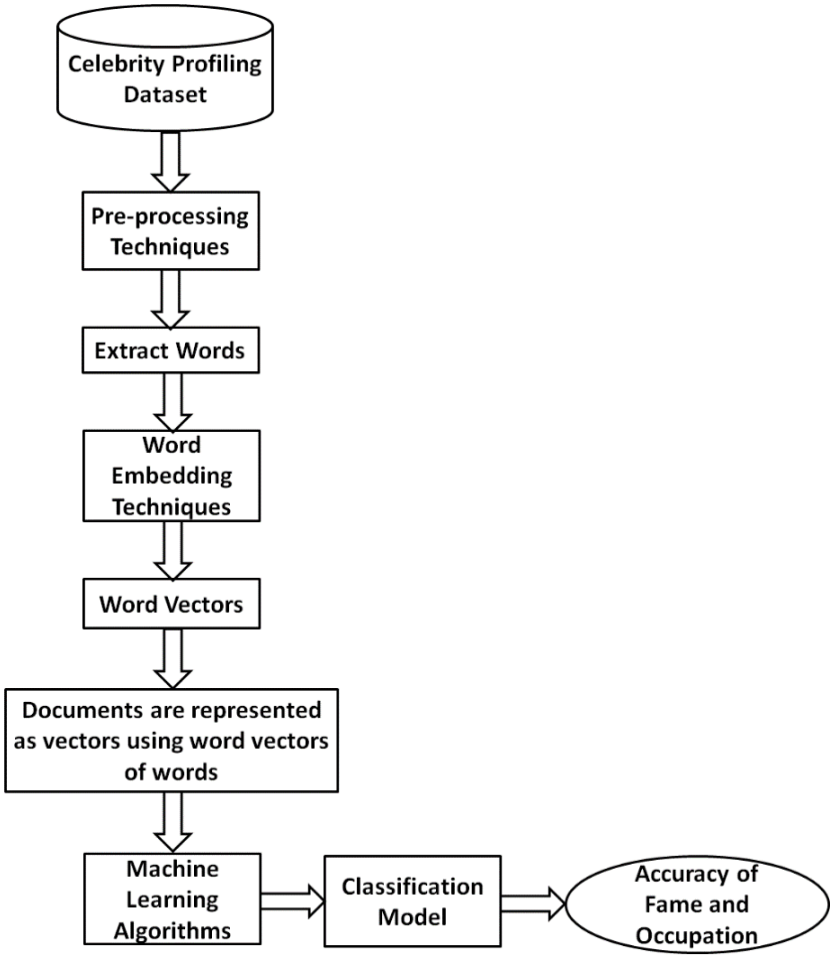
Profile Name	Sub Profile	Number of Tweets
Degree of Fame	Star	25230
	Rising	1490
	Superstar	7116
Occupation	Creator	5475
	Manager	768
	Performer	9899
	Politics	2835
	Professional	525
	Science	818
	Sport	13481
	Religious	35

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, same is the case with occupation, where there are sufficient instances of sports, performer and creator, whereas remaining categories are in minority.

IV. Proposed Approach for Celebrity Profiling

In this work, we proposed a word embedding techniques based approach for celebrity profiling to predict the fame and occupation of authors. The proposed approach is represented in Figure 1.

In the proposed approach, first, we apply suitable pre-processing techniques like punctuation marks removal, stop words removal, and lemmatization to remove the irrelevant and unwanted information from the dataset. After cleaning the dataset, extract all words from the dataset. Forward these words to word embedding techniques to generate the word vectors. These word vectors are used to represent the documents as vectors. The documents vectors are used to train the machine learning algorithms. These algorithms internally generates classification model which is used for predicting the accuracy of fame and occupation prediction.



7

Figure 1: The Proposed Approach

4.1. Word Embeddings

Natural language processing has gained popularity in both research and business over the past years. The increase of computing power enables one to process text on a large scale, quantifying millions of words within hours. Language modelling by the quantification of text allows users to feed natural language as input to statistical models and machine learning techniques. A popular approach to quantify text is to represent each word in the vocabulary by a vector filled with real-valued numbers, called a word embedding. The numbers in these word embeddings represent scores of latent linguistic characteristics. The trained word embeddings capture similarities of words in text data. In this work, we used two word embedding techniques such as Word2Vec and FastText for generation word vectors for words.

4.1.1 Word2Vec Model

A word embedding is based on the words surrounding the word of interest. These surrounding words are the so-called context words. The estimation procedure of Word2Vec is based on the prediction power of words to predict other words in the neighbourhood, the so-called local context window, of those words in the text [10]. Word2Vec model has two variants such as CBOW (Continuous Bag Of Words) and Skip-Gram model. The CBOW variant, on the other hand, compares each word with the average representation of the surrounding words and the word in the center of the context window is predicted given its context. The process of CBOW model is represented in Figure 2.

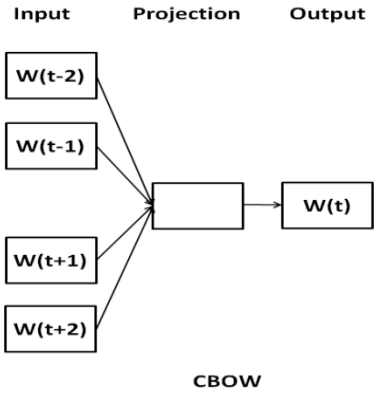


Figure 2: CBOW Model

The Skip-Gram variant of word2vec uses word by word similarity comparisons and tries to predict a word in the local context window given the word in the middle of this context window. The process of Skip-Gram model is represented in Figure 3.

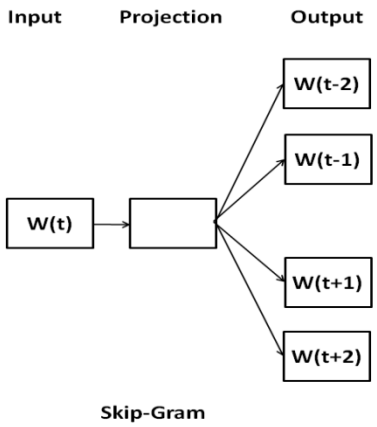


Figure 3: Skip-Gram Model

In this work, we used the skip-gram model for generating word vectors.

4.1.2 FastText

One shortcoming of the Word2Vec framework is that it only assigns distinct embeddings to words. In cases of dataset with many rare words, this may result in missing embeddings for words that have not been in the training corpus (Bojanowski et al., 2017) [11]. To bypass this limitation by providing word embeddings for a high percentage of before unseen words, the concept of implementing the prediction tasks additionally at the level of constituent character n-grams for each word is realized. This technique was developed as an extension of the Word2Vec framework, called FastText (Bojanowski et al., 2017) [11].

The representations of words are learned by summing up the embeddings learnt for the constituent n-grams and the word itself. If one considers the word 'embedding' and character n-grams of length 3, the resulting sequence on which sub-word character ngram embeddings are learnt is made up of <em, emb, mbe, bed, edd, ddi, din, ing, ng>. '<' and '>' are treated as special boundary symbols to delimit the word boundaries. Once a before unseen word is encountered, the FastText framework creates its word embedding by averaging its constituent sub-word embeddings. There by providing a reasonable alternative to assigning 0-vectors or random numbers as word embeddings for these words with the drawback that words may incorporate the same constituent sub-word n-grams without being semantically related.

4.2. Machine Learning Algorithms

The machine learning algorithms are used to evaluate the performance of proposed approaches. These algorithms display the performance by using different performance evaluation measures. In this work, three machine learning algorithms such as Naive Bayes Multinomial (NBM), Support Vector Machine (SVM), and Random Forest (RF) are used to evaluate the performance of our proposed approach.

4.2.1 Naive Bayes Multinomial (NBM)

The Naive Bayes Classifier (NBC) is a probabilistic classifier that applies the Bayes theorem while presuming the features are independent [12]. NBC has two parts that are training and testing. In the training part, NBC creates a model with the training data using features/predictors. In the testing part, NBC tests the classification model using testing data. Some of the NBCs are Gaussian NBC, Multinomial NBC and Bernoulli NBC. Depending on the type of the data to classify, different types of NBC's are used. For continuous data Gaussian NBC was used and for discrete data Multinomial NBC was used. Multinomial NBC is widely used in text classification. With Multinomial NBC, the documents are represented as a histogram of words and document d is represented as a word frequency list.

4.2.2 Random Forest (RF)

Decision Tree (DT) is usually used to handle classification tasks. It is a powerful method in ML as it works for both categorical and continuous dependent variables, which is remarkable. Using DT, the population is divided into two or more homogeneous sets. To generate as many distinctive sets as possible, the most meaningful attribute's variables are used [13]. RF and ET are extended versions of DT. RF is a DT approach but an ensemble method, which is because it is made up of a collection of DTs known as a forest. In this algorithm, each tree contributes a classification to a new object based on characteristics, which represents the votes for that class. These trees grow based on the number of samples. If the training set has N instances, a random sample of N instances is taken with replacement. This sample set is going to be the training set that contributes to growing the tree. It is important that there is no pruning in RF, meaning that each tree is grown to the greatest degree feasible [14].

4.2.3 Support Vector Machine (SVM)

This algorithm is used widely for classification tasks. To perform training with SVM, the number of features n in a set of data is calculated for plotting each data item as a point in n-dimensional space, and each feature's value is the coordinate's value. This helps to define a line (separation boundaries or hyperplane) that divides the points into distinctive groups in regions, which are then classified differently. A hyperplane is defined by the distance between two points (called support vectors). If the two closest points are at the furthest distance from a line, then it is the classifier line. The space between the line and each of these points is called the margin [15].

V. Empirical Evaluations

In this work, the experiment carried out for predicting the performance of proposed approach for fame and occupation prediction. The accuracy measure is used for representing the performance of proposed approach.

5.1. Evaluation Measures

The researchers used different evaluation measures like precision, recall, accuracy and F1-score to evaluate the efficiency of the approaches proposed to the Celebrity Profiling. The contingency table for a class C_i is represented in the table 2.

Table 2: Contingency table

Class C_i		**Original labels of documents**	
		Original YES	**Original NO**
Predicted by the system	**Predicted YES**	TP_i (True Positives)	FP_i (False Positives)
	Predicted NO	FN_i (False negatives)	TN_i (True Negatives)

In table 2, TP_i is the number of YES label documents are predicted as YES by the system, FP_i is the number of NO label documents are predicted as YES by the system, TN_i is the number of NO label documents are predicted as NO by the system, FN_i is the number of YES label documents are predicted as NO by the system.

Accuracy is the ratio between number of test documents are predicted correctly and total considered test documents. Accuracy is represented in Equation (1).

$$Accuracy = \frac{TP_i + TN_i}{TP_i + FP_i + FN_i + TN_i} \quad (3)$$

5.2. Accuracies for Fame Detection

In this experiment, the document vectors are represented with the word vectors generated by the word embedding techniques of Word2Vec and FastText. Three classifiers such as NBM, SVM and RF are used to train the model by using these vectors. The Table 3 presents the accuracies of the proposed approach for fame prediction.

Table 3: The accuracies of proposed approach for Fame prediction

Machine Learning Algorithms / Word Embedding Techniques	**NBM**	**SVM**	**RF**
Word2Vec	47.08	60.83	88.33
FastText	42.9	56.66	88.33

In Table 3, the random forest classifier attained best accuracies for fame prediction when compared with the accuracies of other two classifiers such as NBM and SVM. The Word2Vec obtained good accuracies for fame prediction when compared with the accuracies of FastText technique. The random forest classifier with Word2Vec attained best accuracy of 88.33 for fame prediction.

5.3. Accuracies for Occupation Detection

The Table 4 presents the accuracies of the proposed approach for occupation prediction.

Table 4: The accuracies of proposed approach for Occupation prediction

Machine Learning Algorithms / Word Embedding Techniques	**NBM**	**SVM**	**RF**
Word2Vec	90.4	90.8	99.1
FastText	79.58	85.4	97.5

In Table 4, the random forest classifier attained best accuracies for occupation prediction when compared with the accuracies of other two classifiers such as NBM and SVM. The Word2Vec obtained good accuracies for occupation prediction when compared with the accuracies of FastText technique. The random forest classifier with Word2Vec attained best accuracy of 99.1 for occupation prediction.

VI. Conclusion and Future Scope

The Celebrity Profiling is a type of Author Profiling technique to predict the demographic characteristics like gender, fame, birthyear and occupation of the celebrities by analyzing their written texts. The experiment carried on 2019 PAN Competition task of Celebrity Profiling. In this work, we proposed a word embedding techniques based approach for celebrity profiling. In the proposed approach we used two popular word embedding techniques of Word2Vec and FastText for representing informative words as vectors. These word vectors are used for representing each document in the dataset as vector. Three machine learning algorithms are used for evaluating the proposed approach and presenting the accuracies of fame and occupation prediction. The Random Forest classifier with Word2Vec attained best accuracies of 88.33 and 99.1 for fame and occupation prediction respectively.

In future work, we are planning to implement Gated recurrent Unit as classifiers to predict the accuracy of fame and occupation prediction. We are also planning to implement BERT and Glove word embedding techniques for fame and occupation prediction of celebrities.

.

REFERENCES

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

[2] Rangel Pardo, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain. CEUR-WS.org (Sep 2013)

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[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] T. MIKOLOV, I. SUTSKEVER, K. CHEN, G. S. CORRADO, AND J. DEAN, Distributed representations of words and phrases and their compositionality, in Advances in neural information processing systems, 2013, pp. 3111–3119.

[11] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017). Enriching word vectors with subword information, Transactions of the association for computational linguistics 5: 135–146.

[12] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "Profile specific Document Weighted approach using a New Term Weighting Measure for Author Profiling ", International Journal of Intelligent Engineering and Systems, 9 (4), pp. 136-146, Nov 2016.

[13] J. Ali, R. Khan, N. Ahmad, and I. Maqsood. Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5):272, 2012.

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

[15] V. Vapnik. The nature of statistical learning theory. Springer science & business media, 2013.

[16] Karunakar Kavuri, Kavitha, M. (2020). "A Stylistic Features Based Approach for Author Profiling". In: Sharma, H., Pundir, A., Yadav, N., Sharma, A., Das, S. (eds) Recent Trends in Communication and Intelligent Systems. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-0426-6_20.

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

[18] Karunakar. Kavuri and M. Kavitha, "A Term Weight Measure based Approach for Author Profiling," 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), 2022, pp. 275-280, doi: 10.1109/ICESIC53714.2022.9783526.