Journal of the Maharaja Sayajirao University of Baroda
ISSN :0025-0422

ANALYSIS OF CELEBRITY PROFILING WITH WORD EMBEDDING TECHNIQUES

Padmavathi V, M.Tech-Student, Department of Computer Science and Engineering, Swarnandhra College of Engineering and Technology, Narasapur, India.
G Archana, Assistant Professor, Department of Computer Science and Engineering, Swarnandhra College of Engineering and Technology, Narasapur, India.
K. Rajesh Kumar, Associate Professor, Department of Computer Science and Engineering, Swarnandhra College of Engineering and Technology, Narasapur, India.
Dr. P Pandarinath, Professor, Department of Computer Science and Engineering, Swarnandhra College of Engineering and Technology, Narasapur, India.

Abstract: The Celebrity profiling is a variety of author profiling, which involves textual content analysis to identify an author's profile, such as gender, birth year, fame, and occupation. The majority of celebrities express their interests and plans to their fans and well-wishers using several social media platforms. 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. We used word embedding techniques Word2Vec for representing words as vectors. 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 Support Vector Machine, Random Forest, Convolutional Neural Networks (CNN), Long short-term memory (LSTM). In this work, we concentrated on the fame and occupation of celebrity authors. The proposed approach CNN and Word2Vec attained best accuracies for Fame and Occupation prediction.

Keywords **- Celebrity Profiling, Fame detection, Occupation Detection, Word Embeddings**

1. INTRODUCTION

On social media sites like Instagram and Twitter, celebrities typically share their personal images along with their thoughts on current events and society concerns. The demographic attributes of superstars pique the keen interest of their audience. Celebrity profiling is an intriguing field of study that uses text analysis to determine a celebrity's gender, level of renown, profession, and birth year. In 2019, the PAN competition organised the Celebrity Profiling assignment [1]. The purpose of this work is to forecast the profile attributes of famous writers, including their gender, birth year, level of renown, and profession. There are three subprofiles for the gender trait: nonbinary, female, and male. The age range for the birth year is 1940–2011. There are three sub-profiles for the degree of fame: rising, star, and super star. There are eight sub-profiles for this occupation: creator, performer, sports, professional, science, manager, religious and politics. The competition provides a dataset with 48335 celebrity user tweets written in 50 different languages. 33836 celebrity tweets make up the training data; the remaining celebrity tweets are regarded as test data.

Changes in the writing styles of celebrities are utilised in celebrity profiling to determine the demographic features of the authors. The organisers of the PAN competition unveiled Author Profiling as a new field of study in 2013 [2]. Author profiling is a technique that uses an analysis of an author's writing skills to forecast the author's gender, age, location, educational background, and nativity language. One type of author profiling is celebrity profiling. The necessity to develop different approaches to meet the document category is greatly increased by the abundance of information available on the internet. Assigning a label to documents from a list of established candidate class labels is the overall goal of document categorization and classification.

Applications for celebrity profiling include marketing and forensic investigation, among others. In the field of marketing, celebrities endorse items from various companies and provide textual reviews of those things on social media platforms such as blogs, Twitter, and discussion forums. It's possible that some people are unaware of every celebrity on social media. People select the product to purchase based on the celebrity's popularity. Celebrity profiling is used in this context to examine written texts by celebrities to learn more about them. In forensic analysis, celebrity profiling is used to examine cybercrime-related incidents such as identity theft, sexual harassment messages, and threatening messages in order to identify the perpetrator's fundamental information.

Each author adheres to a specific writing style while creating content for social media, blogs, forums, and reviews. Generally speaking, writers never alter their style of writing throughout their lives. One field of study that look into author variances in writing style is stylistometry. To distinguish between the writers' writing styles, the researchers began identifying several stylistic elements, such as word-based, character-based, structural, syntactic, and semantic features. Most scholars in the field of author profiling employ these stylistic characteristics to analyse written texts and predict the author's gender, age group, geography, educational background, and nativity language. The researchers examined a range of datasets and noted various variations in the authors' styles. The majority of researchers in the field of celebrity profiling distinguished between celebrities' writing styles using stylistic characteristics.

We employed content-based elements in this work, such as the text's informative terms. In conventional methods, the document vectors are directly represented by the informative words. These methods do not appropriately take into account a word's value when representing the texts. Following the development of word embedding techniques, contextualised information is taken into account by word representation, which represents words as vectors. Using the word vectors produced by word embedding techniques, we created a method for celebrity profiling. Words are represented as vectors via the embedding method Word2Vec. Word vectors that are contained in the documents are aggregated to represent the documents as vectors. The Support Vector Machine, Random Forest, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) are used to train these document vectors in order to predict the level of fame and occupation of famous writers.

There are six sections in this study. The various efforts that have been offered for the field of celebrity profiling are discussed in Section 2. In section 3, the features of the dataset are outlined and clarified. Section 4 describes the suggested approach and its components, which include machine learning algorithms and word embedding techniques. Section 5 discusses the empirical evaluations of the suggested method for predicting fame and occupation. This paper's section 6 offers suggestions for future improvements to raise the accuracy of celebrity profiling.

2. EXISTING SOLUTIONS FOR CELEBRITY PROFILING

In order to distinguish between the writing styles of the writers in authorship and celebrity profiling, the majority of research studies employed a set of stylistic characteristics. Using supervised lexicon extraction features, supervised cross entropy, KL divergence measure, and cross entropy measure, as well as supervised corpus statistics including gender score measure and corpus statistic features like IR features (TF-IDF and IDF), stylistic and corpus statistic features, lexical, and bayes score, Maria De-Arteaga et al. [3] extracted a set of features and found that unsupervised corpus statistics were not good predictors in comparison to supervised corpus statistics, which are more accurate in predicting gender. They also noted that lexical and stylistic features are more appropriate for age prediction.

They introduced a novel method in [4] for identifying celebrity profiles from tweets. The tweets produce a set of sociolinguistic characteristics that are fed into several classifiers. Pre-processing, standardisation and transformation, feature extraction, classifier design, and testing are some of the phases that the suggested system goes through. The experiment was done to predict the accuracy of celebrity profiling using a variety of classifiers, including Random Forest, Gaussian NB, Complement NB, Multinomial NB, and Logistic Regression. Among all classifiers, Multinomial NB, Logistic Regression demonstrated good accuracy in predicting celebrity attributes, according to the authors' observations. For the predictions of frame, gender, and birth year, the logistic classifier had

an accuracy of 0.65, 0.88, and 0.387, respectively. For occupation prediction, the Multinomial NB attained an accuracy of 0.567. To represent the document vectors, they retrieved eighteen features. The most common words in the corpus appeared at least six times in the bag of words model used for the experiment. They merged a single user's tweets into a single document during the pre-processing stage, replacing all hashtags with label_hashtag, URLs with label_url, mentions with label_mention, and emojis with label_emoji. They experimented with many artistic elements, and their strategy helped them place second in the competition.

The authors of [5] presented a technique that generates the model using a random forest classifier and extracts features using the TFIDF measure. Their primary focus was on pre-processing techniques, wherein they employed various text normalisation techniques such emoji transformation, lemmatization, and URL replacement. They experimented with artificial oversampling methods in an attempt to address the issue of class imbalance. They conducted experiments using the celebrity profiling corpus that the PAN 2019 competition had introduced. They employed 10-fold cross validation as a testing approach in their solution. By compressing the repeated occurrences of the same letter in a word, they were able to minimise the dimensionality of words and eliminate all handles from the Twitter dataset. The URLs are changed to url tokens, Unicode emojis are changed to descriptions that aid in comprehension, all text is changed to lowercase, accents and stop words are eliminated, and all text is converted to lowercase. They used Tomek connections in conjunction with the Synth CSOB etic Minority Oversampling Technique (SMOTE) to eliminate overlapping sounds. They decreased the number of characteristics in the range from 3000 to 30000 and experimented with word n-grams (n range from 1 to 7). The authors noted that their method did not yield satisfactory results and that it required more memory and processing time due to the increased feature usage.

A model was created by researchers in [6] for the Celebrity Profiling challenge, which is part of the PAN 2019 competition. The 33,836 celebrities who tweet in 50 different languages are included in the corpus of the Celebrity Profiling challenge. Predicting a celebrity's birth year (1940–2011), subprofiles of popularity (rising, star, and superstar), gender (male, female, and nonbinary), and occupation (performance, sports, creator, management, professional, scientist, religious, and political) is the challenge at hand. The authors created models to predict various profiles, including gender, celebrity, birth year, and occupation of celebrity profiling, by using word distance information as input to various classifiers. The experimentation uses six different machine learning methods, including Gaussian Naive Bayes, Decision Tree, Logistic Regression, K-Nearbors, Random Forest, and SVC. Machine learning techniques were implemented using the sklearn library. Twenty percent of the corpus was utilised for model evaluation and eighty percent was used for model training. Each language has its own model that has been prepared. They construct 200 (4 * 50) models in total, 50 models for each characteristic of a star. For instance, a gender prediction model for the English language is generated by taking into account English tweets. 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. The gender of a new tweet that is entirely in English is predicted using this model. To extract a good collection of terms from the tweets, they used a variety of preprocessing approaches, including removing stopwords, punctuation marks, alphanumeric words, numerals, links / URLs, all escape characters, @, hashtag (#), brackets, and spaces.

Authors in [7] employed a logistic regression classifier and basic n-grams as features. Using the corpus provided for the PAN 2019 celebrity profiling problem, they conducted experiments. They use their study to forecast celebrities' fame, gender, birth year, and employment on Twitter. According to their observations, their method performed best at guessing gender and poorly at identifying birth year. They also noted that their technique felt unreliable when it came to forecasting

career and celebrity. The authors placed third in the competition after creating four classification models for four profiling traits. In order to compile the paper, they took into account the first 100 tweets posted by the celebrities. If the celebrities had more than 100 tweets, they combined their tweets into a single document. The authors saw that the process of reducing the number of tweets lowers the complexity of space and time, and they felt that 100 tweets would be enough to anticipate their own profiles. In their experiment, various preprocessing techniques are used, including removing stopwords and punctuation, replacing all mentions with @MENTION, all URLs with HTTPURL, and all hashtags with #HASHTAG. The dataset is used to extract three different types of n-grams features: word unigrams, suffix character tetragrams, and word bound character tetragrams. These features are then normalised using the Scikit-learn library's MinMaxScaler. In the experiment, several classifiers were used, including Linear SVM, Random Forest, Gradient Boosting, Logistic Regression, and SVM with RBF kernel. It was discovered that the Logistic Regression classifier produced better profile prediction accuracy than the other classifiers.

In [8], authors used the TF-IDF technique for the PAN CLEF 2019 Celebrity Profiling Competition, which was based on character n-grams and word bigrams. The objective is to determine the renown, gender, birth year, and occupation of the author of a given tweet. The dataset includes eight subclasses for occupation, three subclasses for renown and gender, and birthyear spans from 1940 to 2012. Various preprocessing procedures are used to the dataset, such as removing retweets, removing special symbols other than @, #, numbers, and letters, replacing hyperlinks with , replacing user tags with , and replacing numerous continuous white spaces with a single white space. Based on TF-IDF, they determined the top 10,000 word bigrams to use as tweet vector representations. For each trait prediction, a combination of logistic regression and SVM is employed. For vector representations of tweets, the experiment using the top TF-IDF scored 10,000 character n-grams (where n = 3, 4). When compared to the outcomes of character n-grams, the writers noted that word bigrams produced good results. Multilayer perceptrons are used in place of logistic regression and linear SVM to avoid the overfitting issue.

In [9], scientists put into practice a transfer learning-based approach that was assessed using four classifiers, one for each attribute, to predict the characteristics of the writers, including gender, notoriety, birth year, and occupation. Twitter is used as the basis for training the classifiers. ULMFiT and Google BERT are two well-liked methods for transfer learning. In this experiment, Pelzer employed ULMFiT by taking into account the hardware specifications of the two methods. Wikipedia serves as the foundation for the pre-trained English model known as ULMFiT. ULMFiT recommends using a one-cycle policy for training all four classifiers. For renown, occupation, gender, and birth year, they had accuracy scores of 0.39, 0.51, 0.68, and 0.32, respectively.

3. DATASET DESCRIPTION

The PAN competition generates a platform for finding investigators to work in different text mining fields. The PAN competition organisers will first choose the study field, create the training and testing data sets, and make them available to the competitors. The corpus used in this study came from the Celebrity Profiling track of the PAN 2019 competition [1]. The English tweets in the corpus' training data include the author's fame, gender, and employment. The corpus has 48835 user profiles and their average number of tweets, which is 2181 per person. In this paper, we focused on predicting writers' fame and occupation. Table 1 presents the attributes of the dataset.

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 failed consistency. When it comes to fame, stars make up a large percentage of user accounts, but the frequency of superstars and rising stars is extremely low in comparison to stars. In the same way, there are lots of instances of sports, performer and creator in work, whereas the other categories are underrepresented.

4. PROPOSED APPROACH FOR CELEBRITY PROFILING

In this paper, we suggested a celebrity profile method based on word embedding techniques to predict authors' fame and occupation. The suggested methodology is depicted in Figure 1.

The suggested method starts by using appropriate pre-processing techniques to eliminate unnecessary and irrelevant data from the dataset, such as stop word removal, lemmatization, and punctuation mark removal. Extract every word from the dataset after it has been cleaned. To create the word vectors, pass these words through word embedding techniques. The documents are represented as vectors by means of these word vectors. The learning algorithms are trained using the vectors from the documents. The categorization model that these algorithms internally produce is used to forecast the accuracy of the fame and occupation predictions.

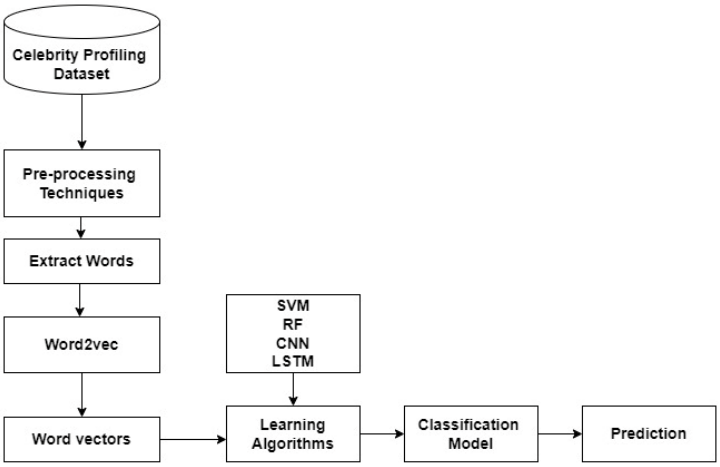


Figure 1: The Proposed Approach

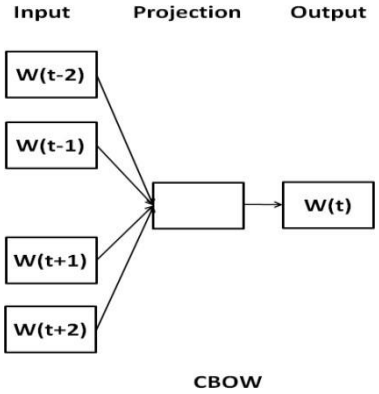
4.1. Word Embeddings

Over the past few years, natural language processing has become more and more popular in both industry and research. With the growth in processing capacity, millions of words can be quantified in a matter of hours through large-scale text processing. Users can feed natural language into statistical models and machine learning approaches by quantifying text and modelling language. One common method for quantifying text is to use a vector called a word embedding, which is made up of real-valued values, to represent each word in the lexicon. These word embeddings' numerical values correspond to scores of latent linguistic features. Word similarities in text data are captured by the learned word embeddings. In this work, we generated word vectors for words using the Word2Vec

word embedding technique.

4.1.1 Word2Vec Model

The words that surround the term of interest serve as the basis for a word embedding. The words that surround these are known as context words. The Word2Vec estimation process is predicated on a word's ability to anticipate additional words in its immediate vicinity, or what is known as its "local context window," in a given text [10]. There are two variations of the Word2Vec model: the Skip-Gram model and the Continuous Bag of Words (CBOW). In contrast, the CBOW variation predicts



the word in the centre of the context window based on its context by comparing each word with the average representation of the surrounding words. Figure 2 depicts the CBOW model's procedure.

Figure 2: CBOW Model

Using word-by-word similarity comparisons, the Skip-Gram variation of word2vec predicts a word in the local context window given the word in the middle of this context window. Figure 3 depicts the Skip-Gram model's operation. In this work, we used the skip-gram model for generating word vectors.

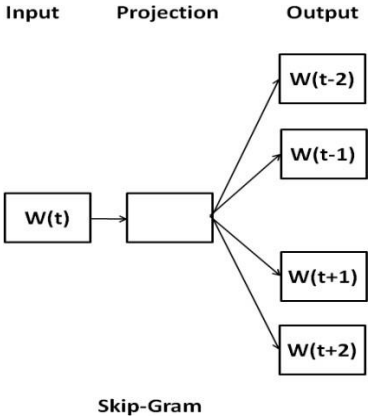


Figure 3: Skip-Gram Model

4.2. Learning Algorithms

The effectiveness of the suggested strategies is assessed using the learning algorithms. These methods use many performance evaluation metrics to display the performance. The performance of our suggested method is assessed in this work using four learning algorithms: Random Forest, Support Vector Machine, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM).

4.2.1 Support Vector Machine (SVM)

A lot of tasks involve classification use this approach. In order to plot each data item as a point in n-dimensional space and execute SVM training, the number of features (n) in the data set is determined. The value of each feature is equal to the coordinate of that data item. This aids in the definition of a line (separation boundaries or hyperplane) that separates the points in regions into distinguishable groupings that are subsequently assigned unique classifications. The distance between two locations (referred to as support vectors) defines a hyperplane. A line is the classifier line if the two nearest points are the ones that are farthest from it. The margin is the distance [15] between the line and each of these places.

4.2.2 Random Forest (RF)

Classification jobs are typically handled using Decision Trees (DT). It's amazing that this powerful machine learning method can be applied to both continuous and categorical dependent variables. The population is split up into two or more homogeneous sets using DT. The variables associated with the most significant attribute are used to produce as many unique sets as possible [13]. Extended forms of DT are RF and ET. Because RF is composed of a group of DTs known as a forest, it is a DT approach but an ensemble method. Each tree in this process provides a classification, or the votes for that class, to a new item based on its features. The number of samples determines how big these trees go. A random sample of N instances is taken with replacement if there are N occurrences in the training set. The training set that aids in the tree's growth will be this sample set. It is crucial that every tree in RF be grown to the maximum extent possible without any pruning [14].

4.2.3 Long Short-Term Memory (LSTM)

Recurrent neural networks (RNNs) include Long Short-Term Memory (LSTM). Text, audio, and time series are examples of sequential data that LSTM can process and analyse. They regulate the information flow with gates and a memory cell. An LSTM network is made up of a sequence of LSTM cells, each of which has a set of input, output, and forget gates to regulate the information entering and leaving the cell. The LSTM can preserve long-term dependencies in the input data by using the gates to selectively forget or keep information from earlier time steps.

4.2.4 Convolutional Neural Networks (CNN)

Multi-layered artificial neural networks called convolution neural networks (CNN) can recognise complicated elements in data. Three different kinds of layers are present in a typical neural network:

Input layer: This layer is where we feed data into our model. The entire number of characteristics in our data is equal to the number of neurons in this layer.

Hidden Layer: The hidden layer receives the input that was previously given into the input layer. Several hidden layers may exist, contingent on the amount of the data and our model. The number of neurons in each hidden layer varies, but they are usually more than the number of characteristics. The network is nonlinear because each layer's output is calculated by multiplying the output of the layer before it by the learnable weights of that layer, adding learnable biases, and then applying the activation function.

Output Layer: The probability score for each class is obtained by feeding the output of the hidden layer into a logistic function such as the sigmoid or softmax.

5. EMPIRICAL EVALUATIONS

This study used an experiment to forecast how well the suggested method for forecasting fame and occupation will perform. The suggested approach's performance is represented by the accuracy measure.

5.1. Evaluation Measures

A variety of evaluation metrics, including precision, recall, accuracy, and F1-score, were employed by the researchers to assess the effectiveness of the methods suggested for celebrity profiling. Table 2 displays the contingency table for a class Ci.

Table 2: Contingency table

Class Ci		**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)

Table 2 displays the number of YES label documents that the system predicts to be YES (TPi), the number of NO label documents that the system predicts to be YES (FPi), the number of NO label documents that the system predicts to be NO (TNi), and the number of YES label documents that the system predicts to be NO (FNi).

The ratio of accurately predicted test documents to the total number of test documents considered is known as accuracy. Equation (1) represents accuracy.

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

5.2. Accuracies for Fame Detection

In this experiment, Word2Vec's word embedding technique produces word vectors that are used to represent the document vectors. These vectors are utilised to train the model using four classifiers: SVM, RF, CNN, and LSTM. The suggested method's accuracy in predicting celebrity is shown in Table 3.

Table 3: The accuracies of proposed approach for Fame prediction

Learning Algorithms / Word Embedding Technique	SVM	RF	CNN	LSTM
Word2Vec	53.3	54.6	90.8	90

When compared to the accuracies of other classifiers, such as LSTM, SVM and RF, Table 3 shows that the CNN classifiers achieved the greatest results for fame prediction. For predicting renown, the CNN classifier using Word2Vec had the highest accuracy of 90.8.

5.3. Accuracies for Occupation Detection

The accuracy of the suggested method for occupation prediction is shown in Table 4.

Table 4: The accuracies of proposed approach for Occupation prediction

Learning Algorithms / Word Embedding Technique	SVM	RF	CNN	LSTM
Word2Vec	90	90.3	96.6	96.2

Comparing the occupation prediction accuracy of the CNN classifiers to that of the LSTM, SVM, and RF classifiers, Table 4 shows that the CNN classifiers achieved the highest accuracy. The CNN classifier using Word2Vec achieved the highest occupation prediction accuracy of 96.6.

6. CONCLUSION AND FUTURE SCOPE

Celebrity profiling is a form of author profiling technique that uses text analysis to predict the demographic traits of celebrities, such as gender, renown, birth year, and occupation. The experiment focused on the Celebrity Profiling task from the 2019 PAN Competition. In this paper, we suggested a celebrity profiling method based on word embedding techniques. In the suggested method, we represented instructive words as vectors using the well-liked Word2Vec word embedding technique. Every document in the collection is represented as a vector using these word vectors. The suggested method is assessed using learning algorithms, which are also utilised to display the accuracy of the fame and occupation predictions. For the purposes of predicting renown and occupation, the Word2Vec CNN classifier had the highest accuracy rates of 90.8 and 96.6, respectively.

We intend to use Gated Recurrent Units as classifiers in future work to estimate the precision of renown and occupation predictions. In order to forecast celebrities' renown and occupation, we also want to use the BERT and Glove word embedding algorithms.

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, Transactionsof 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.