IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

A Survey on Author Profiling Techniques

Vivitha Vijayan[1*], Sharvari Govilkar[2]

[1,2] Department of Computer Engineering, Pillai College of Engineering, Mumbai University, New Panvel, Maharashtra, India

[*]*Corresponding Author: vivitha.vijayan@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i3.10651069 | Available online at: www.ijcseonline.org

Accepted: 19/Mar/2019, Published: 31/Mar/2019

***Abstract*—** In this information age, Internet is growing exponentially with large amount of information through social media networks like Facebook, Blogs, Twitter, LinkedIn, etc. Most of the text we have seen in Internet are anonymous in nature. Analysis of such kind of text became crucial nowadays. Author Profiling is a technique which is used to analyse the anonymous text in Internet for finding out the characteristics of author like age, gender, country, native language, educational background etc. Style of writing of each author is utilized for the analysis of different characteristics of author's profile. Researchers experimented with different types of features to improve the accuracy of prediction. The final accuracy of prediction depends on the feature which is extracted and on the machine learning algorithm used for prediction. The various application domains of author profiling are forensics, security, marketing and education. In this paper the various author profiling approaches and techniques are explained and their performances are analysed.

***Keywords*—**Author Profiling, Stylometric Features, Machine Learning Algorithms, Features, Accuracy

I. INTRODUCTION

Social media websites became an inevitable part of our lives nowadays. The text which we have seen in social media are unstructured and anonymous. This anonymous text leads to cybercrimes like cyberterrorism, cyberbullying etc. So, the need for analyzing such kind of text to find out the profile characteristics of the author increases. The researchers developed a tool to find out the profile characteristics of authors by exploiting their writing styles.

Author profiling is a text classification technique which is used to predict the demographic characteristics of the author like age, gender, country, educational background, native language, and personality type by analyzing their text [12]. Researchers used different approaches like style-based approach, content-based approach, topic-based approaches and hybrid approaches for distinguishing the writing styles of authors.

Author profiling is an important area which has applications in different domains like forensics, security, marketing, education etc. In forensics, the findings about the author or the profile characteristics of author is considered as valuable evidence in forensic investigation. In security, author profiling is used to find the details of the source of threatening mails and messages. In marketing, the company used this as an aid to build business strategies by analyzing the reviews of customers. In education, this is used to evaluate the knowledge level and unique talent of each student in an educational forum.

Section II of this paper includes the related works in author profiling. Section III briefly explains the methodology of author profiling. In Section IV the different approaches of author profiling are explained. Section V explains machine learning algorithms implemented in author profiling and their performance are analyzed. Finally, section VI concludes the paper.

II. RELATED WORK

Many researchers focused on predicting the age and the gender of authors. Braja Gopal Patra, Kumar Gourav Das, and Dipankar Das [1] identified the gender of twitter authors. The machine learning algorithms used are SVM and Logistic regression. Raju Nadimpalli V. G, Gopala Krishna. P, Yelleni Mounica, V. Sahithi [2] tried to identify the gender using stylometric features. Among the different supervised algorithms, Random Forest classifier achieved good performance. Satya Sri Yatam, T. Raghunatha Reddy [5] determined the age, gender and mother language by using both content-based features and style-based features. Roy Bayot, Teresa Goncalves [4] have done experiment in twitter tweets in English and Spanish to find age and gender using Support Vector Machine algorithm.

III. AUTHOR PROFILING METHODOLOGY

The general work flow of author profiling consists of six phases. Figure. 1, shows the different phases involved in author profiling methodology.

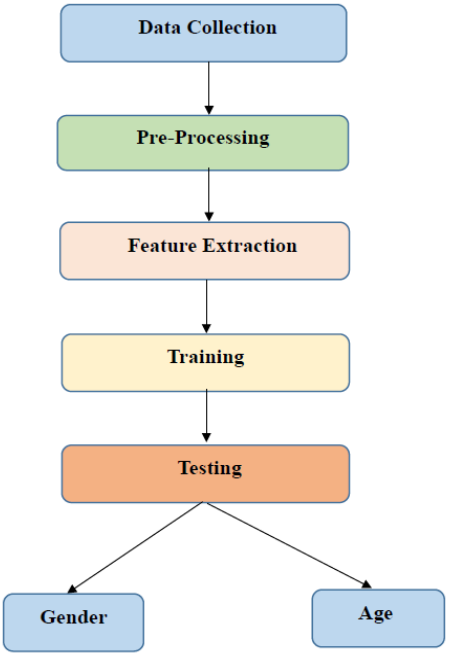

Figure.1 Block Diagram of Author Profiling Methodology.

A. Data Collection

Data collection is the first phase in every author profiling task. In data collection the data is collected from different sources like Facebook, reviews, Twitter, Blogs, social media etc. Researchers built the corpus with the text data from these sources. Most of the researchers used PAN dataset for their research purpose in author profiling. PAN dataset is the labelled dataset which is provided by the competition organizers.

B. Pre-Processing

The data which is present in the corpus has to be pre-processed since it contains noisy data, missing data, incomplete data etc. The motive of pre-processing is to clean data or to reduce the volume of data without losing the required information for effective feature extraction. It includes several techniques like,

- *Data cleaning*: This involves removing unnecessary data like images, stop words, URL etc.
- *Tokenizing*: This involves splitting the whole text into tokens (small meaningful elements). It can be words, phrases, symbols etc.
- *Normalization*: In this the text will be transformed into single canonical form.

C. Feature Extraction and Selection

The processed data is used for feature extraction This is the actual dimensionality reduction process. In this phase the data which is processed is reduced into small groups called features. In author profiling task stylometric features, content-based features and topic-based features are extracted and used for building the feature vector. Bag-Of Words (BOW) model and TF/IDF are commonly used for feature extraction in author profiling.

- *BOW model*: In this model, the vocabulary of each unique word is created and then these words will be tagged with the frequency of occurrence of that word in the document.
- *TF/IDF*: In this technique, TF represents the term frequency and IDF represents the inverse document frequency. Each word will have TF score and IDF score and the product of these scores will be the TF*IDF weight of that particular word. Let ‘t’ be a term in a document ‘d’, then the weight of that term, ie, Wt,’d’ is given by,

$$Wt,d = TFt,d \log (N/DFt)$$

Where, TFt,d - the frequency of occurrence of ‘t’ in ‘d’
DFt - the number of documents containing ‘t’
N - the number of documents in the corpus

D. Training

In training process, an algorithm is constructed which has the capability to learn and make predictions on the data given Such kind of data is called training data. One mathematical model will be generated and the test data will be feed to this model for prediction. For evaluating the performance of this model many techniques are used. K-Fold-Cross Validation is one technique used. In this technique data is divided randomly into ‘k’ equal parts. Among that one part will be taken as validation data and is used for testing the model. The remaining k-1 part is used as training data. Then the model performance is evaluated. This will be repeated for ‘k’ times.

E. Testing

The testing phase is the actual prediction part. The test data is given to the model and model will predict the output. The accuracy of the prediction is evaluated. The researchers carried out evaluation using accuracy measures. Accuracy is the ratio between total number of correct predictions (nc) over total number of predictions (np) [11].

IV. AUTHOR PROFILING APPROACHES

In the initial phases of author profiling task, features are extracted from the pre-processed datasets. These features possess an important part in predicting the profile characteristics of authors. In author profiling different types of features are extracted. Figure. 2 shows the different author profiling approaches.

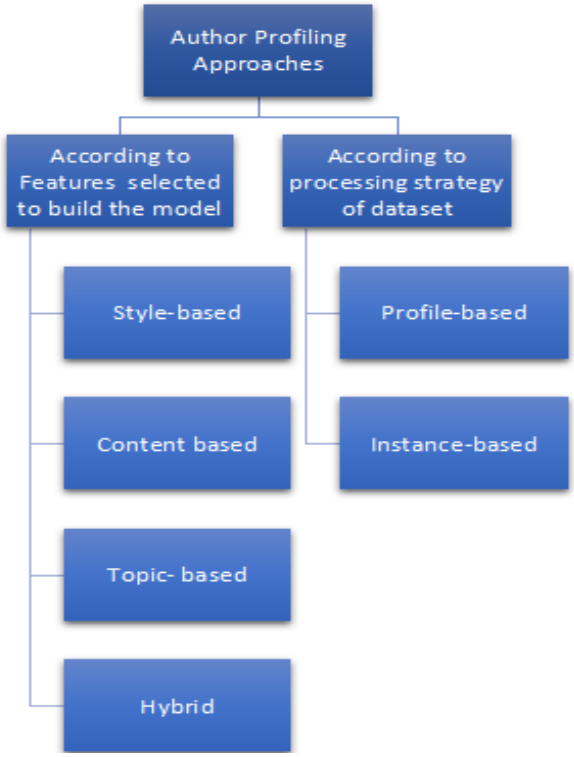

Figure.2 Classification of Author Profiling Approaches

According to the type of features extracted, the author Profiling has four different approaches as following,

A. Style-Based Approach

Every human being has unique style of writing. This varies with both gender and age. For example, males use more determiners, adjectives, 'of' modifiers than females and females use more pronouns, present tense and negations than males. In style-based approach this individual style is utilized to form the feature set. Those features are called stylometric features. The different stylometric features used are syntactic, lexical and structural. Syntactic features include punctuations, parts-of-speeches, function words etc. In Lexical features the habit of using words and characters like number of characters or words, frequency of each kind of characters or words, word length, sentence length etc are analysed in the text are measured. In structural features the author's style of organizing sentences, paragraphs etc are analysed. Style-based approach has been used by numerous researchers who have worked on the topic of author profiling in the past for identifying different demographic traits of a writer [13].

B. Content-Based Approach

In content-based approach, content-based features are utilized to find the demographic details of the author. Content based features are words which are used frequently in a domain. Researchers used these as the key element for differentiating males, females and different age groups. The choice varies with gender and different age groups. For example, males mostly used to talk about politics, sports etc whereas females used to talk about shopping, cooking, children etc. The preference of different age groups is also different. For instance, teenagers like to talk about school, video games etc, authors in 20's used to talk about college life, movies etc, individuals in 30's used to talk about marriage life, mortgage etc, and old age groups likes to talk about juvenile, pension etc.

C. Topic-Based Approach

In this approach, topic-based features are extracted. This is a statistical approach in which the frequency of occurrence of topics in document is used. A document can contain different topics but the ratio of topics varies. If a document contains 5% topics used by males and 10% topics used by females, so there would be 90% male relates words. A topic model uses the same idea for extracting hidden patterns from a collection of documents. Topic-based representations are effective in capturing the content–thematic–information of documents, and therefore that they could be appropriate for the task of author profiling in social media domains [10].

D. Hybrid Approach

In hybrid approach, the combination of style, content and topic-based features are used to increase the accuracy of prediction. Many researchers used the combination of stylometric features and content-based features to obtain maximum accuracy.

According to the processing strategy of dataset, there are two different approaches in author profiling namely profile-based approach and instance-based approach.

E. Profile Based Approach

In profile-based approach all the text which is written by a particular author is concatenated to a single file and from this file the features are extracted to find the profile characteristics of the author. In training phase, the profile of all the candidate authors will be formed. In the testing phase the text from the test data will be compared with that single file of each author and finally the author will be predicted. In Figure. 3 the generation of profiles and prediction of author is shown.

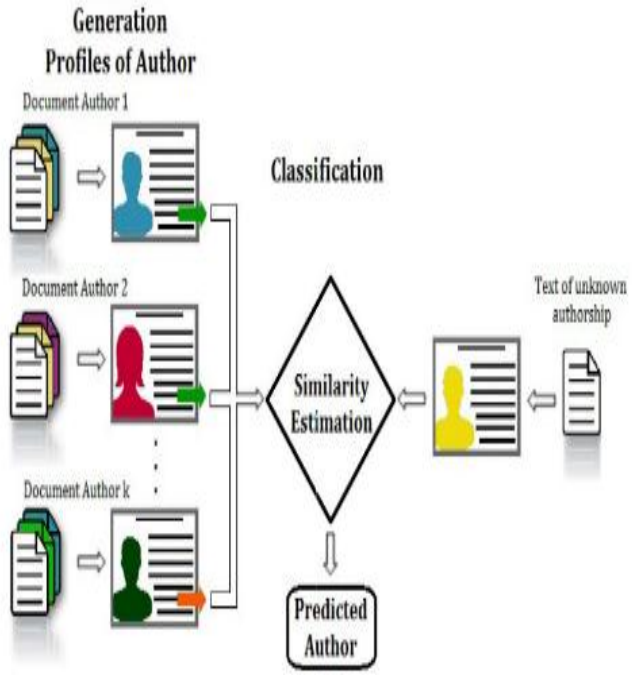

Figure.3 Generation of Profiles and Prediction of Author [8]

F. Instance Based Approach

In instance-based approach, every text sample from the training corpus is represented as a vector of attributes and a chosen classification algorithm is trained based on the instance set of known authorship (training set) in order to develop a profiling model [9]. This model is used for predicting the profile characteristics of authors from the test data. The performance of classifier increases as the number of instances in training dataset increases. In figure. 4 the classification strategy in instance-based approach is shown.

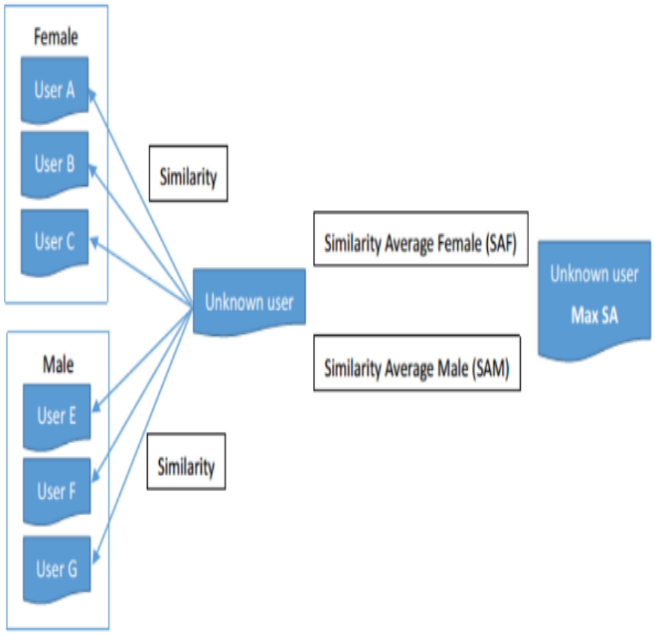

Figure. 4 Instance-Based Classification Strategy [7]

V. MACHINE LEARNING ALGORITHM

Researchers used different machine learning algorithms for classification. Some of the machine learning algorithms which is used in author profiling is briefly described below.

A. Support Vector Machine (SVM)

SVM is supervised learning algorithm which is used in author profiling to predict the different demographic features of author. This classifier is trained using the training dataset and a model is generated for prediction. For classification, SVM tries to construct a hyperplane which separates data into different categories. It selects the optimum hyperplane in such a way that the distance between the hyperplane to the support vectors should be maximum. SVM gives good accuracy in author profiling.

B. Decision Tree

Decision Tree is a machine learning algorithm which is used for both classification and regression. In author profiling researchers used this classifier for classification. It builds models in the form of tree in top-down manner. The root node and internal nodes represents the questions and the leaf node represents the prediction results. Decision Tree uses entropy and information gain for building the trees. Entropy

is used to find out the homogeneity of data. Its value will be zero if the data is homogeneous and will be one if the data is equally divided. Information gain is the difference between the entropy of sample before partition and the sum of entropy of all branches of the tree after partition. The attribute (node) with maximum information gain will be partitioned first. So, we can generalize that in a decision tree the root will have the maximum information gain and the leaf nodes will have the minimum entropy.

C. Random Forest

Random Forest is an ensemble classifier, which constructs multiple Decision Trees and uses the prediction of each trees to get the more accurate prediction. In author profiling, Random Forest got reasonable accuracy.

A comparative study of the above three machine learning algorithms in author profiling is given below.

Table 1. Performance Analysis of Machine Learning Algorithms in AP

Sr. No	Classifier	Advantage	Dis - advantage	Performa-nce in AP
1	SVM	Good for high dimensional data and overfitting is less	Takes long training time for large data	Performs well
2	Random Forest	Efficiently run on large data	Overfits to some data	Reasonable accuracy
3	Decision Tree	Implementation is easy and handles missing values	Overfitting is high	Less Accuracy

.

VI. CONCLUSION

The motive of author profiling is to find out the maximum information of the writer by analyzing his text so that it would be useful in different domains like forensic, security, marketing etc. For effective author profiling, the selection of good features and suitable classification algorithm plays an important role. It was observed that while using both style-based features and content-based features together the accuracy of prediction increases. Researches shown that in instance-based approach the classification is more stable. As the volume of training dataset increases the accuracy also increases. Researchers are continuously trying to improve the accuracy of prediction in author profiling by experimenting with different features and classification algorithms.

REFERENCES

[1] Braja Gopal Patra, Kumar Gourav Das, and Dipankar Das, "*Multimodal Author Profiling for Twitter*," Notebook for PAN at CLEF, 2018.

[2] Raju.Nadimpalli.V. G, Gopala Krishna. P, Yelleni Mounica, V Sahithi," *Authorship Profiling in Gender Identification on English editorial documents using Machine Learning Algorithms*" International Journal of Engineering Trends and Technology (IJETT), April 2017.

[3] Dang Duc Pham, Giang Binh Tran, Son Bao Pham, "*Author Profiling for Vietnamese Blogs*", International Conference on Asian Languages Processing, 2009.

[4] Roy Bayot, Teresa Goncalves, "*Multilingual Author Profiling using Word Embedding Averages and SVMs*", 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA),2016.

[5] Satya Sri Yatam, T. Raghunatha Reddy, "*Predicting Gender and Age from Blogs, Reviews & Social media*", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Vol. 3 Issue 12, December-2014.

[6] K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma, "*Author Profiling: Predicting Age and Gender from Blogs*", Notebook for PAN at CLEF, 2013.

[7] Yaritza Adame-Arcia, Daniel Castro-Castro, Reynier Ortega Bueno, Rafael Muñoz, "*Author Profiling, instance-based Similarity Classification*", Notebook for PAN at CLEF ,2017.

[8] Ma. Jos´e Garciarena Ucelay, Ma. Paula Villegas, Dario G. Funez, Leticia C. Cagnina1, Marcelo L. Errecalde, Gabriela Ram´ırez-de-la-Rosa, and Esa´u Villatoro-Tello," *Profile-based Approach for Age and Gender Identification*", Notebook for PAN at CLEF 2016.

[9] T. Raghunadha Reddy, B. Vishnu Vardhan, P. Vijayapal Reddy, "*A Survey on Authorship Profiling Techniques*", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 5 (2016) pp 3092-3102, 2016

[10] Miguel A. Álvarez-Carmona, A. Pastor López-Monroy, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, Ivan Meza. "*Chapter 13 Evaluating Topic-Based Representations for Author Profiling in Social Media*", Springer Nature, 2016.

[11] Mehwish Fatima, Komal Hasan, Saba Anwar, Rao Muhammad Adeel Nawab. "*Multilingual author profiling on Facebook*", Information Processing & Management, 2017.

[12] T Raghunatha Reddy, B. Vishnu Vardhan, Vijayapal Reddy. "*Author profile prediction using pivoted unique term normalization*", Indian Journal of Science and Technology, 2016.

[13] F. Rangel, P. Rosso, M. Koppel, and E. Stamatatos, "*Overview of the Author Profiling Task at PAN 2013*," in Notebook Papers of CLEF, 2013.

[14] Sumit Goswami, Sudeshna Sarkar, Mayur Rustagi, "*Stylometric Analysis of Blogger's Age and Gender*", Proceedings of the Third International ICWSM Conference, 2009

