Author Profiles Prediction Using Syntactic and Content-Based Features

Chapter · January 2020

DOI: 10.1007/978-981-15-1097-7_23

4 authors, including:

Tejaswi Reddy
Niloufer Hospital
130 PUBLICATIONS **1,134** CITATIONS
SEE PROFILE

Manam Srilatha
Velagapudi Ramakrishna Siddhartha Engineering College
3 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Sreenivas Mekala
Sreenidhi Institute of Science & Technology
12 PUBLICATIONS **26** CITATIONS
SEE PROFILE

Author Profiles Prediction Using Syntactic and Content-Based Features

T. Raghunadha Reddy, M. Srilatha, M. Sreenivas and N. Rajasekhar

Abstract In digital forensics, the forensic analysts raised the major questions about the details of the author of a document like identity, demographic information of authors and the documents which were related these documents. To answer these questions, the researchers proposed a new research field of stylometry which uses the set of linguistic features and machine learning algorithms. Information extraction from the textual documents has become a popular research area in the last few years to know the details of the authors. In this context, author profiling is one research area concentrated by the several researchers to know the authors' demographic profiles like age, gender, and location by examining their style of writing. Several researchers proposed various types of stylistic features to analyze the style of the authors writing. In this paper, the experiment was performed with combination of syntactic features and content-based features. Various machine learning classifiers were used to evaluate the performance of the prediction of gender of reviews dataset. The proposed method achieved best accuracy for profiles prediction in author profiling.

Keywords Gender prediction · Author profiling · PDW model · Syntactic features · Content-based features

T. Raghunadha Reddy (✉)
Department of IT, Vardhaman College of Engineering, Hyderabad, India
e-mail: raghu.sas@gmail.com

M. Srilatha
Department of CSE, VR Siddhartha Engineering College, Vijayawada, India
e-mail: srilatha.manam@gmail.com

M. Sreenivas
Department of IT, Sreenidhi Institute of Science and Technology, Hyderabad, India
e-mail: mekala.sreenivas@gmail.com

N. Rajasekhar
Department of IT, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India
e-mail: n.rajasekhar@griet.ac.in

K. S. Raju et al. (eds.), *Data Engineering and Communication Technology*,
Advances in Intelligent Systems and Computing 1079,
https://doi.org/10.1007/978-981-15-1097-7_23

1 Introduction

The Internet is increasing rapidly with the huge amount of text day by day through reviews, blogs, documents, tweets, and other social media content. The researchers need the automated tools to process the information which is dynamic in nature. In this process, sometimes it is necessary to identify the owner who has created the text or document. Authorship analysis is one such area paying attention by the many researchers to find the author details of the text [1]. Authorship analysis is a process of reaching to a conclusion by understanding the characteristics of a piece of written document and thereby analyzing the author whose roots are coming from stylometric, which is a linguistic research field. The characteristics of the texts play an important role in the procedure of the authorship analysis. In general, the authorship analysis is categorized into three categories namely authorship attribution, authorship verification, and authorship profiling [2].

Every author has their own writing style but by looking at the writing style one can predict certain profiling characteristics of the text. The following are the profiling characteristics to analyze writing styles of the authors and their age, gender, location, personality traits, native language, occupation, educational background, etc. [3]. The writing style of a human being will never change throughout his lifetime. That is the reason the style of writing of author is same either he tweets, writes in a blog or writes in a document.

There is an important phenomenon observed while understanding the writing styles of male and female. It has been witnessed that the female writing style is different from male writing style, and female are more expressive and involve emotionally in their writings. The female writing expresses both positive and negative comments on an object or a person. Researchers have given another inference on male writing wherein, the male tends to narrate new stories and focus on what had happened but whereas the female expresses how they felt [1].

Koppel et al. assumed [4] that men used more determiners and quantifiers in their writings and women writings contain more pronouns by analyzing different types of corpuses. The female authors are interested in the topics like shopping, beauty, jewelry, and kitty party. In contrast to that male authors' interested in topics related to technology, sports, women, and politics. In the past, some of the researchers found that more number of prepositions [1, 4] was found in female author's writings than the male author's. In general, the writing style is defined as a set of grammar rules, words of choice, and clubbed with the selection of topics. In another observation [5], females use more adjectives and adverbs and talk on shopping and wedding styles and male author's writings consisting of technology and politics.

In this paper, we addressed predicting the profiling characteristic of gender of the authors from reviews dataset. The existing researchers extracted different set of features to predict the author profiles. In this work, two types of features were used with PDW (Profile specific Document Weighted) model for gender prediction of the authors using different classifiers.

This work is planned in six sections. The existing works in author profiling are analyzed in Sect. 2. The characteristics of dataset of the reviews were represented in Sect. 3. The PDW model was discussed in Sect. 4. Section 5 explains the experimental results PDW model with combination of content-based features and syntactic features. Section 6 concludes this work with future scope.

2 Related Works

To allocate predefined set of classes to text documents, classification process uses a set of features and variety of machine learning procedures. Since there is no direct mechanism to process the raw text, each document is specified in the form of vector. To identify the importance of each feature in a given document, traditional feature frequency measure is used. As numerous researchers addressed, frequency of a feature is not sufficient to discover the significance of different features. To overcome this difficulty, an information retrieval-based measure TF-IDF is used to find the feature weight based on its frequency and the number of text documents that contains these feature in the given corpus [6]. A variety of weight measures are proposed by many researchers to determine the weight of these features. The count of these features and its associated weight measures along with different machine learning algorithms involved in the prediction accuracies of the author profiles in author profiling.

In the experiments of Dang Duc [7], 298 features such as character-based and word-based features were extracted from approximately 3500 web pages of 70 Vietnamese blogs and concluded that word-based features are more intended to predict the gender than the features based on characters. Stylometric features are extracted from 1000 blog posts of 20 bloggers by many Greek researchers which also included most frequent n-grams such as word n-grams and character n-grams. They observed that character n-grams and word n-grams with more length are producing good accuracy in gender prediction when experimented with SVM.

The experiment of Soler and Wanner [8] on opinion blogs of New York Times corpus, features with different combinations were tried. These combinations include sentence-based, word-based, character-based, syntactic features, and dictionary-based features for prediction of gender, and the combination of all these features achieved good accuracy. A considerable amount of drop in the accuracy is noticed with the application of bag-of-words approach on 3000 words having most TF-IDF values.

Koppel et al. [9] experimented with British National Corpus (BNC), from which 566 text documents were collected, 1081 features were extracted and good accuracy is achieved when predicting the gender. Argamon et al. [10] composed a corpus from various blogs which includes different authors (approximately 19,320), by using both stylistic features and content-based features the accuracy is calculated. Argamon, S. et al. also concluded that stylistic features are contributing more in predicting and discriminating the gender.

Palomino-Garibay Alonso et al. stated that the character level n-grams with the value of n ranging from 2 to 6 were producing better accuracies over the languages like Spanish and Dutch with the combination of abovementioned features. In order to classify the styles of writing of the different authors, some researchers suggested different composition of features like readability, semantic, syntactic, structural, character-based, and lexical features [11]. Estival et al. [12] experimented with approximately 9800 emails and collected 689 features and different types of classification algorithms, and they observed that SMO algorithm produced good accuracy in predicting the gender.

3 Dataset Characteristics

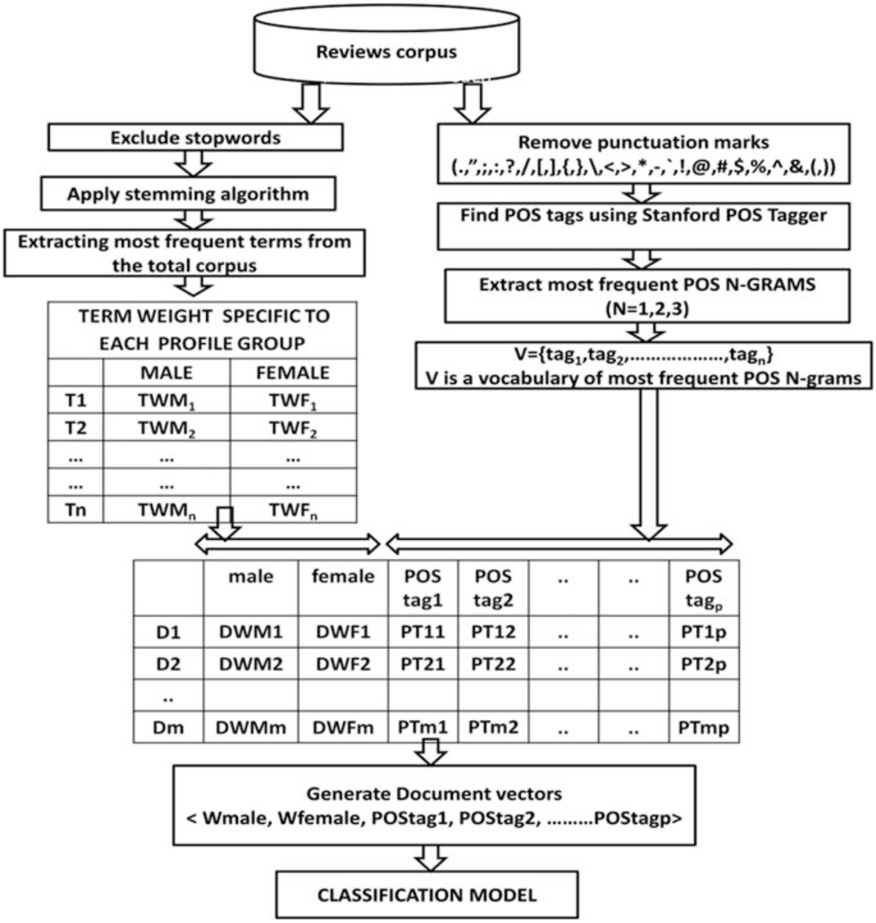
In this work, the dataset contains 4000 English reviews of different hotels and it was gathered from www.TripAdvisor.com. The dataset for gender profile was balanced, i.e., both female and male profile groups include 2000 documents each.

The researchers used different types of performance measures such as recall, precision, accuracy, and F1-measure were used to determine the efficiency of their proposed system in the approaches of author profiling problem. In this paper, accuracy measure is used to compute the performance of the system. In this context, the accuracy is defined as the ratio among the number of test documents was predicted their author gender correctly and the total number of test documents considered for testing [13].

4 PDW Model

In our view of knowledge, few researchers concentrated on different representations of text in author profiling. At present, bag-of-words (BOW) model is used by most of the researchers to represent the document vectors in author profiling. In BOW model, the document vector values are computed by the frequency of content/stylistic features. The existing document vector representation techniques face many problems. To overcome those difficulties and improving the prediction accuracy of author profiles, Raghunadha Reddy et al. [14] proposed a new approach named PDW which uses the relationship among the documents to author profile group and the terms to documents. In this work, we used the combination of content-based features such as most frequent terms and syntactic features like most frequent POS n-grams to represent the document.

Figure 1 shows the model of PDW approach. In this model, the training dataset consists of 4000 reviews of male and female, and the dataset is balanced that is each gender group contains 2000 documents. The dataset is used for extracting content-based and syntactic features. In content-based features extraction, two preprocessing methods like stop word removal and stemming were used to prepare


Reviews corpus
Exclude stopwords
Apply stemming algorithm
Extracting most frequent terms from the total corpus
Remove punctuation marks (.,”,;,:,?,/,[,],{,},\,<,>,*,-,`,!,@,#,$,%,^,&,(,))
Find POS tags using Stanford POS Tagger
Extract most frequent POS N-GRAMS (N=1,2,3)
V={tag_1,tag_2,………………,tag_n}
V is a vocabulary of most frequent POS N-grams
Generate Document vectors
< Wmale, Wfemale, POStag1, POStag2, ………POStagp>
CLASSIFICATION MODEL

TERM WEIGHT SPECIFIC TO EACH PROFILE GROUP		
	MALE	FEMALE
T1	TWM_1	TWF_1
T2	TWM_2	TWF_2
…	…	…
…	…	…
Tn	TWM_n	TWF_n

	male	female	POS tag1	POS tag2	..	..	POS tag_p
D1	DWM1	DWF1	PT11	PT12	..	..	PT1p
D2	DWM2	DWF2	PT21	PT22	..	..	PT2p
..							
Dm	DWMm	DWFm	PTm1	PTm2	..	..	PTmp

Fig. 1 PDW model for prediction of gender using content-based features and syntactic features

the data for efficient features extraction. Once the data is cleaned then extract the most frequent terms from the entire training dataset. After extraction of most frequent terms, each document is represented with these terms as document vector. In this model, term weight measure was used to compute the terms weight. Each term's importance is measured by the weight of the term against each gender group. These term weights of individual gender group were used to compute the weight of the documents by using document weight measure. The document weight is also computed against to each gender group.

In syntactic features extraction, first remove the punctuation marks to clean the text for efficient features extraction. We used Stanford POS tagger to identify the part of speech of words. The most frequent POS n-grams were extracted from the dataset as syntactic features. In this paper, the most frequent POS n-grams where n range from 1 to 3 (POS unigrams, bigrams, trigrams) was considered as syntactic features. {tag1, tag2, …, tag$n$} is a set of POS tags, {male, female} is a set of sub-profile groups in a gender profile. In this model, DWM_m, DWF_m are the document D_m weights in male and female documents, respectively. $\mathrm{PT}_{\mathrm{mp}}$ is the weight of POS tag 'p' in document 'm.'

Finally, the vectors of documents were represented with the weights of the document computed in the previous step and the frequencies of the most frequent POS n-grams identified in the later step. The machine learning algorithms used these document vectors to generate the classification model. The gender of a new document is predicted by using this classification model. The term weight measures and the document weight measures influence the prediction accuracy of the gender. In this paper, we used a supervised term weight measures proposed in [15].

4.1 Supervised Term Weight Measure (STWM)

The general idea is the term which appears in most of the documents has least power to distinguish from different classes to that of the terms which appears in a single document or few documents. Also the computed weights play a crucial role in assigning the class labels. The terms which are appearing only in single document or in few documents carry most representative information called unique terms. The terms with more concentrated inter-class distribution have strong distinguishing power. Therefore, this weight measure allocates weight to the terms according to their inner document, inter-class and intra-class distributions [15]. Equation (1) shows the STWM.

$$\begin{aligned} W\left(t_i, d_k \in \mathrm{PG}_p\right) = {} & \frac{\mathrm{tf}\left(t_i, d_k\right)}{\mathrm{DF}_k} \times \frac{\sum_{x=1, d_x \in \mathrm{PG}_p}^{m} \mathrm{tf}\left(t_i, d_x\right)}{1 + \left(\sum_{y=1, d_y \notin \mathrm{PG}_p}^{n} \mathrm{tf}\left(t_i, d_y\right)\right)} \\ & \times \frac{\sum_{x=1, d_x \in \mathrm{PG}_p}^{m} \mathrm{DC}\left(t_i, d_x\right)}{1 + \left(\sum_{y=1, d_y \notin \mathrm{PG}_p}^{n} \mathrm{DC}\left(t_i, d_y\right)\right)} \end{aligned} \quad (1)$$

In this measure, tf (t_i, d_k) represents the no. of times the term t_i appeared in document d_k, DF_k represents total number of terms in a document d_k.

4.2 Document Weight Measure (DWM)

In this thesis, the document weight against to author groups is calculated by using a DWM proposed by Raghunadha Reddy et al. [16]. The DWM is represented in Eq. (2).

$$W\left(d_k, \mathrm{PG}_p\right) = \sum_{t_i \in d_k, d_k \in \mathrm{PG}_p} \mathrm{TFIDF}(t_i, d_k) * W\left(t_i, \mathrm{PG}_p\right) \tag{2}$$

In this measure, $W(d_k, \mathrm{PG}_p)$ is the document weight of d_k in PG_p profile group. This DWM is a product of TF-IDF weight of a term and the weight of a term against to author group computed in previous step of the model.

5 Experimental Results

5.1 PDW Model with Content-Based Features and Syntactic Features

The major steps in PDW model are finding suitable features, identifying suitable sub-profiles, computing feature weights and document weights, and document representation. Table 1 presents the gender prediction accuracies of PDW model, when content-based and syntactic features were used to represent the vector of document and different types of machine learning algorithms such as Naïve Bayes Multinomial (NBM), Simple Logistic (SL), Logistic (LOG), IBK, Bagging (BAG), and Random Forest (RF) were used. In this work, the experiment performed with combination of 8000 most frequent terms and the number of POS n-grams varies with an interval of 500 from 500 to 3000 for computing the document weight.

The PDW model produced highest accuracy of 93.25% for prediction of gender when Random Forest classifier used 8000 most frequent terms and 3000 most frequent POS n-grams. The Random Forest classifier achieved best accuracy when compared with other classifiers. It was observed that the accuracies were increased when the number of features was increased.

Table 1 The accuracies of PDW model when content-based features and syntactic features were used as features

Classifier/number of features	NBM	SL	LOG	IBK	BAG	RF
8000 terms + 500 POS n-grams	84.51	78.63	80.54	70.65	64.56	87.65
8000 terms + 1000 POS n-grams	86.74	79.74	81.71	71.78	65.74	88.89
8000 terms + 1500 POS n-grams	87.20	81.32	83.32	73.94	67.91	90.12
8000 terms + 2000 POS n-grams	89.76	82.87	84.89	74.09	68.87	91.56
8000 terms + 2500 POS n-grams	90.94	84.21	86.54	75.74	69.41	92.71
8000 terms + 3000 POS n-grams	91.36	84.89	87.08	76.36	71.34	93.25

6 Conclusions

In this work, a PDW model was experimented with combination of syntactic and content-based features. The PDW model with the combination of both content-based features and syntactic features obtained 93.25% for gender prediction. It was observed that with the constant number of content-based features, the accuracy of gender prediction is increased when the number of POS n-grams was increased. It was also observed that the syntactic features influence is more when compared with content-based features to improve the accuracy of gender.

References

1. Koppel, M., Argamon, S., Shimoni, A.: Automatically categorizing written texts by author gender. Literary Linguist. Comput. 401–412 (2003)
2. Raghunadha Reddy, T., Vishnu Vardhan, B., Vijayapal Reddy, P.: A survey on author profiling techniques. Int. J. Appl. Eng. Res. **11**(5), 3092–3102 (2016)
3. Raghunadha Reddy, T., Vishnu Vardhan, B., Vijayapal Reddy, P.: N-gram approach for gender prediction. In: 7th IEEE International Advanced Computing Conference, Hyderabad, Telangana, pp. 860–865, 5–7 Jan 2017
4. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: unmasking pseudonymous authors. J. Mach. Learn. Res. **8**, 1261–1276 (2007)
5. Newman, M.L., Groom, C.J., Handelman, L.D., Pennebaker, J.W.: Gender differences in language use: an analysis of 14,000 text samples. Discourse Process. **45**(3), 211–236 (2008)
6. Pradeep Reddy, K., Raghunadha Reddy, T., Apparao Naidu, G., Vishnu Vardhan, B.: Term weight measures influence in information retrieval. Int. J. Eng. Technol. **7**(2), 832–836 (2018)
7. Dang Duc, P., Giang Binh, T., Son Bao, P.: Author profiling for vietnamese blogs. Asian Language Processing, 2009 (IALP '09), pp. 190–194 (2009)
8. Company, J.S., Wanner, L.: How to use less features and reach better performance in author gender identification. The 9th edition of the Language Resources and Evaluation Conference (LREC), 26–31 May 2007
9. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging. In: Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, vol. 6, pp. 199–205, Mar 2006
10. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM **52**(2), 119 (2009)
11. Palomino-Garibay, A., Camacho-Gonzalez, A.T., Fierro-Villaneda, R.A., Hernandez-Farias, I, Buscaldi, D., Meza-Ruiz, I.V.: A random forest approach for authorship profiling. In: Proceedings of CLEF 2015 Evaluation Labs (2015)
12. Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B.: Author profiling for English emails. In: 10th Conference of the Pacific Association for Computational Linguistics (PACLING, 2007) (2007)
13. Raghunadha Reddy, T., Vishnu Vardhan, B., Vijayapal Reddy, P.: Author profile prediction using pivoted unique term normalization. Indian J. Sci. Technol. **9**(46) (2016)
14. Raghunadha Reddy, T., Vishnu Vardhan, B., Vijayapal Reddy, P.: Profile specific document weighted approach using a new term weighting measure for author profiling. Int. J. Intell. Eng. Syst. **9**(4), 136–146 (2016)

15. Sreenivas, M., Raghunadha Reddy, T., Vishnu Vardhan, B.: A novel document representation approach for authorship attribution. Int. J. Intell. Eng. Syst. **11**(3), 261–270 (2018)
16. Raghunadha Reddy, T., Vishnu Vardhan, B., Vijayapal Reddy, P.: A document weighted approach for gender and age prediction. Int. J. Eng. –Trans. B: Appl. **30**(5), 647–653 (2017)