A Feature Selection Technique–Based Approach for Author Profiling Using Word Embedding Techniques

Karunakar Kavuri and M. Kavitha

1 Introduction

On the internet, textual data is exponentially increasing through various platforms, such as reviews, emails, blogs, Twitter tweets, Facebook posts, etc. This textual content provides a noteworthy and considerable source of data. Researchers have concentrated on this data to extrac hidden information. Author profiling (AP) is one technique developed by the research community to extract demographic information of authors by examining their written texts. In AP, there is a possibility to categorize authors into different groups by considering similar features or patterns collected from their written text and learned through machine learning (ML) algorithms. This is in high demand because the count of users is high on online platforms [1]. Studies suggested that the gender, age, and other profiles of every author are indirectly or directly associated with their style of writing and it is very crucial to extract these profiles from a given document or post [2]. As a result, researchers have widely used this author profiling technique to address different varieties of problems across different domains [2].

In 2013, PAN competition started concentrating on different platforms of social networking sites to extract various types of datasets for predicting the demographic profiles of authors by using author profiling techniques. These techniques are mostly applied on social network platforms because most of the users are anonymously posting messages [3]. Author profiling become an active research field because of its variety of applications in extensive areas, such as forensics, marketing, education, and security [4]. In the domain of security, author profiling tasks are

Karunakar K. (✉) · M. Kavitha
CSE Department, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India
e-mail: karunakar.mtech@gmail.com; kavitha@veltech.edu.in

F. M. Lin et al. (eds.), *Accelerating Discoveries in Data Science and Artificial Intelligence I*, Springer Proceedings in Mathematics & Statistics 421,
https://doi.org/10.1007/978-3-031-51167-7_72

trying to reveal the authors' identity by classifying and predicting their profiles, which are helpful for detecting the source of terrorists [5] and protecting users from identity theft or online harm [6]. In the field of marketing, author profiling systems help to enhance strategies of marketing by permitting companies to learn about the demographic characteristics of online customers who gave reviews about their products or services, and also helpful to recognize appropriate customers for advertising their services or products online [3]. In digital forensics analysis, author profiling techniques provide benefits to detect the demographics of the perpetrator who is involved in the crime. Author profiling tasks are also more helpful in the education domain to identify the exceptional talents of students based on the writing style of students [6].

Most of the researchers paid more attention to predicting the gender and age of authors than other demographic profiles like native language and personality traits. Other demographics like mental health of the author, region of origin, and level of education are not extensively detected in the approaches of author profiling [7]. Most of the researchers used stylistic features to differentiate the author's writing style. In this Article, we developed a feature selection technique–based approach for author profiling by using word embedding techniques. In the proposed approach, we experimented with four embedding techniques Word2Vec, Glove, fastText, and BERT, to generate word embeddings. Group similar words as clusters by using the similarity score among the word embeddings. Select important words from clusters and represent documents as vectors by using these selected words. These vectors are trained with two ML algorithms, SVM (Support Vector Machine) and RF (Random Forest). The experiment was carried out with two PAN competition datasets for predicting age and gender.

This chapter is divided into six sections. Section 2 discusses the different approaches proposed for the task of author profiling. The information about the dataset is described in Sect. 3. The description of the proposed method, word embedding techniques, similarity measure, TWM (Term Weight Measure) and ML algorithms are given in Sect. 4. The empirical evaluations of the proposed method are displayed in Sect. 5. Section 6 concludes this chapter with future plans to enhance the prediction accuracy of age and gender.

2 Literature Survey on Existing Author Profiling Approaches

In this twenty-first century, artificial intelligence and machine learning techniques have becomewidely developed and provide high-quality services to humans in several research domains. As a result of these improvements, several companies utilized predictive models for assessing the behavior of customers. Also, with the extensive utilization of social media, most of the companies deliver their services and products to customers through these accounts of social media. But most of the customers are not showing interest in all services and products. Every customer has their own interests. The gender of the customer plays a crucial role when it comes to selecting a service or product. If the gender of a user in social media is known, then the companies are able to offer most suitable services or products. Ozer ÇELİK et al. proposed [8] a method for predicting the gender of the commenter. The main aim of this study was the estimation of gender by using machine learning algorithms. The method applied was the comments that are posted on Facebook by the companies. In the dataset, the gender of the commenter was labeled by using the name of the commenter. Among different machine learning algorithms, the logistic regression classifier attained the highest accuracy of 74.13% for gender prediction.

Jesus Silvaa et al. presented [9] a method for predicting the age and gender profiles of authors. This method was developed by utilizing a set of dialogues that are written in two different languages, such as Spanish and English, which are provided in a PAN "Uncovering Plagiarism, Authorship, and Social Software Misuse" evaluation forum 2018 task of author profiling. They developed a two-stage classification system by using syntactic, semantic, and lexical characteristics of text. This system first predicts the gender profile of an author and then predicts the age of the author. The experimentation results proved that there is a possibility to obtain an accuracy greater than 50% for both age and gender prediction with the available data in the dataset. Further, these accuracy values could be enhanced by using more information on training data.

AP is the process of information extraction about the written texts of the author via linguistic analysis. Scholars such as psychologists, linguists, and data analytics specialists utilize this technique to investigate about the prediction of personality traits, demographics, native language, and education of authors. In this context, gender is a very important characteristic of the author. A few research works reported that the accuracy of gender prediction was as high as 80%, and even higher. However, there are still several issues that need to be addressed. The first issue is many existing research works concentrated on English texts only. The second issue is many research articles focused on the content-based features that were imitated easily. The third issue is several latest papers used the concepts of machine learning algorithms to highlighting the accuracy of profile prediction but did not emphasize the differences between the writing styles of females and males. Tatiana Litvinova et al. developed [10] a mathematical model to detect the gender of the author by using only topic-independent high-frequency text parameters. The main purpose of their work is to disclose the writing style differences between female and male texts that are written in Russian. They paid special attention to the comparison of writing style differences in written texts of males and females with those differences obtained earlier for Russian and other languages.

Miguel A et al. developed [11] a multimodal method for extracting information from images and written messages that are shared by users. The same approach was used for both text and images by converting images to text. The extracted information from images and text was explored by using various distributional term representations in order to detect the topics discussed by the user. The experimental results show that the images' textual descriptions consist of more useful information for the task of AP, and the fusion of information that was extracted from images with the textual information enhances the accuracy of the author profiling task.

Table 1 PAN 2014 Reviews dataset details of gender and age profiles

	Gender		Age group				
	Female	Male	18–24	25–34	35–49	50–64	65–xx
Number of documents	2080	2080	360	1000	1000	1000	800
Total number of documents	4160		4160				

Table 2 The characteristics of PAN 2016 English training dataset

	Gender		Age group				
	Female	Male	18–24	25–34	35–49	50–64	65–xx
Number of documents	218	218	28	140	182	80	6
Total number of tweets	113,972	149,059	363,031				

3 Dataset Characteristics

In this chapter, experimentation was performed on two PAN competition datasets, which are PAN 2014 Reviews dataset [12] and PAN 2016 Tweets datasets [13], for age and gender prediction. The descriptions of these datasets are displayed in Tables 1 and 2.

4 Proposed Feature Selection Technique–Based Approach for Author Profiling

In this chapter, a new feature selection technique–based approach is proposed for age and gender prediction. Figure 1 shows the proposed approach. In this approach, apply different preprocessing techniques like tokenization, lowercase conversion, punctuation marks removal, stop-words removal, and lemmatization to prepare the dataset for extracting features. Preprocessing standardizes the text, allows common patterns to be detected across similar items in the text, and removes inconsistent or unknown tokens and formatting from the full dataset. Lowercasing converts all the letters in the full corpus to lowercase letters. Tokenization splits the full corpus into standard tokens. A stop-word is a predefined common word like determiners, prepositions, articles, etc., which adds little value to the meaning of a corpus. Lemmatization is the practice of replacing each word with its root. For example, “computed” and “computing” would both be replaced with “compute.”

After cleaning the dataset, then extract all words from the dataset. Forward all words to word embedding techniques to generate word vectors. Find the similarity among the word vectors by using the similarity measure. Grouping similar words into clusters based on the similarity score. Identify important words in each cluster based on the TFIDF (“Term Frequency and Inverse Document Frequency”) weight of words. Utilize these identified words for document vector representation. Use a term weight measure to represent the word value in the

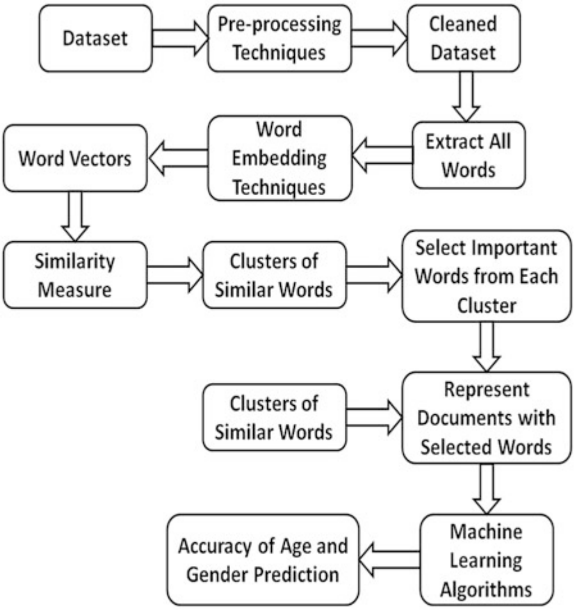
Dataset
Pre-processing Techniques
Cleaned Dataset
Extract All Words
Word Embedding Techniques
Word Vectors
Similarity Measure
Clusters of Similar Words
Select Important Words from Each Cluster
Clusters of Similar Words
Represent Documents with Selected Words
Machine Learning Algorithms
Accuracy of Age and Gender Prediction

Fig. 1 Proposed feature selection technique–based approach for author profiling

document vector representation. These vectors are trained with ML algorithm to yield the classification model. This model detects the prediction accuracy of age and gender. In the proposed approach, different word embedding techniques, similarity measures, term weight measures, and ML algorithms are used for predicting the age and gender of the author.

4.1 Word Embedding Techniques

A language model allows us to represent and work with language. There are many ways to model language, varying in detail and scope. Many models have a notion of a vocabulary that is represented by the model. The vocabulary is a finite list of words. Each vocabulary word is represented in the language model. Existing models represent each word in the vocabulary and guarantee a unique identifier for each word. However, this model does not capture any information about the relationships between words. It is possible to capture more information with the way we represent the vocabulary words. One such possibility is word embeddings. A word embedding is a vector that represents a word. Each word in the vocabulary is denoted by a unique vector [14]. All the vectors are the same length. The embeddings are learned from input text. As such, words that are more similar will have more similar word embeddings. Words that are less similar will have less similar word embeddings. In the proposed approach, we experimented with four word embedding techniques, which are Word2Vec [15], Glove [16], FastText [17], and BERT [18].

4.2 Similarity Measure

Similarity measures determine the similarity among pairs of vectors. In this chapter, we used the cosine similarity measure (CSM) for finding the similarity among pairs of word embeddings. If all vectors have the same length, then the similarity between two vectors is determined by how similarly they are positioned. In the context of word embeddings, it is common to use the cosine similarity measure [19]. The cosine similarity measure is calculated by using Eq. (1).

$$\mathrm{CSM}\,(D_a, D_b) = \frac{\sum_{k=1}^{n} W\,(T_k, D_a) \times W\,(T_k, D_b)}{\sqrt{\sum_{k=1}^{n} (W\,(T_k, D_a))^2} \times \sqrt{\sum_{k=1}^{n} (W\,(T_k, D_b))^2}} \tag{1}$$

where $\{T_1, T_2, \ldots T_n\}$ is a set of terms, $\{D_1, D_2, \ldots, D_m\}$ is a set of documents in the training dataset. CSM(D_a, D_b) is the cosine similarity among the document D_a and D_b vectors. $W(T_k, D_a)$ is the term T_k weight in document D_p and $W(T_k, D_b)$ is the weight of term T_k in D_b document.

The cosine similarity measure computes the cosine of the angle among two vectors that are normalized. The cosine angle varies from 0 to 180° for the values ranging from -1 to $+1$, respectively. Using this metric, the closer the cosine similarity of two word embeddings is to 1, the more similar the word embeddings are. The closer the cosine similarity is to 0 for two word embeddings, the less similar the embeddings are. Hence, higher cosine similarity is better when looking for similar words.

4.3 Term Weight Measure (TWM)

The TWM determines the term importance in a document [20]. In this chapter, we used a TWM proposed in our previous work [20]. Eq. (2) is used to compute the weight of terms by using Proposed Term Weight Measure (PTWM).

$$\begin{aligned} \mathrm{PTWM}\left(T_i, D_k \in C_j\right) = \frac{\mathrm{TF}\,(T_i, D_k)}{\mathrm{TNTD}_k} * \frac{\mathrm{TF}\left(T_i, D_k \in C_j\right)}{\mathrm{TF}\left(T_i, D_k \notin C_j\right)} * \frac{A+D}{B+C} \\ * \left(\frac{A}{A+B} - \frac{C}{C+D}\right) \end{aligned} \tag{2}$$

where A and C are counts of documents that contain the T_i term in C_j class of documents and other than C_j class of documents, respectively. B and D are counts of documents that don't contain the T_i term in C_j class of documents and other than C_j class of documents, respectively. TF(T_i, D_k) is the term T_i frequency in D_k

Table 3 The prediction accuracies of age and gender on the dataset of PAN 2014

	SVM		RF	
Word embedding technique/ML techniques – profiles	Gender	Age	Gender	Age
Word2Vec	0.8512	0.8164	0.8623	0.8327
Glove	0.8489	0.8232	0.8759	0.8416
FastText	0.8551	0.8386	0.8914	0.8474
BERT	0.8738	0.8445	0.9082	0.8638

document, TNTDk is the total count of terms in D_k document. TF(T_i, $D_k \in C_j$) is count of T_i term in a C_j class of documents. TF(T_i, $D_k \notin C_j$) is count of T_i term in other than C_j class of documents.

4.4 Machine Learning (ML) Algorithms

The method of features extraction from input data is the main difference between machine learning and deep learning techniques. In traditional ML techniques, features are manually extracted by an expert of dataset which is called as feature engineering, whereas in deep learning models, the network automatically extracts a set of features. In this research article, we conducted an experiment with two ML techniques such as RF [21] and SVM [22].

5 Empirical Evaluations

In this chapter, we proposed a novel feature selection technique–based approach for author profiling by using word embedding techniques. The performance of the proposed approach is represented by using an evaluation measure of accuracy. In the context of AP, accuracy is the number of test samples of the dataset that correctly predicted their age and gender from the total test samples. Two classification algorithms, which are RF and SVM, are used for determining the accuracy of the proposed approach. Table 3 presents the accuracies of age and gender prediction on the PAN 2014 dataset when the experiment was conducted with different word embedding techniques and different machine learning techniques.

The BERT model accomplished good accuracies of 0.8738 and 0.8445 for gender and age prediction, respectively, when the experiment was performed with the SVM classifier. The RF classifier with the BERT model attained the highest accuracies of 0.9082 and 0.8638 for gender and age prediction, respectively. The BERT model shows the best performance when compared with the efficiency of other word embedding techniques. The RF classifier performance is better than the performance of SVM in all cases. Table 4 presents the accuracies of age and gender prediction

Table 4 The prediction accuracies of age and gender on the dataset of PAN 2016

	SVM		RF	
Word embedding technique/ML techniques – profiles	Gender	Age	Gender	Age
Word2Vec	0.8319	0.7328	0.8478	0.7562
Glove	0.8347	0.7419	0.8518	0.7637
FastText	0.8253	0.7462	0.8547	0.7689
BERT	0.8476	0.7586	0.8639	0.7753

on the PAN 2016 dataset when the experiment was conducted with different word embedding techniques and different machine learning techniques.

The BERT model accomplished good accuracies of 0.8476 and 0.7586 for gender and age prediction, respectively, when the experiment was performed with the SVM classifier. The RF classifier with the BERT model attained the highest accuracies of 0.8639 and 0.7753 for gender and age prediction, respectively. The BERT model shows the best performance when compared with the efficiency of other word embedding techniques. The RF classifier performance is better than the performance of SVM in all cases.

6 Conclusions and Future Scope

The main goal of AP techniques is to detect the author’s demographic profiles by analyzing their written texts. Most of the research works succeeded in obtaining good accuracies for gender and age prediction by using content-based features. For this chapter, the experiment was carried out with content-based features in the dataset. We proposed a feature selection technique–based approach for author profiling by using word embedding techniques. In this approach, different word embedding techniques are used in the experiment, and it was observed that the BERT model accomplished the highest accuracies for age and gender prediction. The RF classifier with the BERT model attained the best accuracies of 0.8638 and 0.9082 for age and gender prediction, respectively, on PAN 2014 dataset. For the 2016 dataset, the RF classifier with the BERT model accomplished the highest accuracies of 0.7753 and 0.8639 for age and gender prediction, respectively.

In future work, we are planning to consider autoencoders for selecting relevant features to improve the prediction accuracy of gender and age. We are also planning to apply the proposed approach to predict other demographic profiles.

References

1. T. Raghunadha Reddy, B. Vishnu Vardhan, P. Vijayapal Reddy, A survey on author profiling techniques. Int. J. Appl. Eng. Res. **11**(5), 3092–3102 (2016)

2. L. Kaati, E. Lundeqvist, A. Shrestha, M. Svensson, Author profiling in the wild, in *2017 European Intelligence and Security Informatics Conference (EISIC)*, (2017), pp. 155–158
3. H. Markov, J.P. Gomez-Adorno, G. Posadas-Duran, Sidorov, A. Gelbukh, Author profiling with doc2vec neural network-based document embeddings, in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10062 LNAI*, (2017), pp. 117–131
4. T. Raghunadha Reddy, B. Vishnu Vardhan, P. Vijayapal Reddy, Profile specific document weighted approach using a new term weighting measure for author profiling. Int. J. Intell. Eng. Syst. **9**(4), 136–146 (2016)
5. G.S. Reddy, T.M. Mohan, T.R. Reddy, Author profiling approach for location prediction, in *First International Conference on Artificial Intelligence and Cognitive Computing*, (Springer, 2019), pp. 389–395
6. R. Bayot, T. Goncalves, Multilingual author profiling using word embedding averages and SVMs, in *SKIMA 2016–2016 10th International Conference on Software, Knowledge, Information Management and Applications*, pp. 382, 2017–386
7. T. Raghunadha Reddy, B. Vishnu Vardhan, P. Vijayapal Reddy, A document weighted approach for gender and age prediction. Int. J. Eng. Trans. B Appl. **30**(5), 647–653 (2017)
8. Ç.E.L.İ.K. Ozer, A.S.L.A.N. Ahmet Faruk, Gender prediction from social media comments with machine learning. Sakarya Uni. J. Sci. **23**(6), 1256–1264 (2019)
9. J. Silvaa, S. García, M.A. Binda, F.M. Gonzalez, R. Barrios, B.L. Castro, L. Castro, A method for detecting the profile of an author. Proc. Comput. Sci. **170**, 959–964 (2020)
10. T. Litvinova, P. Seredin, O. Litvinova, O. Zagorovskaya, Identification of gender of the author of a written text using topic-independent features. Pertanika J. Soc. Sci. Hum. **26**(1), 103–112 (2018)
11. A. Miguel, A. Carmona, E.V. Tello, M. Montes, L.V.P. Gomez, Author profiling in social media with multimodal information. Computación y Sistemas **24**(3), 1289–1304 (2020). https://doi.org/10.13053/CyS-24-3-3488
12. F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, W. Daelemans, Overview of the 2nd author profiling task at pan 2014, in *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers*, (Sheffield, 2014), pp. 1–30
13. F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, B. Stein, Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. CEUR Workshop Proc. **1609**, 750–784 (2016)
14. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information. arXiv:1607.04606 [cs] (2017) arXiv: 1607.04606. [Online]. Available: http://arxiv.org/abs/1607.04606, Visited on 05/26/2021
15. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
16. J. Pennington, R. Socher, C. Manning, Glove: global vectors for word representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Association for Computational Linguistics, Doha, 2014), pp. 1532–1543
17. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
18. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Association for Computational Linguistics, Minneapolis, 2019), pp. 4171–4186
19. A. Jain, A. Jain, N. Chauhan, V. Singh, N. Thakur, Information retrieval using cosine and Jaccard similarity measures in vector space model. Int. J. Comput. Appl. **164**(6), 28–30 (2017)
20. K. Kavuri, M. Kavitha, A term weight measure based approach for author profiling, in *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), Chennai, India*, (2022), pp. 275–280. https://doi.org/10.1109/ICESIC53714.2022.9783526

21. C. Swathi, K. Karunakar, G. Archana, T. Raghunadha Reddy, A new term weight measure for gender prediction in author profiling. Proc. Adv. Intell. Syst. Comput. **695**, 11–18 (2018)
22. V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, 2013)