Author Profiling with Doc2vec Neural Network-Based Document Embeddings (preprint version)

Conference Paper *in* Lecture Notes in Computer Science · August 2017

DOI: 10.1007/978-3-319-62428-0_9

Ilia Markov
University of Antwerp
41 PUBLICATIONS **236** CITATIONS

Helena Gomez Adorno
Universidad Nacional Autónoma de México
40 PUBLICATIONS **411** CITATIONS

Grigori Sidorov
Instituto Politécnico Nacional
232 PUBLICATIONS **1,526** CITATIONS

Alexander Gelbukh
Instituto Politécnico Nacional
506 PUBLICATIONS **4,703** CITATIONS

Author Profiling with doc2vec Neural Network-Based Document Embeddings Preprint version*

Ilia Markov[1], Helena Gómez-Adorno[1], Juan-Pablo Posadas-Durán[2],
Grigori Sidorov[1], and Alexander Gelbukh[1]

[1] CIC, Instituto Politécnico Nacional (IPN), Mexico City, Mexico
markovilya@yahoo.com, helena.adorno@gmail.com,
sidorov@cic.ipn.mx, www.gelbukh.com
[2] ESIME-Zacatenco, Instituto Politécnico Nacional (IPN), Mexico City, Mexico
jpposadas@gmail.com

Abstract. To determine author demographics of texts in social media such as Twitter, blogs, and reviews, we use doc2vec document embeddings to train a logistic regression classifier. We experimented with age and gender identification on the PAN author profiling 2014–2016 corpora under both single- and cross-genre conditions. We show that under certain settings the neural network-based features outperform the traditional features when using the same classifier. Our method outperforms existing state of the art under some settings, though the current state-of-the-art results on those tasks have been quite weak.

Keywords: document embeddings, doc2vec, neural networks, machine learning, author profiling

1 Introduction

The author profiling (AP) task aims at identifying author demographics, such as age, gender, personality traits, or native language, basing on the analysis of text samples. This research area has experienced an explosive increase in interest in recent years. It contributes to marketing, security, terrorism prevention, and forensic applications, among other.

The approaches that tackle the task of AP from the machine-learning perspective view the task as a multi-class, single-label classification problem, when the set of class labels is known *a priori*. Thus, AP is modeled as a classification task, in which automatic methods have to assign class labels (e.g., male, female) to objects (texts).

Machine-learning algorithms require input data to be represented in the form of a fixed-length feature vector. Approaches that have been used to obtain such

* The final publication is available at Springer: http://dx.doi.org/10.1007/978-3-319-65813-1_15

vector include bag-of-words, bag-of-n-grams, etc., models. In this work, we apply the doc2vec algorithm [1] to obtain the fixed-length feature vector, that is, we learn neural network-based document embeddings (also known as document distributed representations or paragraph vectors) in an unsupervised manner from texts. This type of document embeddings allows representing texts as dense vectors, taking into account their semantic and syntactic structure. Furthermore, this representation has been shown to be efficient when dealing with high-dimensional and sparse data [1, 2].

Our motivation was two-fold: first, to suggest an author profiling method better than existing ones; second, to compare the doc2vec features with traditional features when used with the same classifier. We show that using neural network-based document embeddings for the AP task improves the classifier performance in some settings; we conducted experiments not only under single-genre conditions but also under cross-genre AP conditions, when the training and test datasets are from significantly different sources, such as Twitter vs. reviews. Namely, the neural network-based features outperform the baseline features in many cases when used with the same classifier (logistic regression in our case), as well as outperform the state-of-the-art approaches under some AP conditions.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 describes the proposed methodology. Section 4 provides some characteristics of the corpora used. Section 5 describes the conducted experiments. Section 6 presents the obtained results and their evaluation. Section 7 draws the conclusions and points to the possible directions of future work.

2 Related Work

A wide range of approaches have been proposed to tackle the AP task, with a variety of feature types and feature representations used. In order to promote studies on author profiling (AP) and other authorship identification-related tasks, the PAN evaluation campaign,[3] which is held as part of the CLEF conference, has been organized annually since 2013. It is constantly gaining much attention of researchers from around the world. In this section, we will focus on the winning approaches of each edition of the PAN evaluation campaign.

In the first edition of PAN in 2013 [3], the task consisted in identifying the author's age and gender based on blog posts written in the English and Spanish languages. The work by López-Monroy *et al.* [4] is the overall winner of this year competition, even though their system was ranked second in the individual evaluation on both the English and Spanish datasets. Their approach consisted in using the second order representation based on relationships between documents and profiles. The best approach on the English dataset [5] used ensemble-based classification on a large feature set, including structural, part-of-speech (POS),

[3] `http://pan.webis.de` [last access: 17.07.2016]. All other URLs in this document were also verified on this date.

and text difficulty features, when for Spanish, the best performing approach relied on content-based, style-based, and topic-based features [6].

The second PAN edition in 2014 [7] also focused on determining the author's age and gender. The provided dataset was composed of blog posts, tweets, and social media texts written in both English and Spanish, as well as hotel reviews written in English. As in the previous year, the approach that used the second order representation [8] outperformed other submitted systems.

In 2015 [9], the task aimed at predicting age, gender, and five personality traits: extroversion, stability, agreeableness, conscientiousness, and openness. This year task was limited to tweets, but was extended to four different languages: English, Spanish, Dutch, and Italian. Álvarez-Carmona *et al.* [10] who approached the task using second order profiles and latent semantic analysis (LSA) achieved the best results on the English, Spanish, and Dutch datasets. The best results on the Italian dataset were obtained using stylistic features represented by character and POS n-grams [11].

The focus of the recent 2016 shared task [12] has shifted towards cross-genre age and gender identification covering the English, Spanish, and Dutch languages, that is, the training corpus was on one genre (tweets), while the test set was on another genre (blog posts for English and Spanish, and reviews for Dutch). The best performing system [13] used combinations of stylistic features such as function words, POS, emoticons, punctuations marks, along with the second order representation.

As one can see, feature representation plays a crucial role in achieving high performance in this task. The second order representation based on relationships between documents and profiles led to the best results in all PAN editions. Taking it into account, we focus on an alternative feature representation based on a neural network, which we explain in detail in the next section.

The only work in all PAN editions that used distributed representations of words, namely, word2vec embeddings [14, 15], to tackle the AP task is that by Bayot and Gonçalves [16]. They used the word2vec model trained only on Wikipedia dumps without using the training corpus; this may be the cause for their modest results.

The doc2vec algorithm for learning neural network-based document embeddings is widely used in natural language processing (NLP) tasks, e.g., in text classification, sentiment analysis, information retrieval, etc. [1, 2]. However, to the best of our knowledge, the only work that has been done on AP using document embeddings is our previous research [17]. While the primary goal of that work was to evaluate the effect of pre-processing on learning document embeddings, in this work we focus on evaluating different parameters of the doc2vec algorithm itself, on comparing the doc2vec method with the state of the art, and on comparing the neural network-based features with the traditional features when using the same classifier. In addition, in this paper we address both single- and cross-genre AP settings.

Here, we learn the doc2vec model for author profile identification of anonymous texts; however, this approach can be also used for author demographics identification from other types of textual data, such as source codes [18, 19]. Identification of author's personality from his/her source code is gaining much interest nowadays in automatic source code analysis, which led to the organisation of the first shared task in this filed.[4]

3 Methodolody

The pre-processing performed in this work include standardizing non-standard language expressions, that is, replacing slang words, contractions, abbreviations, and emoticons by their corresponding normalized language expressions. In our previous research [17], we showed that this pre-processing strategy improves the quality of a neural network-based feature representation when used for the AP task.

In order to obtain neural network-based document embeddings, we use the doc2vec algorithm introduced in [1]. It learns features from the corpus in an unsupervised manner and provides a fixed-length feature vector as output. Then, the output is fed into a machine-learning classifier. A framework for learning document vectors is shown in Figure 1.

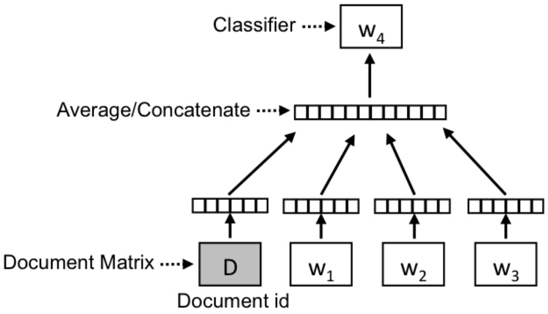

Fig. 1. Framework for learning document vectors. Adapted from [1].

Document vectors are asked to contribute to the prediction task of the next word given many contexts sampled from the document. Each document is mapped to a unique vector represented by a column in a document matrix D. Typically, the document vectors are initialized randomly and in the process of training capture semantics as a side effect of the prediction task.

It is usually recommended to train the doc2vec model several times with unlabeled data while exchanging the input order of the documents. Each iteration of the algorithm is called an epoch, and its purpose is to increase the quality of the output vectors. The selection of the input order of the documents is usually done by a random number generator.

[4] `http://www.autoritas.es/prsoco/`

In this work, instead of initializing the vectors randomly, we use a fixed number generator (fixed seed). Moreover, we apply the Fisher-Yates shuffle algorithm [20] with a fixed seed for exchanging the order the documents are input in each epoch of the training process. In this way, we ensure the reproducibility of the experiments.

4 Datasets

For the evaluation of neural network-based document embeddings in single-genre author profiling (AP), first, we conducted experiments on the PAN AP 2015 training corpus [9] under 10-fold cross-validation. The corpus is composed of Twitter messages in English, Spanish, Dutch, and Italian.

The English and Spanish training datasets are labeled with age and gender, whereas the Dutch and Italian datasets are labeled only with gender. The following age classes are considered: 18–24, 25–34, 35–49, and 50+. The distribution of age and gender over the number of authors can be seen in Table 1.

Table 1. Age and gender distribution over the PAN AP 2015 training corpus.

		English	Spanish	Dutch	Italian
Total		152	100*	34	38
Age	18–24	58	22	–	–
	25–34	60	46	–	–
	35–49	22	22	–	–
	50+	12	10	–	–
Gender	Male	76	50	17	19
	Female	76	50	17	19

* PAN AP 2015 overview [9] mentions 110; however, in fact the corpus contains 100 documents, and other papers in PAN AP 2015 proceedings, such as [10, 11, 21, 22], report 100 as well.

Then, the experiments were conducted on the PAN AP 2016 training corpus [12] under single-genre setting (with 10-fold cross-validation). The PAN AP 2016 corpus consists of Twitter messages in English, Spanish, and Dutch. The PAN 2016 English and Spanish training datasets are labeled with age and gender, when the Dutch dataset is labeled only with gender. The distribution of age and gender over the number of authors in the PAN AP 2016 corpus can be seen in Table 2.

As one can see comparing Tables 1 and 2, the PAN AP 2016 corpus contains more documents, and there are more age classes than in the PAN AP 2015 corpus. Both corpora are perfectly balanced in terms of represented gender classes; however, they are highly unbalanced in terms of age classes. Figure 2 presents the alternative view of age distribution over the PAN AP 2015 and 2016 training corpora.

Table 2. Age and gender distribution over the PAN AP 2016 training corpus.

		English	Spanish	Dutch
Total		426*	250	384
Age	18–24	26	16	–
	25–34	135	64	–
	35–49	181	126	–
	50–64	78	38	–
	65+	6	6	–
Gender	Male	213	125	192
	Female	213	125	192

* PAN AP 2016 overview [12] mentions 428; however, the corpus contains two empty documents.

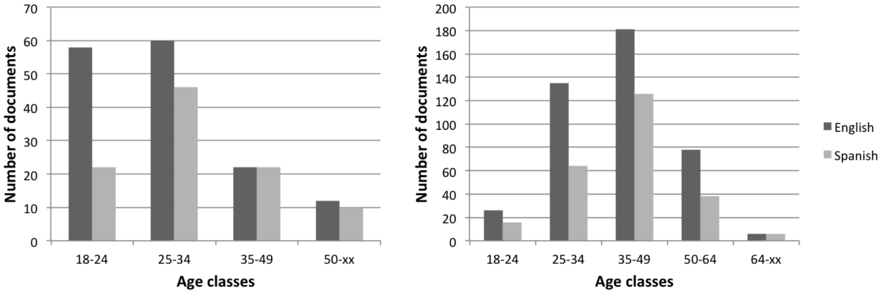

Fig. 2. Age distribution over the PAN AP 2015 (left) and 2016 (right) training corpora.

Finally, we carried out experiments under cross-genre AP conditions, that is, we used a training corpus on one genre, while the test set was on another genre. As training corpus we used the PAN AP 2016 corpus and a subset of the PAN AP 2014 training corpus as test dataset, since the PAN 2016 test corpus is currently not available due to the policies of the PAN organizers. The used subset of the PAN AP 2014 corpus is composed of English and Spanish blog posts and social media, as well as of English reviews. The distribution of age and gender over the number of authors in the used subset of the PAN AP 2014 corpus is shown in Table 3. As one can see, this corpus is also highly unbalanced in terms of age classes.

5 Experimental Settings

In order to standardize non-standard language expressions mentioned in Section 3, we used the dictionaries of shortened vocabulary introduced in [23]. Moreover, we converted all characters to lowercase and separated each document with a new line.

Table 3. Age and gender distribution over the PAN AP 2014 training corpus.

		Blog posts		Social media		Reviews
		English	Spanish	English	Spanish	English
Total		147	88	7,746	1,272	4,160
Age	18–24	6	4	1,550	330	360
	25–34	60	26	2,098	426	1,000
	35–49	54	42	2,246	324	1,000
	50–64	23	12	1,838	160	1,000
	65+	4	4	14	32	800
Gender	Male	74	44	3,873	636	2,080
	Female	73	44	3,873	636	2,080

In the case of cross-gender setting, we learned the doc2vec model from the training corpus (PAN AP 2016) and obtained a neural network-based distributed representation for each training text sample with the doc2vec algorithm [1]. In run-time, we inferred a vector for each previously unseen text sample in the test corpus (PAN AP 2014) using the doc2vec model previously learned from the training data.

However, for the single-genre setting, instead of using disjoint test and training corpora, we used one corpus (PAN AP 2015 or 2016) with 10-fold cross-validation experimental design. In this case we learned the doc2vec model from the whole corpus (not from the 90% of the corpus used for training in each fold of the 10-fold cross-validation setting) and obtained a neural network-based distributed representation for each text sample. We realize that this is not a clean experimental design, but we believe that other systems conducted their experiments in this way (given that typically WEKA [24] was used in those works for 10-fold cross-validation [25, 26]), and replicated this experimental design to be able to compare our results with the previous work.

Then, these distributed vector representations were used to train a classifier. We conducted experiments using the Scikit-learn [27] implementation of the LR classifier. This classifier with default parameters has previously given good performance on high-dimensional data [28, 29]. We generated different classification models for each of the aspects of an author profile, i.e., one model for the age profile and another one for the gender profile. For the 10-fold cross-validation experiments, we used the function from the Scikit-learn Python module that returns the accuracy for each of the 10 folds. The overall accuracy was calculated as the average of the 10 scores.

As we have mentioned in Section 3, the doc2vec method implements a neural network-based unsupervised learning algorithm that builds distributed representations of fixed length from texts [1]. In this work, we used a freely available implementation of the doc2vec algorithm included in the GENSIM[5] Python module. The implementation of the doc2vec algorithm requires the following three

[5] https://radimrehurek.com/gensim/

parameters: the number of features to be returned (length of the vector), the size of the window that capture the adjacent words, and the minimum frequency of words to be included into the model. The values of these parameters depend on the corpus.

In order to narrow down the search of the parameters of the algorithm, similarly to [15, 17], we performed a grid search over the following fixed ranges: vector length from 100 to 300 with step 100, window size from 3 to 15 with step 1, and minimum frequency from 3 to 4 with step 1. The optimal parameters for the PAN 2015 and 2016 corpora are shown in Table 4. The optimal parameters for age do not always correspond to the optimal parameters for gender. For a given corpus, we selected a single set of parameters that provided the best average accuracy between age and gender, i.e., we did not optimize the parameters for age and gender separately. When evaluating the document embeddings performance on the test dataset—the PAN AP 2014 training corpus, we used the parameters selected by 10-fold cross-validation on the training data—the PAN AP 2016 training corpus.

Table 4. Optimal parameters of the doc2vec algorithm for the PAN AP 2015 and 2016 corpora.

Parameter	Vector length	Window size	Minimum frequency
PAN AP 2015 corpus			
English	100	14	4
Spanish	100	5	3
Dutch	100	4	3
Italian	100	12	3
PAN AP 2016 corpus			
English	100	10	3
Spanish	100	9	4
Dutch	100	5	4

6 Experimental Results

In order to evaluate the efficiency of the doc2vec model on the AP task, we compared it with existing state-of-the-art approaches. Since there is no official competition of AP approaches with the settings considered in our work, and since undertaking an exhaustive literary research to identify the top-performing approaches was infeasible, we considered several top-performing systems of the PAN 2015 or 2016 AP competition and examined the corresponding papers to see whether, in addition to the official PAN AP task results, they reported results with the settings addressed in our work. Some of them did; of these, we selected those that reported the highest results:

- For the single-genre AP task on the PAN AP 2015 corpus: the work by Sulea and Dichiu [30], the fifth top system at PAN 2015, which used the author type/token ratio (verbosity rate) and tf-idf weighting scheme.
- For the single-genre AP task on the PAN AP 2016 corpus: the work by Busger *et al.* [13], the winning system at PAN 2016, which used stylistic features and second order feature representation.
- For the cross-genre AP task trained on the PAN AP 2016 corpus and tested on the PAN AP 2014 corpus: the work by Modaresi *et al.* [31], the second top system at PAN 2016, which used logistic regression classifier with stylistic and lexical features.

As baselines, we considered word unigram-based (bag-of-words model) and character 3-gram-based (bag-of-cn-grams model, $n = 3$) approaches, which are commonly believed to be highly predicative for the AP task, independently of language [9, 32]. We used our own implementation of the bag-of-words and bag-of-c3-grams approaches, using a logistic regression (LR) classifier.

Table 5 compares our results with the state-of-the-art approaches on the PAN AP 2015 training corpus under single-genre setting (with 10-fold cross-validation) in terms of accuracy for age and gender classification for each language. Here, our method in all but one cases outperformed bag-of-words and bag-of-c3-grams baselines. However, it was below the state of the art [30].

Table 5. Single-genre results (accuracy, %) for age and gender classification on the PAN AP 2015 training corpus. LR stands for logistic regression classifier. The best results for age and gender for each language are in bold.

Approach	English		Spanish		Dutch	Italian
	Age	Gender	Age	Gender	Gender	Gender
Sulea and Dichiu [30]	**75.65**	**78.94**	**73.00**	**88.00**	**76.47**	**78.94**
LR on bag of words	57.71	61.96	47.00	66.00	44.17	63.33
LR on bag of c3-grams	63.00	58.75	52.00	70.00	49.17	65.83
LR on doc2vec (our)	65.00	69.08	56.00	62.00	56.67	70.00

Table 6 shows comparison of the results on the PAN AP 2016 corpus under single-genre setting. In this experiment, our method outperformed both the baseline and the state-of-the-art approaches for all considered cases; note that the state-of-the-art approach here performed very weakly in comparison with the baselines for gender classification, while we achieved higher improvement for gender than for age.

Tables 7 and 8 show cross-genre results, for the English and Spanish languages, respectively. Unlike under single-genre conditions, in cross-genre setting improvement in accuracy was mostly achieved for age and not for gender, regardless of the language or genre of documents. However, in this setting the state-of-the-art method performed weakly for age classification.

Table 6. Single-genre results (accuracy, %) for age and gender classification on the PAN AP 2016 training corpus.

Approach	English		Spanish		Dutch
	Age	Gender	Age	Gender	Gender
Busger *et al.* [13]	45.73	70.67	48.99	70.85	72.13
LR on bag of words	41.55	72.56	47.63	72.00	71.56
LR on bag of c3-grams	39.20	72.80	48.82	66.40	74.97
LR on doc2vec (our)	**46.01**	**76.98**	**50.44**	**77.20**	**75.54**

Table 7. Cross-genre results (accuracy, %) for English age and gender classification trained on the PAN AP 2016 corpus and tested on the PAN AP 2014 corpus.

Approach	Blog posts		Social media		Reviews	
	Age	Gender	Age	Gender	Age	Gender
Modaresi *et al.* [31]	38.78	**84.35**	20.00	**60.00**	15.24	**60.67**
LR on bag of words	35.37	65.31	26.61	50.50	23.85	55.34
LR on bag of c3-grams	**46.26**	55.10	27.49	50.81	22.57	58.49
LR on doc2vec (our)	35.37	54.42	**29.80**	49.85	**23.92**	51.75

Table 8. Cross-genre results (accuracy, %) for Spanish age and gender classification trained on the PAN AP 2016 corpus and tested on the PAN AP 2014 corpus.

Approach	Blog posts		Social media	
	Age	Gender	Age	Gender
Modaresi *et al.* [31]	40.91	**77.27**	16.27	**59.51**
LR on bag of words	29.55	70.45	32.23	55.66
LR on bag of c3-grams	30.68	61.36	**32.39**	56.84
LR on doc2vec (our)	**42.20**	64.77	31.29	55.90

Table 9 summarizes the comparison with the state of the art and the baselines. As we have stated in the introduction, our contribution is two-fold: on the one hand, we suggest a new method meant to outperform the state of the art; on the other hand, we suggest the features meant to improve the performance of a given classifier. Accordingly, in this table, *Method* refers to the comparison of our results with the state of the art; *Features* refers to the comparison of our results with the baseline features using the same classifier (LR). *Extraction* refers to whether the feature extraction was performed on the entire corpus, before splitting it into training and test corpora, or only on the training portion of the corpus. *Size* refers to the size of the training corpus. The value of '+' indicates that our results were better, '–' that they were worse, and '±' indicates varying comparison results. As we have mentioned above, in some cases the published state-of-the-art results are below our experiments with the baseline features; for those cases, when our doc2vec method outperformed the baseline features, it automatically outperformed those state-of-the-art results. Such trivial success cases are marked as '(+)'.

Table 9. Summary of the comparison.

				Method		Features	
Extraction	Setting	Size	Corpus	Age	Gender	Age	Gender
entire	single	small	2015	−	−	+	+
		large	2016	(+)	(+)	+	+
training	cross		2016/2014	(+)	−	±	−

From this table one can observe that the doc2vec features outperformed the baseline features in single-genre setting, for which the experiment design included feature extraction from the entire corpus, both training and test portions, and gave varying results in cross-genre setting with feature extraction from the training corpus only.

As to the state of the art, our method outperformed only very weak methods, which, still, are the best ones reported in the literature so far. This is not surprising because we did not optimize our method but mainly aimed only at a clear comparison of the neural network-based features with the traditional ones, applying a commonly-used classifier.

7 Conclusions and Future Work

Author profiling (AP) is the task of identifying author demographics based on his or her writings. This is useful for security, marketing, and forensics applications. Recently, the interest in the task of AP has increased steadily, which to a large extent is caused by the annual organization of the PAN AP shared task with a high number of participating teams.

Machine-learning methods are commonly used to identify common stylistic patterns of the authors that share the same profiling aspects. In this work, we applied an approach based on neural network-based document embeddings for the identification of author's age and gender. We used the doc2vec algorithm to learn neural network-based document embeddings and evaluated their performance on the PAN AP 2014–2016 corpora under both single- and cross-genre AP conditions. Our method in certain settings outperformed the state-of-the-art approaches for the AP task.

The contribution of this work is two-fold:

- First, we compare the document-embedding features with traditional features, using the same machine-learning algorithm, and show that the former ones are better in some settings;
- Second, we compare our method with the state-of-the-art approaches and show that it outperforms those approaches under some AP conditions.

The obtained results, in line with the previous work in the field, indicate that feature representation is important for obtaining high performance in this task. Given the same learning algorithm (logistic regression), neural network-based document embeddings used as feature representation in many cases improved the results as compared with the baseline features. Namely, our features outperformed the baseline features in single-genre settings; in those experiments, features were extracted form the entire corpus, including the training and test portions. Moreover, our method outperformed state of the art in those cases when that state of the art was weaker than our baselines.

One of the directions for future work will be to examine the robustness of neural network-based document embeddings on other AP corpora. We will also evaluate the doc2vec-based AP methods using other types of feature as input data representation for the doc2vec method, such as n-grams of words, semantic relations of different types [33, 34], syntactic dependency-based n-grams of various types [35–37], part-of-speech tags [38], and different categories of character n-grams [39, 40].

Acknowledgments

This work was partially supported by the Mexican Government (CONACYT projects 240844 and 20161958, SNI, COFAA-IPN, SIP-IPN 20151406, 20161947, 20161958, 20151589, 20162204, and 20162064).

References

1. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning. ICML '14 (2014) 1188–1196
2. Dai, A., Olah, C., Le, Q.: Document embedding with paragraph vectors. CoRR **abs/1507.07998** (2015)
3. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: CLEF 2013 Labs and Workshops, Notebook Papers. Volume 1179. (2013)
4. López-Monroy, A.P., Montes-y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Villatoro-Tello, E.: INAOE's participation at PAN'13: Author profiling task. In: Working Notes Papers of the CLEF 2013 Evaluation Labs. CLEF '13, CEUR (2013)
5. Meina, M., Brodzińska, K., Celmer, B., Czoków, M., Patera, M., Pezacki, J., Wilk, M.: Ensemble-based classification for author profiling using various features. In: Working Notes Papers of the CLEF 2013 Evaluation Labs. CLEF '13, CEUR (2013)
6. Santosh, K., Bansal, R., Shekhar, M., Varma, V.: Author profiling: Predicting age and gender from blogs. In: Working Notes Papers of the CLEF 2013 Evaluation Labs. CLEF '13, CEUR (2013)
7. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd author profiling task at PAN 2014. In: CLEF 2014 Labs and Workshops, Notebook Papers. Volume 1180. (2014) 898–927

8. López-Monroy, A.P., Montes-y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L.: Using intra-profile information for author profiling. In: Working Notes Papers of the CLEF 2014 Evaluation Labs. CLEF '14, CEUR (2014)
9. Rangel, F., Celli, F., Rosso, P., Pottast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at PAN 2015. In: CLEF 2015 Labs and Workshops, Notebook Papers. Volume 1391., CEUR (2015)
10. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y-Gómez, M., Villaseor-Pineda, L., Jair-Escalante, H.: INAOE's participation at PAN'15: Author profiling task. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. Volume 1391 of CLEF '15., CEUR (2015)
11. González-Gallardo, C.E., Montes, A., Sierra, G., Núñez-Juárez, J.A., Salinas-López, A.J., Ek, J.: Tweets classification using corpus dependent tags, character and POS n-grams. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. Volume 1391 of CLEF '15., CEUR (2015)
12. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2016)
13. Busger Op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., Nissim, M.: GronUP: Groningen user profiling. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. Volume 1609 of CEUR Workshop Proceedings., CLEF and CEUR-WS.org (2016) 846–857
14. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. Computing Research Repository **abs/1301.3781** (2013)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems 26. (2013) 3111–3119
16. Bayot, R., Gonçalves, T.: Author profiling using SVMs and word embedding averages. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. Volume 1609 of CEUR Workshop Proceedings., CLEF and CEUR-WS.org (2016) 815–823
17. Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J., Sanchez-Perez, M.A., Chanona-Hernandez, L.: Improving feature representation based on a neural network for author profiling in social media texts. Computational Intelligence and Neuroscience **2016** (2016)
18. Sidorov, G., Ibarra Romero, M., Markov, I., Guzman-Cabrera, R., Chanona-Hernández, L., Velásquez, F.: Detección automática de similitud entre programas del lenguaje de programación Karel basada en técnicas de procesamiento de lenguaje natural [Automatic detection of similarity of programs in Karel programming language based on natural language processing techniques (in Spanish, abstract in English)]. Computación y Sistemas **20** (2016) 279–288
19. Sidorov, G., Ibarra Romero, M., Markov, I., Guzman-Cabrera, R., Chanona-Hernández, L., Velásquez, F.: Measuring similarity between Karel programs using character and word n-grams. Programming and Computer Software **43** (2017) (in press)
20. Ronald, F., Frank, Y.: Statistical tables for biological, agricultural and medical research. 3rd edn. Oliver & Boyd, London (1948)
21. Kocher, M.: UniNE at CLEF 2015: Author profiling. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. Volume 1391 of CLEF '15., CEUR (2015)

22. Nowson, S., Perez, J., Brun, C., Mirkin, S., Roux, C.: XRCE personal language analytics engine for multilingual author profiling. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. Volume 1391 of CLEF '15., CEUR (2015)
23. Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J., Fócil-Arias, C.: Compilación de un lexicón de redes sociales para la identificación de perfiles de autor [Compiling a lexicon of social media for the author profiling task] (in Spanish, abstract in English). Research in Computing Science **115** (2016)
24. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explorations **11** (2009) 10–18
25. Villena Román, J., González Cristóbal, J.C.: DAEDALUS at pan 2014: Guessing tweet author's gender and age. In: CLEF 2014 Labs and Workshops, Notebook Papers. Volume 1180 of CLEF '14. (2014) 1157–1163
26. De-Arteaga, M., Jimenez, S., Duenas, G., Mancera, S., Baquero, J.: Author profiling using corpus statistics, lexicons and stylistic features. In: CLEF 2013 Labs and Workshops, Notebook Papers. Volume 1179 of CLEF '13. (2013)
27. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning. (2013) 108–122
28. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "How old do you think i am?"; A study of language and age in Twitter. In: Proceedings of the 7$^{\text{th}}$ International AAAI Conference on Weblogs and Social Media, AAAI Press (2013)
29. Maharjan, S., Solorio, T.: Using wide range of features for author profiling. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. Volume 1391 of CEUR Workshop Proceedings., CEUR (2015)
30. Sulea, O.M., Dichiu, D.: Automatic profiling of Twitter users based on their tweets. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. Volume 1391 of CEUR Workshop Proceedings., CEUR (2015)
31. Modaresi, P., Liebeck, M., Conrad, S.: Exploring the effects of cross-genre machine learning for author profiling in PAN 2016. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. Volume 1609 of CEUR Workshop Proceedings., CLEF and CEUR-WS.org (2016) 970–977
32. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. AAAI (2006) 199–205
33. Markov, I., Mamede, N., Baptista, J.: A rule-based meronymy extraction module for Portuguese. Computación y Sistemas **19** (2015) 661–683
34. Markov, I., Mamede, N., Baptista, J.: Automatic identification of whole-part relations in Portuguese. In: Proceedings of the 3$^{\text{rd}}$ Symposium on Languages, Applications and Technologies. Volume 38., Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik (2014) 225–232
35. Posadas-Durán, J., Markov, I., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Gelbukh, A., Pichardo-Lagunas, O.: Syntactic n-grams as features for the author profiling task. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. Volume 1391 of CEUR Workshop Proceedings., CEUR (2015)
36. Gómez-Adorno, H., Sidorov, G., Pinto, D., Markov, I.: A graph based authorship identification approach. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. Volume 1391 of CEUR Workshop Proceedings., CEUR (2015)

37. Sidorov, G., Gómez-Adorno, H., Markov, I., Pinto, D., Loya, N.: Computing text similarity using tree edit distance. In: Proceedings of the Annual Conference of the North American Fuzzy Information processing Society and 5^{th} World Conference on Soft Computing. NAFIPS '15 (2015) 1–4
38. Gómez-Adorno, H., Pinto, D., Montes, M., Sidorov, G., Alfaro, R.: Content and style features for automatic detection of users' intentions in tweets. In: Proceedings of the Ibero-American Conference on Artificial Intelligence, Springer (2014) 120–128
39. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL: Human Language Technologies. NAACL-HLT '15, Association for Computational Linguistics (2015) 93–102
40. Markov, I., Gómez-Adorno, H., Sidorov, G., Gelbukh, A.: Adapting cross-genre author profiling to language and corpus. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. Volume 1609 of CEUR Workshop Proceedings., CLEF and CEUR-WS.org (2016) 947–955