Custom Document Embeddings Via the Centroids Method: Gender classification in an Author Profiling task

Notebook for PAN at CLEF 2018

Roberto López-Santillán[1], Luis Carlos González-Gurrola[1], and Graciela Ramírez-Alonso[1]

Facultad de Ingeniería, Universidad Autónoma de Chihuahua, Circuito No. 1, Nuevo Campus Universitario, Apdo. postal 1552, Chihuahua, Chih., México. C.P. 31240
{jrlopez,lcgonzalez,galonso}@uach.mx

Abstract. According to *Smart Insights*[1], out of the 7.5 billion persons in total population of the world, there are 4 billion Internet users, and out of those an outstanding 3.19 billion are active social media users. In a report by the U.S. Internet Crime Complaint Center, only in 2016 *Identity theft*, *Extortion* and *Harassment or violence threads* stand out among the most frequently reported cyber-crime events[2]. The ***Author Profiling*** (AP) task might be useful to counteract this phenomena by profiling cyber-criminals. AP consists in detecting personal traits of authors within texts (i.e. gender, age, personality).
In the current report we describe a method to address the AP problem, which is one of the three shared tasks evaluated, as an exercise in digital text forensics at PAN 2018 within the CLEF conference (Conference and Labs of the Evaluation Forum). Our approach blends *Word Embeddings* (WE) and the *Centroids Method* to produce *Document Embeddings* (DE), that deliver competitive results predicting the gender of authors, over a dataset comprised of text posts from Twitter®. Specifically, in the testing dataset our proposal achieve an accuracy of 0.78 for English language users, and on average (for English, Spanish and Arabic languages users) it reaches an Accuracy score of 0.77.

Keywords: Author Profiling · Word Embeddings · Document Embeddings.

1 Introduction

Author Profiling (AP) is the task of discovering features like gender, age and psychological traits in persons, by analyzing their language expressions. A methodology to obtain personal profiles of individuals with high accuracy would be useful

[1] https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

[2] https://www.statista.com/statistics/184083/commonly-reported-types-of-cyber-crime/

in areas like customer service, attention of neurological disorders (e.g. autism) and cyber-crimes among others [1]. The proficiency to accurately profile the author(s) of plagiarism, identity theft, Internet sexual predatory activities or even terrorist attacks, has become a matter of critical importance. On the other hand, influential companies such as Amazon, Netflix, Google, Apple amidst other, use several Machine Learning (ML) algorithms to address and create new demand among their clients and to attract new ones, based on profiling their customers [2].

A fair amount of works have attempted solving this problem, several shared task events are held each year to test accuracy of new algorithms. These works report competitive results when predicting the gender or age group of individuals [2]. In an attempt to standardize and set a context framework, world wide conferences in the field of Natural Language Processing (NLP) are organized frequently. Among those events, the PAN evaluation lab on digital text forensics, organized within the CLEF Initiative (Conference and Labs of the Evaluation Forum), is held each year. For the 2018 edition[3] the conference posed shared tasks in mainly 3 different efforts: *Author Identification*, *Author Obfuscation* and ***Author Profiling*** [3].

Machine learning (ML) is a computer science field that has re-gained strength in the last years with the advent of Deep Learning (DL), a subfield of ML that has obtained strong achievements in image and speech recognition, computer vision and as of lately NLP. NLP deals with the problem of how computers can understand Human natural language [4]. AP may be viewed as a sub-task of NLP, and it should be approached as such.

Departing from traditional NLP strategies, we propose a different approach. Word embeddings (WE) are a type of DL application that project natural language words (vectors) into a n-dimensional space. The distance between vectors represent a similitude value within a semantic context. A WE algorithm proposed by Mikolov et al. [5] called ***Word2Vec*** is currently used in multiple NLP tasks with amazing results. Word2Vec uses a Bag of Words approach, but it retains the order of the tokens within the original text. This method provides additional semantic information, richer in inner structure components, which might increase the accuracy of ML algorithms to predict personal traits in people.

Although WEs deliver state-of-art results in NLP tasks, such as text classification or language translation, more difficult assignments like Sentiment Analysis (SA) (which tries to identify positive from negative user opinions) or AP, are not benefited in the same way. WEs capture syntactic and semantic information, nonetheless the latter is caught with less sensitivity. To vectorize whole documents, literature suggests other techniques, such as the *Centroids Method* [6]. This approach considers a document as the sum of its words, hence the WE of each word in a sentence, a paragraph or even a whole document is composed by an aggregate function like maximum, minimum, or a weighted average, producing a single vector for the entire document, with a similar dimension shape as the WE of its words. This design delivers good results when training ML algorithms

[3] https://pan.webis.de/clef18/pan18-web/author-profiling.html

to learn the target of such documents, like topics or themes. To identify the gender, the age or the personality profile of the person behind such manuscript, the *Centroids Method* is more limited.

The fields of *Computer Science* (CS), *Linguistics* and *Psychology* must come together in a transversal strategy to tackle this problem. Some studies have merged the power of *computation* and *linguistics* to identify words or combinations of them, so several large lexicons have been developed, with tokens of statistical significance that can identify gender, age or personality of authors with high accuracy [7]. Such exercises give the certainty that words have a discerning power to differentiate men from women, youngsters from old persons or introverts from extroverts. Then it is hypothesize that WEs could lend their potential to an aggregate strategy that could produce document vectors as well. Taking this into account, our approach to the AP task blends WEs in an aggregate strategy, to produce distributed representations of whole texts (DEs), which we use to train our model.

2 Related Work

Age and gender are often associated with the style of writing. This means that the style might change with age and is strongly associated with the gender of persons. The size of the texts available on a single author for training is known to affect the outcome of the classifiers. Even though larger texts are preferred, short samples like the ones found in social media platforms like Twitter®, might be useful in the AP task. Several studies show that the accuracy to predict gender and age in short texts declined little when compared to larger samples available [2]. Adorno et. al. proposed in [8] an approach that joined the Doc2Vec algorithm (a spin-off of Word2Vec focusing on sentences rather than words [9]) and the use of a Neural Network (NN). Their hypothesis relies on the idea that standardizing nonstandard language expressions used by several authors could render better accuracy performance on the AP task. Moreover the detection of emotions in posts from social media outlets, could result in better accuracy predicting features, as proposed by [10], where the emotions detected in on-line posts might help to predict gender accurately.

The 2015 PAN AP shared task included *personality* as a feature to predict [2]. This endeavor is better addressed as a regression problem since the prediction is a rational value. At that year several strategies were chosen to addresses the problem. For instance the approach of Pervaz et. al. focused on the stylistic and thematic properties of the dataset [11]. On the other hand lvarez-Carmona et. al. [12] chose Second Order Attributes (SOA) and Latent Semantic Analysis (LSA) to enhance the discriminative skills of their algorithms. Most of the teams attained competitive results using state-of-the-art ML algorithms. The most frequent classifiers used were: Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF) [2].

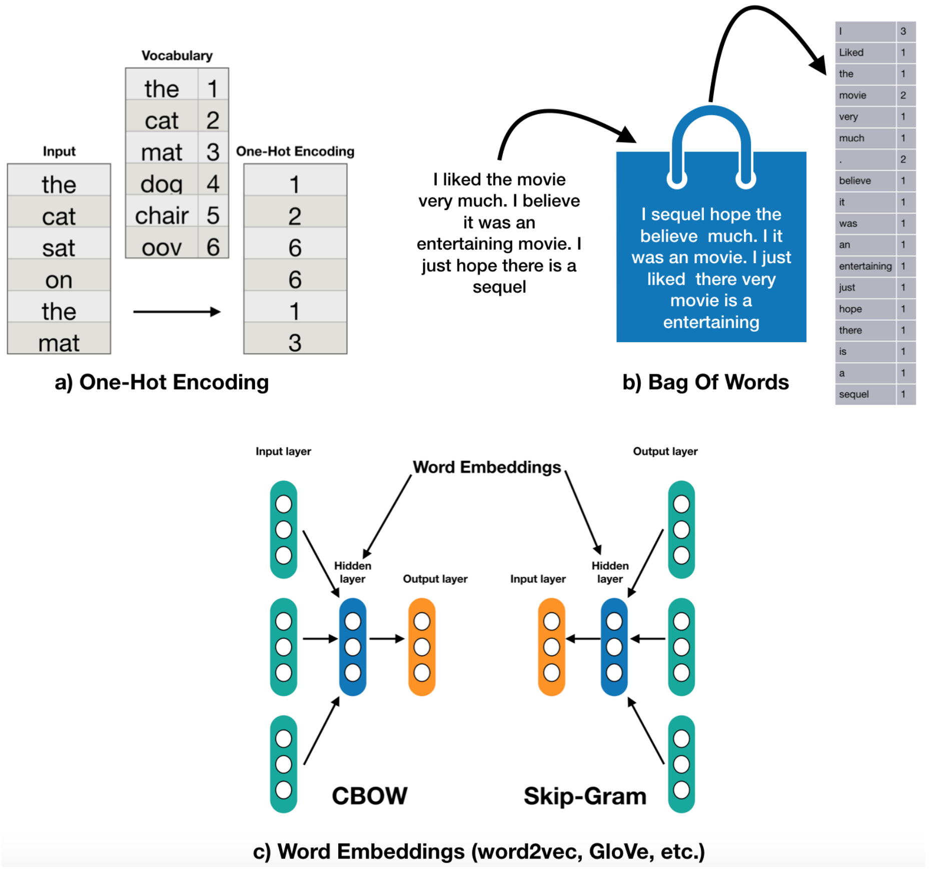
To properly approach the AP problem, it is essential to understand the fundamental blocks of language. The vast majority of currently spoken idioms around the world, are build around tokens known as words. Even though syntax, grammar and other language constructs may vary from one dialect to another, the single elements like "words" remain the same. Nonetheless, Western Latin idioms are based on similar character sets, it is possible to "tokenize" languages like Arabic in order to obtain single tokens equivalent to the former. All this is relevant because works like the one published by Schwartz et. al. [7], have established a statistically significant set of words for the English language, that have the power to categorize authors that use frequently these terms, by gender, age group and personality traits. They propose a mixture of *Linguistics* and *Computer Science* algorithms to determine which words in the English idiom are statistically significant, to be used with discriminative enough capabilities. As it is implied in this study, the popularity of social media has produced huge amounts of available data from all kinds of persons around the globe. This enormous potential datasets allows to develop new ML algorithms which might reveal the intricateness within the written language.

In order to perform NLP tasks, words need to be treated like numeric entities, not only as identifiers, but they must represent through their values something meaningful, to the term and to the context they are being used on. There are several ways to represent words in ML/NLP tasks, for instance *One Hot Encoding*, *Bag-Of-Words* or *Word Embeddings* are 3 of the most used architectures to represent words in NLP efforts. Figure 1 demonstrates the basic structure of these word-coding methods.

For *text classification*, *textual similarities* or even *translation* activities, WEs by themselves deliver extraordinary results. In tasks such as *sentiment analysis* or *author profiling*, it is required to generate vectors for phrases or whole documents. One of the most straightforward techniques for vectorization of whole documents is called the *Centroids Method*. As proposed by Kusner et. al. in [6], in order to perform accurate document classifications, texts should be projected into a n-dimensional space where a distance between them can be computed. A document embedding or vector is generated by calculating a weighted average of its WEs, then vectorial distances (Euclidean, Canberra, etc) between them allows for proper document classification.

As of late, WEs are the preferred method to engage NLP tasks, since they deliver state-of-the-art results in tasks such as *text classification*, *translation*, *text generation* or *sentiment analysis* (SA). Seyed et. al. in [13] tried to enhance WEs in a SE task by adding useful information to the word vectors. Their idea consists in calculating vectors for Part of Speech (POS) elements within the text. A POS is the category of the word, for instance *nouns*, *verbs*, *adjectives* or *pronouns*. Moreover, a lexicon vector is also computed and concatenated as well; there are several lexicons, particularly in the English language with proven discriminative properties as demonstrated by [7].

Our approach is partially based on the aforementioned studies. We differentiate our strategy by creating our own distributed representations for the vocabulary and the POS tags. Also, a preliminary experiment showed that lexicons did



a) One-Hot Encoding

b) Bag Of Words

c) Word Embeddings (word2vec, GloVe, etc.)

Fig. 1: Several ways to encode words from texts. In *a)*, *oov* (out of vocabulary) is used when a token does not belong to the dictionary.

not add accuracy when joined to the DEs, hence they were discarded. A more in depth explanation is presented in the next section.

3 Methodology

The goal for the AP shared task at PAN 2018 was to classify subjects by gender in three different languages: *English*, *Spanish* and *Arabic*. In our implementation only a SVM classifier was employed, due to delivering the best results in preliminary runs. The model was trained offline on the provided training dataset, then was uploaded into a Virtual Machine (TIRA), available at a server in Bauhaus University at Weimar Germany. This environment allows different models (heterogeneous programming languages) to run on a common platform, thus making the evaluation of the shared task easier to assess [14].

The employed strategy is an assemble approach which fuses *a)* WEs, *b)* the *Centroids Method* to produce DEs and *c)* the *tf-idf* (term frequencyinverse document frequency) weighting scheme. *tf-idf* is a statistical value computed for each term in a document. It helps establish the importance of a term within a corpus. The more a word appears in a text, the more the *tf-idf* value increases. For terms with high frequency but few discriminative power (e.g. *the*, *and*), an offset value is computed. The *tf-idf* value is vastly used in *information retrieval* tasks [15]. Equation (1) shows how to compute a *tf-idf* value within a corpus.

$$\text{tf-idf}_{t,d} = (1 + \log \text{tf}_{t,d}) \cdot \log \frac{N}{\text{df}_t} \tag{1}$$

A new vocabulary of WEs was crafted from the dataset using the *Skip-gram* algorithm, then a *tf-idf* value was calculated for each term within the context of each collection of posts of all individuals. For example the *tf-idf* value for the word *"drink"* will be different among persons in the dataset (people use words differently). Next the DEs were generated for each person in the dataset (one per person), by computing the weighted average (using the *tf-idf* value of each word) of all WEs in every set of documents, as shown in formula (2). Finally the DEs were used to train the SVM classifier. A distribution of specimens in the training dataset is depicted in table 1

$$\text{DEs-WAvg} = \frac{\sum_{n=1}^{i}(w_n * tfidf[w_n])}{\sum_{n=1}^{i} tfidf[w_n]} \tag{2}$$

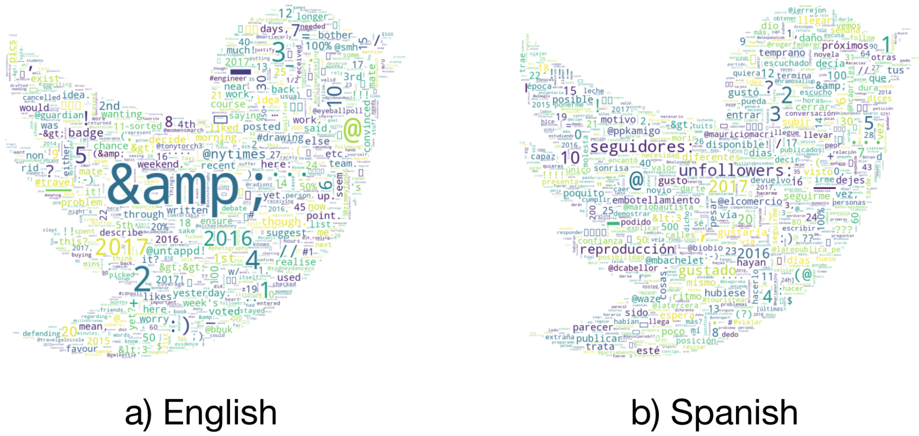
Table 1: Distribution of samples in the PAN-2018 dataset.

	Samples
English	3000
Female	1500
Male	1500
Spanish	3000
Female	1500
Male	1500
Arabic	1500
Female	750
Male	750

Table 2: Parameters used to train the words and their POS tags embeddings, using the *Skip-gram* model.

Word Embeddings	
size	300
min_count	15
window	15
sample	0.05
iter	5
negative	15
POS tag Embeddings	
size	20
min_count	1
window	5
sample	0.05
iter	5

A Tokenization process designed for Twitter® posts[4] was used to produce the individual terms in the dataset. This procedure allows to retrieve each term in the dataset as a single entity. Using a specific tool to produce tokens in a social network environment, allowed us to attain context specific terms such as *”bit.ly”*, *”:-)”* or *”#TuesdayThoughts”*, which might be more discriminative than single words. Figure 2 shows an analysis done on the most frequent Twitter® tokens used on the training dataset.



a) English

b) Spanish

Fig. 2: Words clouds for English and Spanish terms in the training dataset.

Inspired by the work in [13], we performed a POS tagging procedure on the dataset for the English language. For this process we used the NLTK POS tagger[5], which uses a Greedy Averaged Perceptron (GAP) algorithm to compute the POS tags of each word. Subsequently each word in the dataset was replaced by its POS tag and the same *Skip-gram* algorithm was applied to generate embeddings for the POS labels. WEs from the dataset were enhanced by concatenating their POS tags vectors. To produce the Document Embeddings (DE) for every individual on the dataset, a weighted average was also computed for the enhanced WEs in the collection of posts of each individual. This embeddings share the same dimensionality of the enhanced vectors from single words $(300 + 20)$. Figure 3 shows the proposed model. For the *English* language the POS vectors were attached to the original WEs before the weighted averaging. For the *Spanish* and *Arabic* languages, no POS vectors were attached (DEs of 300 dimensions), because multilingual POS tagger tools did not deliver useful labels for these idioms. A translation approach to apply the POS tags strategy in these languages did not deliver an advantageous trade off between accuracy and running time.

[4] https://github.com/dlatk/happierfuntokenizing

[5] https://www.nltk.org/book/ch05.html

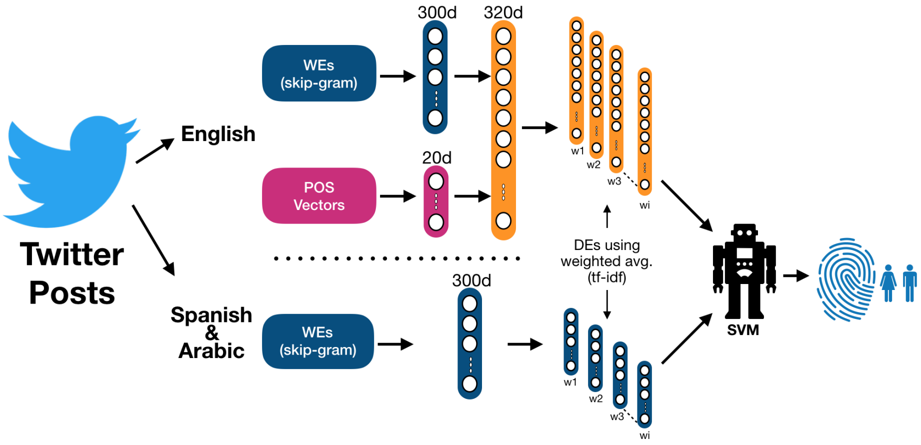


Fig. 3: Architecture of the proposed technique.

Table 2 shows the parameters used to train the WEs of words and their POS. The criterion to choose the parameters was based on a randomized grid search.

For the classification stage, a Support Vector Machine (SVM) algorithm was chosen. Although not entirely new (introduced by Vapnik in the 90's of last century), SVMs are currently acknowledged as one of the most used and effective ML algorithms [16]. A preliminary set of runs to test the best classifier for the AP task, showed that SVM was the fittest option among other methods such as *Random Forest* and *Extra Trees*. SVMs are useful in problems with hard separability, by using the so called *kernel* trick, they approximate a higher dimensional projection by computing variants of the *dot product*, which is both accurate enough and computationally tractable [16]. A full Grid Search (GS) was performed to find the best parameters for the SVM, table 3 shows the best parameters found by the GS.

Table 3: Parameters explored in the Grid search of the SVM classifier.

Parameter	**Range**
C	[0.001, 0.01, 0.1, 0.5, 0.9, 1, 10]
gamma	[0.001, 0.01, 0.1, 1]
kernel	linear’, ’rbf’, ’poly’, ’sigmoid’
degree	[1, 2, 3]
coef0	[0.0 - 10]
best set	(C=10, degree=2, gamma=1, kernel=’poly’, coef0=0.0)

4 Results

In order to evaluate our method in the training phase, a 10-fold cross-validation strategy was selected. Table 4 shows the performance of the SVM classifiers over the training dataset. An average of the accuracy over the three languages was computed to produce an overall performance value.

Table 4: Performance in training dataset.

Language	**Accuracy**
English	0.7990
Spanish	0.7713
Arabic	0.7953
Average	**0.7885**

Table 5: Performance in testing dataset.

Language	**Accuracy**
English	0.7847
Spanish	0.7677
Arabic	0.7760
Average	**0.7761**

For the testing stage, we ran our model through the platform TIRA[6]. The results attained on the testing data showed little decrease compared to the results in the training step, this suggests the method is robust. Table 5 demonstrates the results achieved on testing. In both stages (training and testing), *English* language was best classified, this might be due in part to the POS tagging information that was added only in this idiom. Furthermore the strategy for *Spanish* and *Arabic* classification did not use POS tags, which might suggest the Document Embeddings produced were not as rich in semantic information as those enhanced by POS tags vectors. Despite having more samples (3000 individuals, whilst Arabic has 1500), the *Spanish* task resulted in the lowest accuracy overall, whether it was on training or testing data. This suggests that the use of words and topics in the *Spanish* language might be more uniform than in the *Arabic* idiom.

A more comprehensive list (from all participant teams) of methodologies and results is explained by Rangel et. al. in [17], thus our methodology might be assessed more properly within this context.

5 Conclusion

The Author Profiling task is very important for the current online way of life. The ongoing mission to produce faster and more accurate algorithms, takes us in new directions, to explore new options, to test new approaches. Even tough state-of-the-art results are good enough for some applications, there is still a lot of room for improvement. The method proposed in this report, shows that

[6] http://www.tira.io/

fusing ”old” ways with novel ones, might be a good strategy for the upcoming future. Moreover, given the limited scope of the available datasets for this type of tasks, it might be a good idea to work parallel in devising new techniques to produce more comprehensive datasets in a faster way.

References

1. M. Hildebrandt, S. Gutwirth, M. Hildebrandt, and S. Gutwirth, *Profiling the European Citizen: Cross-Disciplinary Perspectives.* Springer Publishing Company, Incorporated, 1 ed., 2008.
2. F. Rangel, F. C. P. Rosso, M. Potthast, B. Stein, and W. Daelemans, “Daelemans, w.: Overview of the 3rd author profiling task at pan 2015,” in *CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, CEUR-WS.org (Sep 2015)*, 2015.
3. E. Stamatatos, F. Rangel, M. Tschuggnall, M. Kestemont, P. Rosso, B. Stein, and M. Potthast, “Overview of PAN-2018: Author Identification, Author Profiling, and Author Obfuscation,” in *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 9th International Conference of the CLEF Initiative (CLEF 18)* (P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Nie, L. Soulier, E. Sanjuan, L. Cappellato, and N. Ferro, eds.), (Berlin Heidelberg New York), Springer, Sept. 2018.
4. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* The MIT Press, 2016.
5. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), pp. 3111–3119, Curran Associates, Inc., 2013.
6. M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, “From word embeddings to document distances,” in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML’15, pp. 957–966, JMLR.org, 2015.
7. H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, “Personality, gender, and age in the language of social media: The open-vocabulary approach,” *PLOS ONE*, vol. 8, pp. 1–16, 09 2013.
8. H. Gómez-Adorno, I. Markov, G. Sidorov, J. P. Posadas-Durán, M. A. Sanchez-Perez, and L. Chanona-Hernández, “Improving feature representation based on a neural network for author profiling in social media texts,” *Comp. Int. and Neurosc.*, vol. 2016, pp. 1638936:1–1638936:13, 2016.
9. Q. V. Le and T. Mikolov, “Distributed representations of sentences and documents,” *CoRR*, vol. abs/1405.4053, 2014.
10. F. Rangel and P. Rosso, “On the identification of emotions and authors’ gender in facebook comments on the basis of their writing style,” in *Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM 2013) A workshop of the XIII International Conference of the Italian Association for Artificial Intelligence (AI*IA 2013), Turin, Italy, December 3, 2013.*, pp. 34–46, 2013.
11. I. Pervaz, I. Ameer, A. Sittar, and R. M. A. Nawab, “Identification of author personality traits using stylistic features: Notebook for pan at clef 2015.,” in *CLEF*

(Working Notes) (L. Cappellato, N. Ferro, G. J. F. Jones, and E. SanJuan, eds.), vol. 1391 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015.

12. M. Á. Á. Carmona, A. P. López-Monroy, M. Montes-y-Gómez, L. V. Pineda, and H. J. Escalante, "Inaoe's participation at pan'15: Author profiling task," in *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*, 2015.
13. S. M. Rezaeinia, A. Ghodsi, and R. Rahmani, "Improving the accuracy of pre-trained word embeddings for sentiment analysis," *CoRR*, vol. abs/1711.08609, 2017.
14. M. Potthast, T. Gollub, F. Rangel, P. Rosso, E. Stamatatos, and B. Stein, "Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling," in *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)* (E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, and E. Toms, eds.), (Berlin Heidelberg New York), pp. 268–299, Springer, Sept. 2014.
15. J. Ramos, "Using tf-idf to determine word relevance in document queries."
16. S. Marsland, *Machine Learning: An Algorithmic Perspective, Second Edition.* Chapman & Hall/CRC, 2nd ed., 2014.
17. F. Rangel, P. Rosso, M. Montes-y-Gómez, M. Potthast, and B. Stein, "Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter," in *Working Notes Papers of the CLEF 2018 Evaluation Labs* (L. Cappellato, N. Ferro, J.-Y. Nie, and L. Soulier, eds.), CEUR Workshop Proceedings, CLEF and CEUR-WS.org, Sept. 2018.