ISSN 2007-9737

Enhancing Deep Learning Gender Identification with Gated Recurrent Units Architecture in Social Text

Bassem Bsir, Mounir Zrigui

University of Monastir, LATICE Laboratory Research Department of Computer Science, Tunisia

bsir.bassem@yahoo.fr, mounir.zrigui@fsm.rnu.tn

Abstract. Author profiling consists in inferring the authors' gender, age, native language, dialects or personality by examining his/her written text. This paper represent an extension of the recursive neural network that employs a variant of the Gated Recurrent Units (GRUs) architecture. Our study focuses on gender identification based on Arabic Twitter and Facebook texts by investigating the examined texts features. The introduced exploiting a model that applies a mixture of unsupervised and supervised techniques to learn word vectors capturing the words syntactic and semantic. We applied our approach on two corpora of two social media varieties: twitter texts, in which each author is assigned at least 100 tweets, and Facebook corpus containing short texts with an average of 15.77 words per author. The obtained experimental results are comparable to the best findings provided by the best per-forming systems presented in PAN Lab at CLEF 2017.

Keywords. Author profiling, gender identification, deep learning, gated recurrent units (GRUs), twitter, facebook.

1 Introduction

Authorship Analysis aims at extracting information about the authorship of documents from features within these documents. It is based on combining three different techniques, namely Authorship Profiling, Authorship attribution or Identification and Plagiarism Detection. Author profiling is the task of determining the writers' features, such as native language, education, gender, age and personality traits, by understanding their writing styles.

In recent years, author profiling has been a promising field of research in the world of human-machine interaction systems as it presents important opportunities and challenges in various fields such as marketing, analysis of social networks and social computing [40, 8, 31].
For instance, author profiling helps, in crime investigation, identify the perpetrator of a crime by considering the characteristics of his/her writing styles. However, it does also allow studying the behavior of potential terrorists or groups advocating racial segregation [11].

Author profiling is also employed in marketing to detect, for example, the profiles of potential consumers of a website for online sale to help decision makers choose their marketing strategies [24]. Besides, author profiling is applied in e-learning. It allows understanding and detecting the users' behavior in this educational environment and thus permits architects of learning sites and designers of educational tools and materials to create and organize the content of learning according to the learners' needs.

With the development of social networks, Facebook has recently become an interesting target for research providing rich information to study and model user's behavior. Indeed, user's contributions and activities constitutes a valuable insight about individual behavior, opinions, experiences and interests, which makes it an important and rich source for the extraction of corpus serving as the basis for several research works. [31] employed users data on Facebook to explore the feasibility of predictive personality modeling in order to support future intelligent systems. Other studies, such as [37] and [21] utilized Facebook data for sentiment classification, authorship identification [17] and speech recognition [20].

In our approach, we applied deep learning approaches to learn the abstract and higher level

features of the document and identify author characteristics. We explored extension of the recursive neural network that employs a variant of the Gated Recurrent Units (GRUs) architecture to tackle the problem of gender detection.

2 State of the Art

The writing style of the words reflects the authors' mental, social and even the physical as well as the psychological states. Indeed, statistical studies exploring the stylistic features of a text have begun one century ago with T. Mendenhall. Afterwards, several probabilistic machine learning and deep learning approaches were introduced.

The studies of stylometric demonstrated also that individuals can have a footprint linked to their writing style. Indeed, [25] defined an author profile as "a set of length L of the most frequent n-grams with their normalized frequencies". Thus, the profile of an author can be considered as the ordered set of pairs {(x1; f1); (x2; f2)… (xL; fL)} Of the L most frequent n-grams xi and their normalized frequencies fi.

[27] represented 604 documents from the British National Corpus (BNC) in the form of trees whose roots are words sets or POS. They obtained 80 % accuracy to infer the gender of the author.

Researchers in 2002, [12] investigated authorship gender attribution mining from e-mail text documents based on the structural characteristics and gender preferential language features. They obtained 70.2 % precision rate for gender detection.

Different researchers used a variety of powerful machine learning and statistical algorithms to build a classification model by employing these features vectors.

[28] analyzed a corpus of 71,000 blogs incorporating almost 300 million words. They utilized a learning algorithm, called Multi-Class Real Winnow (MCRW), to learn models that classify blogs according to the author's gender and age. They got 43.8% and 86% for age and gender accuracy prediction, respectively.

The authors, in 2016, built a Cross genre Author Profiling System (CAPS) which considered parts of speech, collocations, connective words and various other stylometric features to differentiate between the writing styles of male and female authors as well as between different age groups. Their system attained 74.36% accuracy for gender identification [7].

For age and gender profiling, in 2013 [29] employed (SVM) Classifiers together with Principal Component Analysis (PCA). They concluded that content-based features are more discriminative than other features (style based and content based features). SVM classifiers allowed obtaining 82.6% gender prediction accuracy.

In 2016, [36] approached the task of gender detection with combinations of stylistic features such as function words, parts of speech, emoticons and punctuations signs. They trained their models with SVM and obtained 55.75 accuracy for gender identification in English data PAN 2016 competition.

Liblinear classifier combined with Concise Semantic Analysis (CSA) was used by [4]. It achieved a best accuracy of 65.72 for the age prediction in English blogs data PAN 2013 competition [4].

To predict the author gender, [19] reached a good accuracy equal to 0.8283 by using REP Tree (a fast decision tree learning algorithm) as a classifier and the sentence based, character, syntactic and Word features.

[38] relying on a corpus of 3524 Vietnamese Weblog pages of 73 bloggers and exploiting 298 features, they obtained an accuracy of 82.12 for occupation and 78.00 for location dimension by employing IBK.

[2] used the gensim Python library for LDA topic extraction with SVM classifiers. Their result proved that the topic models are useful in developing author-profiling systems. [10] predicted the gender, age and personality traits of Twitter users in four different languages (Spanish, English, Italian and Dutch). They accounted stylistic features represented by character Ngrams and POS N-grams to classify tweets. They applied Support Vector Machine (SVM) with a linear kernel called LinearSVC and obtained 83.46% for gender detection.

In [13], researchers applied SVM classifier and neural network on TF-IDF and verbosity features. Results showed that SVM classifiers are better for English datasets.

They proved that neural networks performed better for Dutch and Spanish datasets. For English, the best findings were obtained using a TF-IDF at character level combined with the verbosity feature. Their results are almost similar to those provided by [6]. They got 61.5 gender accuracy and 41.03 age accuracy.

Based on the corpus collected from Twitter for four different languages (Arabic, English, Portuguese and Spanish), [3] obtained 85.99 for gender identification accuracy by combining character n-grams (with n between 3 and 5) TF-IDF word n-grams (with n between 1 and 2).

[33] obtained 70.02 by using logistic regression with combinations of character, word and POS n-grams, emojis, sentiments, character flooding, and lists of words per variety in PAN 2017 competition [41].

Deep learning based approaches have recently dominated the state of the art in well-studied problem among NLP researchers. New approaches have also emerged to improve authorship analysis which involves many layers of nonlinear information processing in deep neural networks [22, 49]. Subsequently, deep learning-based approaches have demonstrated remarkable results for text classification and have performed well for phrase level and message level sentiment classification [26, 46].

In 2016 and for the first time, [30] have employed deep learning techniques for author profiling. They described a big gap between traditional machine learning models and deep learning models in the participant teams evaluated in the VarDial2016 workshop. They attempt to narrow this gap using convolutional neural networks (CNN) as a first approach for author profiling.

In 2017, many researches applied deep learning approaches: Recurrent Neural Networks, Convolutional Neural Networks as well as word and character embeddings. However, [17] generated embeddings of the authors' text based on sub word character n-grams. These representations were classified using deep averaging networks. They got 79.19 for gender identification in PAN 2017 competition [41].

[45] used TF-IDF and a Deep Learning model based on Convolutional Neural Networks.

As features, he used a matrix of 2-grams of letters with punctuation marks, beginning and ending with 2-grams. They obtained 72.07 as precision of gender identification in PAN 2017. The same model based on Convolutional Neural Networks was used by [47] who explored parameters such as the size of the input of the network, the size of the convolutional kernels, the number of kernels and the type of input. They found experimentally that sequences of words performed better than sequences of characters as input for the CNN. They obtained 76% of accuracy in the test partition in PAN 2017 competition.

Experiments on automatic classification of users according to latent attributes, such as gender and age, were performed on a wide range of resources including Facebook, telephone conversations [42], blogs [44] and Twitter [16, 51]. The problem that always arises, in this context, is how to label user's profiles to obtain their age and gender. To solve this problem, two techniques were proposed. The first one is based on applying a manual labeling. For instance, [35] constructed manually a dataset by means of a fine-grained annotation effort of more than 3000 Dutch Twitter users.

To determine the author's gender, [9] sampled users from the Twitter stream and used links to blogging sites indicated in their profile. Some approaches used, for example, lists of male and female names by analyzing Facebook texts [15].

The second manner of construction corpora consists in taking into account information provided by the authors themselves. For instance, in blog platforms, [28] studied the effect of age and gender on the style of writing in 71,000 blogs, while [38] used the corpus of 3524 Vietnamese Weblog pages of 73 bloggers or e-mail messages [14].

Though Arabic is spoken by almost 400 million people, research works of authorship analysis performed on Arabic texts are not numerous. For example, the study of [1] focused on author identification. The researchers constructed a corpus written by 20 authors and 20 messages written by each one.

The second work investigating the multilingual messages was carried out by [14].

In their research, authors collected Arabic and English e-mails written by 1033 English people and other e-mails written by 1030 Egyptian Arabic

Table 1. The distribution of comments according to different gender and age categories

	Gender		
	Female	**Male**	**Total**
18-24	214	369	583
25-34	339	415	854
35-49	686	235	921
50-64	468	633	1101
>65	496	489	985
Total	2203	2241	4444

Table 2. PAN CLEF 2017 training corpus statistic for gender detection

Authors	**Tweets**	**Language varieties**	**Genders**
2400	240k	Gulf, Levantine, Maghrebi, Egypt	1200 M; 1200 F

speakers. They studied several demographic and psychometric features for author profiling [48].

In [5], authors construct an Arabic corpus taken from Facebook for age and gender detection. They used different techniques for classification combined with linguistic, stylistic and structural features to determine the gender and age of author. They obtained 71.52 accuracy to infer the gender of the author and 53.08% for age detection.

3 Data

Our study was performed on two varieties of corpus. The first corpus consists of short texts (one comment per author that does not exceed 15 words at the average) proposed by [5]; whereas the second corpus presented by the conference PAN @ CLEF 2017 contains long texts (100 tweet per author) [41].

The first corpus is based on the Web texts, especially Facebook. In fact, they labeled manually gender with the help of a dictionary of proper nouns (ambiguous nouns have been discarded) and by visiting each profile and looking at the photo, description, etc. This first corpus composed of 4444 validated profiles contains 70121 words with an average of 15.77 words per profile. Approximately, 50% of the texts have a text length shorter than 10 words. The corpus was balanced in terms of gender, while it was imbalanced in terms of age.

The second dataset used to train the proposed models is the official PAN@CLEF 2017 Author Profiling Training Corpus. It was collected from Twitter. For each tweet collections, the texts, taken from the Arabic language, are composed of tweets, 100 tweets per authors. In this work, we concatenate the user tweets to have a unique instance. For the Arabic language, four varieties were used in this corpus: Egypt, Gulf, Levantine and Maghrebi.

4 Proposed Approach

The standard recurring networks obtain their strength from their memory capacity for sequence processing thanks to recurrent connections, bringing the context of the preceding element into the sequence, and their ability to be trained through back propagation through time. But, learning long term dependencies using simple recurrent neurons may provoke problems like exploding or vanishing gradients [23].

To solve such issues, recent approaches have modified the simple neuron structure in order to learn more efficiently dependencies over longer intervals. In this study, we evaluate the performance of such neural networks, namely Gated Recurrent Units (GRUs). The latter (GRUs) is a specific recurrent neural network (RNN) architecture that is well employed to learn from experience to predict unknown author's gender.

Gated Recurrent Units (GRUs) use both past and future information stored by both the forward and backward networks, regardless of the type of network.

Indeed, this bi-directional model employs the activation functions of Softmax to calculate its output.

ISSN 2007-9737

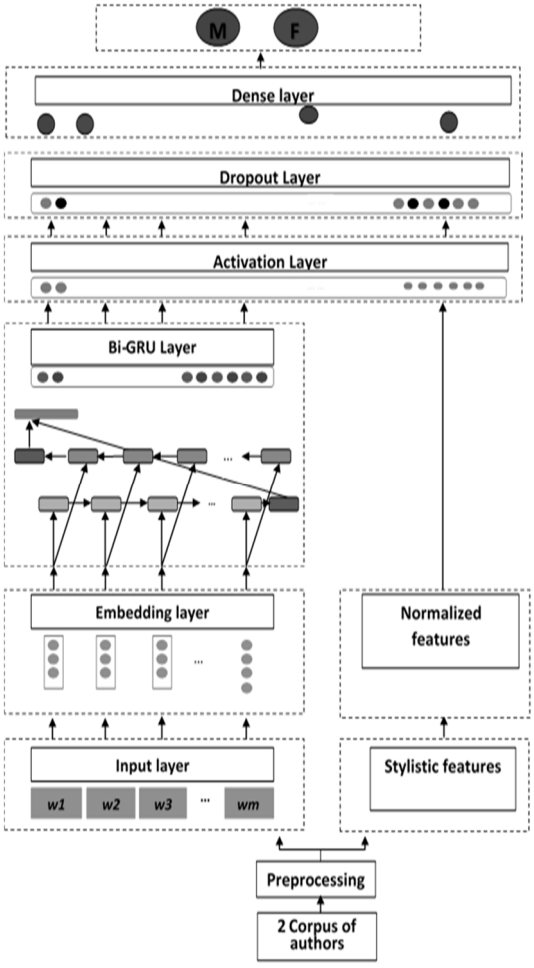

Fig. 1. General architecture of the proposed system

It first computes the hidden state forward by applying the function and the hidden state backwards of an entity in the sequence:

$$h_f^{(t)} = \tanh(U_f x^{(t)} + W_f h_f^{(t-1)} + b_f), \quad (1)$$

$$h_b^{(t)} = \tanh(U_b x^{(t)} + W_b h_b^{(t+1)} + b_b), \quad (2)$$

$$\hat{y}^{(t)} = \mathrm{softmax}(V_b h_b^{(t)} + V_f h_f^{(t)} + c). \quad (3)$$

The transition from hidden a state to another is based on using the Gated recurrent units. Although RNNs can theoretically capture long-term dependencies, they are very hard to actually train. [23] showed that learning long term dependencies in recurrent neural networks through gradient decent is a difficult task. Gated recurrent units are designed to provide more persistent memory, making it easier for RNNs to capture long term dependencies.

In this paper, we apply a bi-directional gated RNN (GRUs) as an optimization algorithm Adam. In the following section, we will present our system of determining the author's gender from Arabic corpus, as shown in Figure1.

4.1 Preprocessing

To model statistically the language, we pre-processed the comments in the extracted corpus utilizing five regular expressions. The first expression permitted identifying Arabic texts and omitting those written in foreign languages. However, in the second expression, we deleted duplicated letters (جداااجدا). The third expression omitted diacritization; whereas the fourth deleted comments in form of hypertexts (advertisements links, images ...). At final two or more successive spaces, resulting from the application of the rules already mentioned or they are founded in the original texts, were replaced by one space.

4.2 Input Layer

In this layer, each unit of the input layer passes its assigned value directly to the Embedding layer.

4.3 Stylistic Features

We considered the comments as a vector in a multi-feature space. Then, we used the obtained labeled vectors to construct our classification model. In fact, six types of style related features were determined [5].

– **Arabic Lexical features:**
These features were obtained by frequency calculations. We distinguished the number of words appearing once and those appearing twice, the average length of sentences, the number of sentences, the number of verbal sentences as well as the number of negative sentences starting with negation. (ل ل،لا،مل،ام،ســي ن).

– **Arabic n-grams:**

We used bigrams and trigrams to detect Arabic authors' profiles.

– **Arabic syntax features:**

It consists in labeling comments and giving each word, in the extracted corpus, its syntactic category (proper name, common noun, verb, adjective, etc).

– **Arabic character frequencies:**

We calculated the frequency of letters, punctuation marks (number of colons, exclamation marks, question marks and commas), uppercase characters, lowercase characters, numerical characters, alphabetical characters , numbers and symbols like (@, #, and, %, *).

– **Bag of smileys:**

We listed 57 manually created emoticons (:-), , : (, ; -), ;) , ;-P, ;P , :P , :-p, :p...) which express different sentiments (happiness, sadness, anger, etc.)

– **Arabic stop words:**

We defined 4 groups of stop words, such as personal pronouns ("he", هو),("she", هي),("they", هم), demonstratives (this هذا, these هؤلاء , there أولئـــك), prepositions (from من , toإلـى , in فـــي , about عن) and interrogatives ("where", أيـن,"who", من).

4.4 Normalized Features

To remove the impact of different scales, we normalized the value of each feature using AMZD normalization [39]:

$$(|X|) - (10^{n-1}) * (|A|) / 10^{n-1}. \quad (4)$$

4.5 Embedding Layer

We employed the default continuous bag of words embedding strained through a shallow neural network, as shown by [51] to initialize the embedding layer of the RNN. This bag represents essentially words by a vector to identify the similarity between words. In fact, the search for similarity is based on the word 2vec techniques. Indeed, word2vec is a combination of two methods (CBOW (Continuous bag of words) and Skip-gram model) which learn weights acting as word vector representations [51].

The Word2Vec model was formed by the corpus of Arabic Wikipedia with 4 million tweets extracted in order to enrich the vocabulary list with words that do not exist in Wikipedia. For training, we used the skip-gram neural network model with a window of size 5 (1 center word + 2 words before and 2 words after), a minimum frequency of 15 and a dimension equal to 300.

4.6 BI-GRU Layer

The GRU cell has two gates: an update gate (z) and a reset gate (r). It reduces the three gates defined in the LSTM networks (the input, forget and output gates).

The following equations represent the gating mechanism in a GRU:

$$z^{(t)} = \sigma(W_z h^{(t-1)} + U_z x^{(t)} + b_z), \quad (5)$$

$$r^{(t)} = \sigma(W_r h^{(t-1)} + U_r x^{(t)} + b_r), \quad (6)$$

$$\hat{h}^{(t)} = \tanh(r^{(t)} \odot W_{\hat{h}} h^{(t-1)} + U_{\hat{h}} x^{(t)} + b_{\hat{h}}). \quad (7)$$

4.7 Activation Layer

Most recent deep learning networks have used rectified linear units (ReLUs) for the hidden layers. Most frameworks, like Tensor Flow and TF Learn, simplify the use of ReLUs on the the hidden layers [31].

It computes the following function:

$$f(x_i) = \max(0, x_i). \quad (8)$$

4.8 Drop Out Layer

As the size of our model is relatively big and to avoid overfitting problem, we applied Dropout to control the size of the network and to change the number of the hidden features in the recurrent layers [34]. In fact, dropout involves randomly removing some hidden units in a neural network during the training step while keeping all of them during the testing step. We employed dropout on our softmax layer with p = 0.5.

Table 3. Gender accuracy for PAN and Facebook corpus

	Facebook Corpus	Pan Corpus
Basile et al., 2017 [42]	--	80.06 %
Our system	62.1%	79%

4.9 Dense Layer

To complete the flow of information throughout the two gates generated by the bi-Gru layer (a value between 0 and 1), we used sigmoid as the activation function.

Sigmoid transform the input as follows:

$$y=\frac{1}{1+e^{-(w^{T}x+b)}}. \quad (9)$$

5 Experimentation and Results

To evaluate the prediction accuracy of our method we used the best results of [42] obtained in PAN@CLEF2017 as a baseline method to assess our technique and show its efficiency. Basile and al [41] obtained the best accuracy result as showing in Table 3.

In our method, we used10-fold cross validation. The dataset, for both corpus, was divided at the note level. We separated out 10 % of the training set to form the validation set. This validation set was used to evaluate our bi-directional RNN model. We also employed a maximum of 20 epochs to train our model. The training was performed on an Intel core i7 machine with 16 GB memory.

To choose the optimization algorithm, we trained our model with 20% of randomly dataset from PAN corpus, in 10 epochs. Then we selected the algorithm which achieving the highest accuracy on the development set. We tested seven algorithms: SGD, Adam, RMSprop, Adagrad, Adadelta, Adamax and Nadam. Experiments shows that our model performed with Adam optimizer and it converges to 80 % of accuracy.

Table 3 shows the performance results of our system in the two test datasets used for gender identification. For PAN corpus, we compared our method with the best accuracy obtained in PAN 2017.

Comparing our result with that achieved by the best participant in PAN@CLEF2017, based on the same corpus, we obtain a very encouraging result that shows the effectiveness of deep learning models. By taking the GRUs model on a large amount of data, it reaches 79% for the age identification task.

To show the relationship between GRUs model and the amount of training data, we changed the training corpus, we based on the second corpus extracted from Facebook, which is a corpus very small compared to PAN corpus, GRUs did not show the same performance and obtained an accuracy of 62.1%.

In general, using stylistic models with word embeddings in a GRUs architecture allowed obtaining the best results and proved that bi-directional deep networks is crucial in author profiling task, especially when it's trained on a huge amount of data.

6 Conclusion

This paper proposes a combination of stylistic models with words embeddings and Deep-Learning (GRUs) Neural Network to predict the gender of Twitter and Facebook Arabic authors.

Our bi-directional recurrent neural networks model shows a good performance on gender identification. The obtained results were encouraging, especially for Facebook corpus. This result not so surprising since neural network models had shown efficiency adapting to natural language processing problems (sentiment analysis, text classification…).

As future works of this study, we plan to extend our detection tool to other attributes for Arabic authors like language variety and personality features.

References

1. **Abbasi, A. & Chen, H. (2005).** Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems,* Vol. 20, No. 5, pp. 67–75. DOI: 10.1109/MIS.2005.81.

2. **Poulston, A., Stevenson, M., & Bontcheva, K. (2015).** Topic Models and n–gram Language Models for Author Profiling. *Proceedings of Notebook for PAN CLEF Evaluation Labs.*
3. **Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., & Nissim, M. (2017).** Is there life beyond n-grams? A simple SVM-based author profiling system. *Working Notes of CLEF 2017-Conference and Labs of the Evaluation Forum*
4. **Alvarez-Carmona, M. A., López-Monroy, A. P., Montes-y-Gómez, M., Villasenor-Pineda, L., & Jair-Escalante, H. (2015).** INAOE's participation at PAN'15: Author profiling task. *Notebook for PAN at CLEF.*
5. **Bassem, B. & Zrigui, M. (2017).** An empirical method for evaluation of author profiling framework. *PACLIC 31.*
6. **Bayot, R. & Gonçalves, T. (2016).** Author Profiling using SVMs and Word Embedding Averages. *Notebook for PAN at CLEF´16*.
7. **Bilan, I. & Zhekova, D. (2016).** CAPS: A Cross-genre Author Profiling System. *CLEF,* pp. 824–835.
8. **Bougiatiotis, K. & Krithara, A. (2016).** Author Profiling using Complementary Second Order Attributes and Stylometric Features. *CLEF.*
9. **Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011).** Discriminating gender on Twitter. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics,* pp. 1301–1309.
10. **González-Gallardo, C. E., Montes, A., Sierra, G., Núñez-Juárez, J. A., Salinas-López, A. J., & Ek, J. (2015).** Tweets Classification Using Corpus Dependent Tags, Character and POS N-grams. *Proceedings of CLEF´15 Evaluation Labs.*
11. **Chen, H., Chung, W., Qin, J., Reid, E., Sageman, M., & Weimann, G. (2008).** Uncovering the Dark Web: A Case Study of Jihad on the Web. *Journal of the American Society for Information Science and Technology,* Vol. 59, No. 8, pp. 1347–1359. DOI:10.1002/asi.20838.
12. **Corney, M., De-Vel, O., Anderson, A., & Mohay, G. (2002).** Gender-preferential text mining of e-mail discourse. *Computer Security Applications Conference Proceedings 18th Annual IEEE,* pp. 282–289. DOI: 10.1109/CSAC.2002.1176299.
13. **Dichiu, D. & Rancea, I. (2016).** Using Machine Learning Algorithms for Author Profiling In Social Media. *CLEF,* pp. 858–863.
14. **Estival, D., Gaustad, T., Hutchinson, B., Pham, S., & Radford, W. (2008).** *Author Profiling for English and Arabic Emails*.
15. **Fink, C. R., Chou, D. S., Kopecky, J. J., & Llorens, A. J. (2011).** Coarse- and Fine-Grained Sentiment Analysis of Social Media Text. *Johns Hopkins APL Technical Digest,* Vol. 30, No. 1, pp. 22–30.
16. **Fink, C., Kopecky, J., & Morawski, M. (2012).** Inferring Gender from the Content of Tweets: A Region Specific Example. *ICWSM.*
17. **Franco-Salvador, M., Plotnikova, N., Pawar, N., & Benajiba, Y. (2017).** Subword-based deep averaging networks for author profiling in social media. *Notebook for PAN at CLEF 2017.*
18. **Gencheva, P., Boyanov, M., Deneva, E., Nakov, P., Kiprov, Y., Koychev, I., & Georgiev, G. (2016).** PANcakes Team: A Composite System of Genre-Agnostic Features for Author Profiling. *Proceedings of CLEF´16 Evaluation Labs.*
19. **González-Gallardo, C. E., Montes, A., Sierra, G., Núnez-Juárez, J. A., Salinas-López, A. J., & Ek, J. (2015).** Tweets Classification using Corpus Dependent Tags, Character and POS N-grams. *CLEF´15.*
20. **Graves, A., Mohamed, A. R., & Hinton, G. (2013).** Speech recognition with deep recurrent neural networks. *Acoustics, speech and signal processing. (ICASSP), International Conference on IEEE,* pp. 6645–6649. DOI: 10.1109/ICASSP.2013.6638947.
21. **Hamouda, S. B. & Akaichi, J. (2013).** Social networks' text mining for sentiment classification: The case of Facebook'statuses updates in the Arabic Springer**.** *International Journal of Application or Innovation in Engineering & Management, (IJAIEM),* Vol. 2, No. 5, pp. 470–478.
22. **Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamedet, A. R., Jaitlyal, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012).** Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine,* Vol. 29, No. 6, pp. 82–97. DOI: 10.1109/MSP.2012.2205597.
23. **Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014).** *Empirical evaluation of gated recurrent neural networks on sequence modeling.* arXiv preprint arXiv: 1412. 3555. pp. 1–9.
24. **Jay, B. S. & Raghavendra, B. K. (2010).** A Neural Network based framework for Customer Profiling for Risk analysis. *International Journal of Advanced Computing (IJAC),* Vol. 2, No. 4.
25. **Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003).** N-gram-based author profiles for authorship attribution. *Proceedings of the conference pacific association for computational linguistics, (PACLING).*

26. **Kim, Y. (2014).** *Convolutional neural networks for sentence classification.* arXiv preprint arXiv: 1408. 5882.
27. **Koppel, M., Argamon, S., & Shimoni, A. R. (2002).** Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, Vol. 17, No. 4, pp. 401–412. DOI:10.1093/llc/ 17.4.401.
28. **Koppel, M., Schler, J., & Argamon, S. (2009).** Computational methods in authorship attribution. *Journal of the Association for Information Science and Technology,* Vol. 60, No. 1, pp. 9–26. DOI: 10.1002/asi.20961.
29. **Lim, W. Y., Goh, J., & Thing, V. L. (2013).** Content-centric age and gender profiling. *Proceedings of the Notebook for PAN at CLEF.*
30. **Malmasi, S., Zampieri, M., Ljubesi, N., Nakov, P., Ali, A., & Tiedemann, J. (2016).** Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,* pp. 1–14.
31. **Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. (2013).** Mining facebook data for predictive personality modeling. *Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM'13),* pp. 23–26.
32. **Mars, M., Antoniadis, G., & Zrigui, M. (2011).** @ rabLearn: a Model of NLP Tools Integration in ICALL Systems. *Proceedings on the International Conference on Artificial Intelligence (ICAI),* The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
33. **Martinc, M., Škrjanec, I., Zupan, K., & Pollak, S. (2017).** Pan 2017: Author profiling-gender and language variety prediction. *Notebook for PAN at CLEF 2017.*
34. **Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014).** Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929–1958.
35. **Nguyen, D. P., Gravel, R., Trieschnigg, D., & Meder T. (2013).** How old do you think I am?. A study of language and age in Twitter. *ICWM.*
36. **op Vollenbroek, M. B., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., & Nissim, M. (2016).** Gronup: Groningen user profiling. *Notebook for PAN at CLEF 2016.*
37. **Ortigosa, A., Martín, J. M., & Carro, R. M. (2014).** Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior,* Vol. 31, pp. 527–541. DOI: 10.1016/j.chb.2013.05.024.
38. **Pham, D. D., Tran, G. B., & Pham, S. B. (2009).** Author profiling for Vietnamese blogs. *Processing, In Asian Language, (IALP'09). International Conference on IEEE.* pp. 190–194. DOI: 10.1109/IALP.2009.47.
39. **Patro, S. & Sahu, K. K. (2015).** *Normalization: A Preprocessing Stage.* arXiv preprint arXiv: 1503. 06462.
40. **Rangel, F. & Rosso, P. (2016).** On the impact of emotions on author profiling. *Information processing & management,* Vol. 52, No. 1, pp.73–92. DOI: 10.1016/j.ipm.2015.06.003.
41. **Rangel, F., Rosso, P., & Chugur, I. (2014).** Overview of the 2nd author profiling task at pan. *CLEF evaluation labs and workshop.*
42. **Rangel, F., Rosso-Potthas, M. & Stein, B. (2017).** Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in twitter. *Working Notes Papers of the CLEF.*
43. **Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010).** Classifying latent user attributes in twitter. *Proceedings of the 2nd international workshop on Search and mining user-generated contents. ACM.* DOI: 10.1145/1871985.1871993.
44. **Rosenthal, S., McKeown, K., & Agarwal, A. (2014).** Columbia NLP: Sentiment Detection of Sentences and Subjective Phrases in Social Media. *SemEval@ (COLING),* pp. 198–202**.**
45. **Sarawgi, R., Kailash, G., & Yein, Ch. (2011).** Gender attribution: tracing stylometric evidence beyond topic and genre. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning,* Association for Computational Linguistics**.**
46. **Schaetti, N. (2017).** Unine at CLEF 2017: Tf-idf and deep-learning for author profiling. *Working Notes of CLEF 2017-Conference and Labs of the Evaluation Forum*.
47. **Aliaksei, S. & Moschitti, A. (2015).** Learning to rank short text pairs with convolutional deep neural networks. *Proceedings of the 38th International (ACM-SIGIR) Conference on Research and Development in Information Retrieval. ACM.*
48. **Sghaier, M. A. & Zrigui, M. (2017).** Tunisian dialect-modern standard Arabic bilingual lexicon. *Computer Systems and Applications (AICCSA), IEEE/ACS 14th International Conference on IEEE,* pp. 973–979.
49. **Sierra, S. & et al. (2017).** Convolutional Neural Networks for Author Profiling. *Working Notes of the CLEF.*

50. Zouaghi, A., Zrigui, M., & Antoniadis, G. (2008). Compréhension automatique de la parole arabe spontanée. *Traitement Automatique des Langues,* Vol. 49, No. 1, pp. 141–166.

51. Zouaghi, A., Merhbene, L., & Zrigui, M. (2012). Combination of information retrieval methods with Lesk algorithm for Arabic word sense disambiguation. *Artificial Intelligence Review*, Vol. 38, No. 4, pp. 257–269.

Article received on 20/01/2018; accepted on 05/03/2018. Corresponding author is Bassem Bsir.