# A Hierarchical Attention Network for Bots and Gender Profiling

## Notebook for PAN at CLEF 2019

Cristian Onose, Claudiu-Marcel Nedelcu, Dumitru-Clementin Cercel, and Stefan Trausan-Matu

Faculty of Automatic Control and Computers
University Politehnica of Bucharest, Romania
{onose.cristian, claudiu.nedelcu.m, clementin.cercel}@gmail.com, stefan.trausan@cs.pub.ro

**Abstract** Author profiling represents the task of detecting various author aspects, for instance age, gender or personality, by analyzing written text. The bot identification issue is particularly important in today's society given the increase in social media usage and the effect of opinion influencing bots on the public. This paper describes our solution for the Bots and Gender Profiling problem, introduced at PAN 2019. The PAN challenge is a two part multilingual problem, namely for the English and Spanish languages. The first task has the goal of identifying if the author is a human or a bot. For the second task, the system has to detect the gender of human authors. Our solution uses a deep learning model based on Hierarchical Attention Networks (HAN) as well as pretrained word embeddings for text representation. For the first task, the official results show that the model achieves an accuracy score of 0.8943 for English and 0.8483 for Spanish. For the second task, our model obtains 0.7485 accuracy for English and 0.6711 for Spanish.



## 1 Introduction

Author profiling refers to the task of identifying different author traits by analyzing the content and style of written text. Such characteristics can include age, gender or even if the author is real of not. Due to the recent increase in the usage of social media, the task of detecting automatically generated text has seen additional interest. A bot can influence the opinion of users in various areas of interest such as politics, commercial interest or religion. For instance, an example of negative influence was observed during the presidential elections in the United States in 2016 [1].

Herein, we present our approach for the Bots and Gender Profiling competition, a novel issue introduced in the 2019 edition of the PAN evaluation campaign [13,4]. This year, the organizers proposed two tasks for this competition. Initially, the task involves

determining the author of a Twitter feed, namely, the classification between bots and humans. Afterwards, in the case of human authors, we are tasked with determining the gender of the author. Traditionally, approaches for solving these problems use machine learning algorithms with hand crafted or content features: punctuation, grammatical errors and their frequencies, bag of words, n-grams, average sentence length, part-of-speech tagging or hyperlinks [14]. Lately, deep learning methods based on language models, such as word or character embeddings, have been proposed. Most popular architectures include Convolutional Neural Networks, Recurrent Neural Networks (RNN) or RNNs with memory cells such as Long Short-term Memory [15].

Motivated by the recent progress with deep learning, we choose a top performing deep learning architecture, used for text classification, as our model. Specifically, we apply the Hierarchical Attention Networks (HAN) [17] model with pretrained embeddings needed to encode the tweets as input. Recently, HAN architectures have been used to efficiently solve diverse text classification tasks such as identification of dialect varieties [11], satirical texts [16] or style change detection [7]. Note that we did not use any additional data to supplement the dataset during training, however, we rely on the word representations trained on additional datasets. In contrast to traditional machine learning models, where hand crafted features are required, our approach learns the features from the data.

We used the model described in this paper to participate in both tasks by considering them as binary classification problems. To test the performance of the HAN architecture, we experimented with different embeddings. According to the official results, our submission achieved the following accuracy scores: 0.8943 and 0.8483 for task 1, in the case of English and Spanish, respectively; 0.7485 and 0.6711 for task 2 also for English and Spanish. Overall, the performance of our system was ranked as average, a more detailed analysis is presented in Section 5.

The remainder of this paper is structured as follows. In Section 2 we describe the competition dataset and the preprocessing steps applied. Section 3 gives an overview of the word embeddings used in order to represent text as input for our model. In Section 4 we briefly review the methodology behind our solution as well as implementation details, while the results are presented in Section 5. Lastly, Section 6 includes our conclusions.

## 2 Data Description and Preprocessing

The dataset provided contains English and Spanish texts collected from the Twitter social media platform. Each user feed has exactly 100 tweets which are supplied raw, without any preprocessing, meaning that retweets are not removed and the language is not guaranteed. In order to avoid overfitting, the dataset is pre-split between training and validation, for each language, as described in Table 1. For both tasks and languages the dataset is perfectly balanced. This is useful as it ensures that classes are not advantaged or disadvantaged based on their proportions. The limitation of the dataset consists in the small number of items which decreases the effectiveness of deep learning solutions.

**Table 1.** Dataset training and validation sample distribution for each language, provided by the organizers.

| | Language | Class | Training set | Validation set | Total |
|---|---|---|---|---|---|
| Task 1 | English | Bot | 1440 | 620 | 2060 |
| | | Human | 1440 | 620 | 2060 |
| | Spanish | Bot | 1040 | 460 | 1500 |
| | | Human | 1040 | 460 | 1500 |
| Task 2 | English | Female | 720 | 310 | 1030 |
| | | Male | 720 | 310 | 1030 |
| | Spanish | Female | 520 | 230 | 750 |
| | | Male | 520 | 230 | 750 |

Additionally, the number of Spanish samples is lower than the English ones. Our experiments confirm this observation as the model performs better on the English subset.

To improve the learning performance, we cleaned up the dataset as follows. First, we preprocess the data by replacing user tags with the *user* string because they can act as biases for our model. We choose not to remove them because multiple user tagging is a method used by bots in order to attract attention. Similarly, we replace hyperlinks with the *url* string. Furthermore, we change emojis into their textual representation, for instance *grinning face*[1]. As a last step, we remove all punctuation and, for every author, we consider each tweet as a sentence by merging them and ending each tweet with a period sign. This representation is necessary since our model receives as input a large portion of text divided into sentences.

## 3 Word embeddings

A word embedding is a method of encoding text in the form of numerical vectors with the goal of maintaining the natural language relationships in the new vector space. For instance, a well constructed model will capture various semantic and syntactic relations such as meaning, morphology or context. While these representations can be as simple as one-hot vectors, lately, complex neural network models for learning such embeddings have been introduced [9].

Given the small size of our dataset, in our experiments, we choose to use pretrained embedding models as follows. For English we use a model [5] with a 400 word vector size that was trained using word2vec [10] on 400 million Twitter posts. The model excludes tokens that have a frequency lower than 5 with the final model having a vocabulary of around 3 million words. Similarly, for Spanish we use a model [2] that was also trained using word2vec with a minimum word frequency of 5. The corpus used during training consists of around 1.5 billion words created from multiple Spanish web resources. The final embedding model contains nearly 1 million word vectors of dimension 300.

[1] We used the python *emoji* package: https://pypi.org/project/emoji/

**Table 2.** Performance overview of our submitted model with respect to the first ranked submissions as well as the solutions provided by the organizers [13]. The top table contains the results for the bot vs. human task and the bottom one for the gender identification task.

| English | | | Spanish | | |
|---|---|---|---|---|---|
| Rank | Team | Accuracy | Rank | Team | Accuracy |
| 1 | Johansson & Isbister | 0.9595 | 1 | Pizarro | 0.9333 |
| 2 | Fernquist | 0.9496 | 2 | Jimenez-Villar et al. | 0.9211 |
| 3 | Bacciu et al. | 0.9432 | 3 | Mahmood | 0.9167 |
| 6 | Char nGrams | 0.9360 | 11 | Char nGrams | 0.8972 |
| 7 | Word nGrams | 0.9356 | 16 | Word nGrams | 0.8833 |
| 31 | LDSE | 0.9054 | 30 | **Our system** | **0.8483** |
| 35 | Word Embeddings | 0.9030 | 31 | Word Embeddings | 0.8444 |
| 39 | **Our system** | **0.8943** | 35 | LDSE | 0.8372 |
| 57 | Majority | 0.5000 | 46 | Majority | 0.5000 |
| 59 | Random | 0.4905 | 47 | Random | 0.4861 |

| English | | | Spanish | | |
|---|---|---|---|---|---|
| Rank | Team | Accuracy | Rank | Team | Accuracy |
| 1 | Valencia et al. | 0.8432 | 1 | Pizarro | 0.8172 |
| 2 | Bacciu et al. | 0.8417 | 2 | Jimenez-Villar et al. | 0.8100 |
| 3 | Espinosa et al. | 0.8413 | 3 | Srinivasarao & Manu | 0.7967 |
| 20 | Word nGrams | 0.7989 | 18 | Char nGrams | 0.8385 |
| 23 | Char nGrams | 0.7920 | 22 | Word nGrams | 0.7244 |
| 26 | Word Embeddings | 0.7879 | 26 | Word Embeddings | 0.7156 |
| 28 | LDSE | 0.7800 | 35 | LDSE | 0.6900 |
| 39 | **Our system** | **0.7485** | 38 | **Our system** | **0.6711** |
| 51 | Majority | 0.5000 | 44 | Majority | 0.5000 |
| 56 | Random | 0.3716 | 47 | Random | 0.3700 |

## 4 Hierarchical Attention Networks

Hierarchical Attention Networks (HAN) [17] were introduced in order to solve document classification problems. They achieve this by modeling the two level hierarchical structure of documents. The first level is represented by the words that are used to build sentences, and the second one, the sentences that form the document. This model is able to distinguish between the importance of different text sections with respect to the context. The first attention mechanism creates sentence embeddings by encoding the sequences of word embeddings using Bidirectional Gated Recurrent Units (Bi-GRU) [3,6]. Similarly, the second attention layers uses Bi-GRU cells to create an encoding for the document based on the representation from the first attention mechanism. Lastly, based on the resulting document encoding classification is performed.

The model uses two hyper parameters in order to maintain a consistent input across different sized documents: maximum number of words in a sentence (tweet in our case) and maximum number of sentences per document (Twitter feed). While the document sentence size is always 100, defined by the dataset structure we discussed, we choose the number of words by investigating the distribution across the entire dataset. This offers

us an initial value for the hyper parameter which we improved through a grid search. We observe that the performance doesn't improve with a value greater than 10 for the maximum tweet length in words. Likewise, we set the size of the attention layers to be 200. Finally, the model is trained using Adam [8] with a learning rate of $\alpha = 0.0005$ and recommended values for the other hyper parameters. Training is done in batches of 64 until the stop condition is met, namely when no improvement in the validation loss function is observed for two epochs.

## 5 Results

Overall, our solution, combined for all tasks and languages, achieves a middle rank in the competition, more precisely position 31 out of 55 teams. As a per task score we obtain a better rank in the case of the Spanish language. However, the accuracy for English is higher than that of Spanish.

Table 2 present the most relevant ranks and scores in order to better view the solutions performance. For the first task, the best three models achieved roughly $0.06$ and $0.08$ higher accuracy when compared to our solution, for English and Spanish, respectively. Similarly, for the second task, the best three models achieved roughly $0.1$ and $0.13$ higher accuracy. For the Spanish language, in the second task, the model performs poorly in comparison to the other task as well as the other language. Also, the baseline solutions [13], namely low dimensionality statistical embedding (LDSE) [12], various embeddings such as word vectors, character and word n-grams achieve a mid performance between the top models and our solution.

In Figure 1 we present the confusion matrices as a method to identify the classification bias of the model. The model has the tendency to miss classify more frequently bots as human as well as males as females then the reversed counterparts.

## 6 Conclusions

This paper describes our approach for solving the problem of Bots and Gender Profiling on Twitter user feeds that was introduced at the PAN 2019 competition. The task consists of two multilingual tasks (English and Spanish): binary classification between bots and humans, and in the case of a human author, classification as male or female. Our proposed solution is a deep learning model based on the Hierarchical Attention Network (HAN) architecture. In order to be consistent with the HAN model assumptions, we view the user feeds as structured documents and tweets are regarded as sentences in said documents. Language is represented as numerical input for the model with the help of pretrained word embeddings. The official and training results show that our model for English outperforms the one for Spanish. We attribute this decrease in performance to the quality of the word embeddings, more precisely the fact the training corpus used is broader unlike the English model that was trained on Twitter posts.

## Acknowledgments

This work was supported by the 2008-212578 LT-fLL FP7 project.

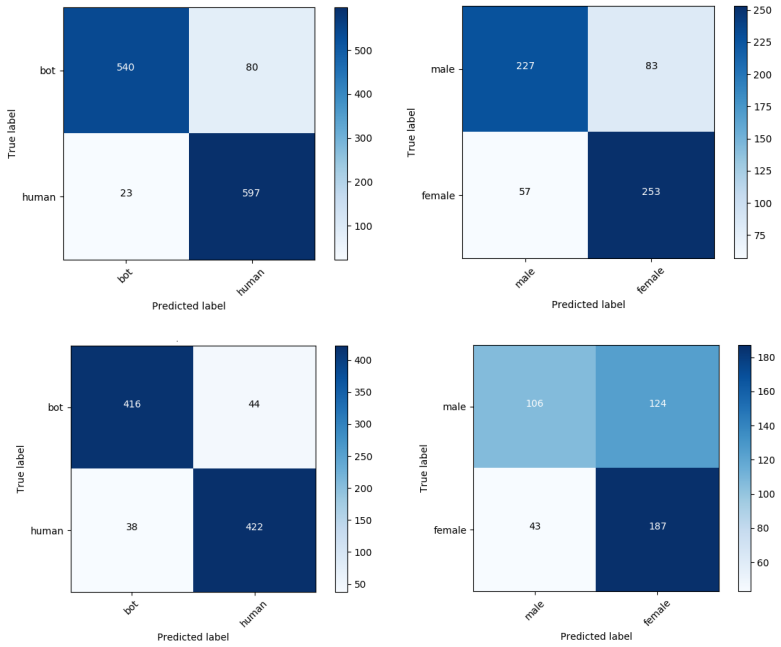


**Figure 1.** Confusion matrices for the English (top row) and Spanish (bottom row); human vs. bot classification (left column) and gender classification in case of human author (right column).

## References


1. Bessi, A., Ferrara, E.: Social bots distort the 2016 US presidential election online discussion. First Monday 21(11-7) (2016)
2. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (March 2016), `https://crscardellino.github.io/SBWCE/`
3. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
4. Daelemans, W., Kestemont, M., Manjavancas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Tschuggnall, M., Wiegmann, M., Zangerle, E.: Overview of PAN 2019: Author Profiling, Celebrity Profiling, Cross-domain Authorship Attribution and Style Change Detection. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D., Heinatz, G., Cappellato, L., Ferro, N. (eds.) Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Springer (Sep 2019)
5. Godin, F., Vandersmissen, B., De Neve, W., Van de Walle, R.: Multimedia lab @ ACL WNUT NER shared task: Named entity recognition for twitter microposts using distributed word representations. In: Proceedings of the Workshop on Noisy User-generated Text. pp. 146–153. Association for Computational Linguistics, Beijing, China (Jul 2015)
6. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks 18(5-6), 602–610 (2005)

7. Hosseinia, M., Mukherjee, A.: A parallel hierarchical attention network for style change detection: Notebook for PAN at CLEF 2018. In: Cappellato, L., Ferro, N., Nie, J., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv:1301.3781 (2013)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
11. Onose, C., Cercel, D.C., Trausan-Matu, S.: SC-UPB at the VarDial 2019 evaluation campaign: Moldavian vs. romanian cross-dialect topic identification. In: Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects. pp. 172–177 (2019)
12. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing. pp. 156–169. Springer International Publishing, Cham (2018)
13. Rangel, F., Rosso, P.: Overview of the 7th author profiling task at PAN 2019: Bots and gender profiling. In: Cappellato L., Ferro N., Müller H, Losada D. (ed.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org (2019)
14. 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 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014. pp. 1–30 (2014)
15. Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at PAN 2018: multimodal gender identification in twitter. Working Notes Papers of the CLEF (2018)
16. Yang, F., Mukherjee, A., Dragut, E.: Satirical news detection and analysis using attention mechanism and linguistic features. arXiv preprint arXiv:1709.01189 (2017)
17. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489 (2016)