Authorship Verification with neural networks via stylometric feature concatenation

Notebook for PAN at CLEF 2021

Antonio Menta [1] and Ana Garcia-Serrano [1]

[1] *E.T.S.I. Informática (UNED), Spain*

Abstract

In the authorship verification task (PAN at CLEF 2021) the main aim is to discriminate between pairs of texts written by the same author or by two different authors. Our work focuses on extracting two stylometric features, character-level n-grams and the use of punctuation marks in the texts. Subsequently, we train a neural network with each of them and finally combine them into a final neural network for the classifier decision making.

Keywords [1]

Stylometric features, neural networks, authorship verification

1. Introduction

In the last twenty years, the ease of publishing texts has increased thanks to different social networks, blogs, news and newspaper webpages. Similarly, the processes of digitalizing literary works started at the end of the last century by the institutions of different countries, has allowed digital access to thousands of textual and multimedia works. This increase in the number of digitized texts, together with the development of new statistical techniques, has served to increase interest in automatic analysis, instead of having to carry out a detailed reading as was done previously [1].

As a result of this interest, different tasks have been created in scientific events such as ImageCLEF [2], [3], [4] or the PAN series [5], [6], [7], [8]. PAN is organized with the aim of carrying out a forensic analysis of digital texts using different techniques including those based on stylometric analysis.

Authorship analysis is based on the study of the author's writing style. People leave a characteristic pattern of how they express their thoughts through written language. These patterns have a cognitive fingerprint that can be detected thanks to the study of stylistic features [9]. These can be divided into several categories. Lexical and character features, where a text is considered as a simple sequence of word symbols or characters, respectively. They may also include syntactic and semantic features that require a deeper linguistic analysis. Finally, the specific features of each domain of study and of the language used [10] may be looked into.

2. Problem statement

From the previous tasks we have focused on the Authorship Verification for the CLEF PAN 2021 conference [11]. The main aim is, given a couple of text documents, to determine whether they have been written by the same author or a different author. The documents have been extracted from the website www.fanfiction.net related to well-known movies, novels or TV series. They belong to the fanfiction genre, texts that borrow characters or stories from famous works (Star Wars, Harry Potter,

[1]CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
EMAIL: amenta1@alumno.uned.es (Antonio Menta); agarcia@lsi.uned.es (Ana Garcia-Serrano)
ORCID: 0000-0002-3172-2829 (Antonio Menta); 0000-0003-0975-7205 (Ana Garcia-Serrano)



CEUR Workshop Proceedings (CEUR-WS.org)

The Avengers), written by fans of the work and not by the authors themselves, with the idea of being shared with the community of followers.

This task is part of a larger three-year research project that started with PAN 2020 [5] and will be completed by next year's edition (PAN 2022). The training datasets proposed were the same as those of the first edition. The main difference between the two editions is that this year the authors and topics of the evaluation texts belong to a subset not found in the training data and, therefore, unknown to the trained models. Two different sizes of training data were used for solving the problem. A smaller one with 52,590 records focused on symbolic machine learning models and a larger one with 275,486 records created to allow the use of approaches based on recent advances in textual representation using neural networks. A test set of 20,000 records was used for evaluation.

3. Approach

Since interest in authorship detection tasks began, hundreds of different stylometric features have been tested. They can be organized in lexical, syntactic, semantic and structural categories. Within the lexical category some examples are the average word length, the richness of the vocabulary or the frequency of character n-grams used. Whereas in the syntactic category, one of the most historically used characteristics is the analysis of the frequency of syntactic tag bigrams [12]. The selection of which ones are more predictive depends on different factors such as the length of the texts, the domain studied, or the language, since the maturity of the software to extract syntactic or semantic features is not the same for each language.

Several studies have shown that the combination of features increases the results obtained in authorship identification [13]. This approach has been demonstrated at the PAN-CLEF 2020 edition in some papers [5]. In our approach we have used character n-grams, a feature widely used in the previous edition [14], [15] and with other different datasets [16]. It is worth noting that in the tests carried out using our neural network architecture, using only this feature, already gave results above the baseline proposed in the competition. In addition, to increase the generalization capabilities of our approach we added another feature of the texts, the use of punctuation marks by the authors. In the tests made alone with this feature it also showed predictive capabilities, although of lower value than the n-grams.

To solve the task, we have chosen to train a supervised classifier based on neural networks where the training data are the aforementioned stylometric features and the labels show whether the texts have been written by the same author or a different one. The size of the dataset makes the use of neural networks possible, as evidenced by the fact that last year the two best approaches were based on them, either using of a siamese network [17] or stylometric features as input [14]. Our work follows this second scheme by extending its use to both the smaller dataset and the larger version. In our opinion, the text length of the smaller dataset, 2,200 words-tokens on average [18] and the total number of records 52,590, allows us to obtain better results using neural networks than other models such as logistic regression or SVM.

To differentiate in each pair of texts those belonging to the same author and those belonging to different authors, we have used the difference between feature vectors as input for each neural network. Given two texts <X,Y>, we define

$$diff(X,Y) =< |x1 - y1|, \ldots., |xn - yn| >, \quad (1)$$

where n is the vector size of the features created [19]. In this manner, the classifier, instead of learning features specific to each author, learns the differences in the use of n-grams and punctuation symbols between authors.

In our opinion, the main point in the classifier is to increase the representativeness of each of the stylometric features used without minimizing the impact of the rest. In many approaches, once the features are extracted, they are all concatenated into a single vector. For example, in our approach the size of the n-grams vector feature has a dimension close to 45,000, while the vector related to the use of punctuation marks is about 32. The high dimensionality of the n-grams vector can minimize the representativeness of the punctuation-marks feature. Different approaches have been applied to solve

this problem in the literature. A first possibility is to use dimensionality reduction techniques such as PCA [20]. However, in our tests their use did not improve the results obtained. Therefore we have preferred to use an approach training a single neural network for each of the features, n-grams and punctuation marks, and subsequently concatenate their final vector representation into another neural network as final decision stage of a deep architecture [21], as shown in the Figure 1.

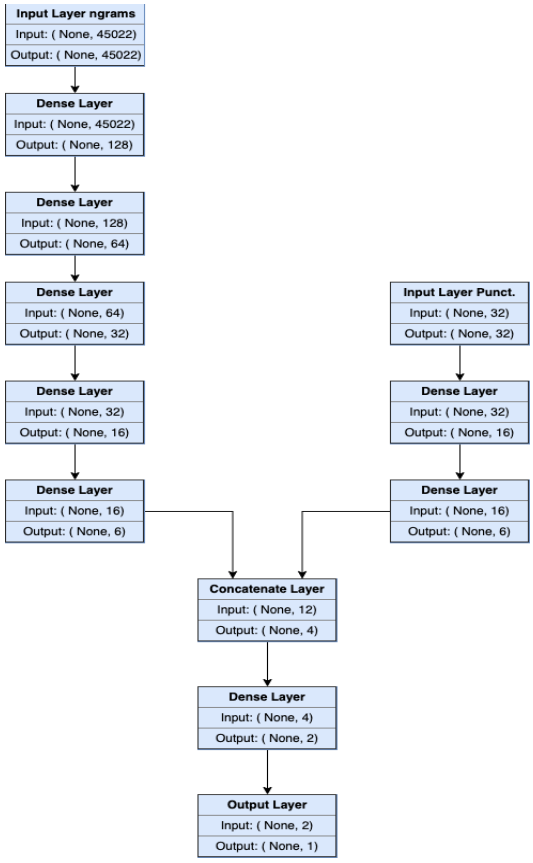

Figure 1: Neural Network Architecture

4. Experiments

A binary classifier based on our neural network architecture was trained for each of the datasets. The major difference between the two models has been the splitting of the data provided in training and validation. In the smaller dataset we have made a split using 5-fold training on four of them and validation on the remaining one. In the larger dataset we preferred to split into training (70%), validation (15%) and test (15%) due to the larger volume of data available, thus reducing the computational time required.

In both models, the same hyperparameters have been used to calculate the TF-IDF values of the selected stylometric features. The n-grams feature has been calculated in a range from 1 to 6, eliminating those elements of the vocabulary that appear in less than 5% of the documents. For the punctuation-marks feature, we have used those elements belonging to string module in Python 3.8.

All experiments were carried out on pytorch 1.6 using python 3.8. As regards the small dataset, we grouped the samples into batches of size 128, we used 50 epochs and a learning rate value of 0.001. With the larger dataset values were 368, 10 and 0.00005 respectively. Selu (Scaled Exponential Linear Unit) has been chosen as the activation function in each layer except in the last one in which a sigmoid function has been used for all experiments.

The final models were trained on a computer with a GeForce GTX 1070i GPU with 8GB of dedicated memory and 16GB RAM. The runtime of preprocessing steps was 2 hours and 10 minutes on the small datasets and 16 hours and 23 minutes on the larger one. Subsequently, the runtime of neural network models was 1 hours and 58 minutes, and 4 hours and 5 minutes respectively.

5. Evaluations

The trained models have been evaluated in TIRA evaluation system [22] from the following metrics: the area under the curve (AUC), c@1 (a variant of the conventional F1 score, which rewards systems that leave difficult problems unanswered), F1-score (the well-known metric without taking non-answers into account), Brier (complement of the Brier score, for evaluating the goodness of binary probabilistic classifiers), and finally, F0.5u, a recently proposed measure that puts more emphasis on the decision of cases by the same author.

The overall score used to generate the final classification is the average of the scores of all the aforementioned evaluation measures. The results obtained by the models are detailed in Table 1.

Table 1
Results

Dataset	AUC	c@1	F1	F0.5u	Brier	Overall
Small	0.9385	0.8662	0.8620	0.8787	0.8762	0.8843
Large	0.9635	0.9024	0.8990	0.9186	0.9155	0.9198

6. Conclusions

In this paper we present an approach based on stylometric feature extraction combined with a neural network architecture where each feature has its own neural network to subsequently concatenate the outputs into a final neural network. We have shown that this type of architecture can obtain good results in the authorship verification task for PAN 2021 on both datasets. Furthermore, our approach has shown that competitive results are possible with only two stylometric features.

As a future line of research, we plan to increase the number and type of stylometric features used. Thus, we will obtain a combination of neural networks where the importance of each stylometric feature can be modified by varying the length of its output vector. Likewise, we plan to carry out a better tuning of the neural network hyperparameters other than manually as we do now.

7. References

[1] P. Juola, "Authorship attribution," Found. Trends Inf. Retr., vol. 1, no. 3, pp. 233–334, 2006.

[2] J. L. Martínez, J. Villena, A. Garcia-Serrano, and J. González, "Combining Textual and Visual Features for Image Retrieval," LNCS, vol. V4022, pp. 680–691, 2006.

[3] A. Garcia-Serrano, X. Benavent, R. Granados, and JM. Goñi, "Some Results Using Different Approaches to Merge Visual and Text-Based Features in CLEF'08 Photo Collection," LNCS, vol. 5706, pp. 568–571, 2009.

[4] J. Benavent, X. Benavent, E. de Ves, R. Granados, and A. Garcia-Serrano, "Experiences at ImageCLEF 2010 using CBIR and TBIR Mixing Information Approaches," CLEF 2010. CEUR WS 1176, 2010.

[5] M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, and B. Stein, "Overview of the Cross-Domain Authorship Verification Task at PAN 2020" CLEF CEUR WS V 2696, pp. 22–25, 2020.

[6] W. Daelemans, M. Kestemont, E. Manjavacas, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Tschuggnall, M. Wiegmann and E. Zangerle. "Overview of PAN 2019: Bots and Gender Profiling, Celebrity Profiling, Cross-Domain Authorship Attribution and Style Change Detection". Springer LNCS 11696. 2019.
[7] E. Stamatatos, F. Rangel, M. Tschuggnall, B. Stein, M. Kestemont, P. Rosso and M. Potthast "Overview of PAN 2018: Author identification, author profiling, and author obfuscation ". Springer LNCS, V 11018. 2018
[8] J. Bevendorff, B.Chulvi, G,Sarracén, M.Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M.Potthast, F.Rangel, P.Rosso, E.Stamatatos, B. Stein, M. Wiegmann, M. Wolska, and Eva Zangerle, "Overview of PAN 2021: Authorship Verification,Profiling Hate Speech Spreaders on Twitter,and Style Change Detection," in 12th International Conference of the CLEF Association (CLEF 2021), 2021.
[9] N. Pokhriyal, K. Tayal, I. Nwogu, and V. Govindaraju, "Cognitive-Biometric Recognition from Language Usage: A Feasibility Study," IEEE Trans. Inf. Forensics Secur., vol. 12, no. 1, pp. 134–143, 2017.
[10] E. Stamatatos, "A survey of modern authorship attribution methods," J. Am. Soc. Inf. Sci. Technol., vol. 60, no. 3, pp. 538–556, Mar. 2009.
[11] M. Kestemont, I. Markov, E. Stamatatos, E. Manjavacas, J. Bevendorff, M. Potthast and Benno Stein, "Overview of the Authorship Verification Task at PAN 2021," in CLEF 2021 Labs and Workshops, Notebook Papers, 2021.
[12] G. Hirst and O. Feiguina, "Bigrams of syntactic labels for authorship discrimination of short texts," Lit. Linguist. Comput., vol. 22, no. 4, pp. 405–417, 2007.
[13] O. Halvani, C. Winter, and A. Pflug, "Authorship verification for different languages, genres and topics," DFRWS 2016 EU - Proc. 3rd Annu. DFRWS Eur., vol. 16, pp. S33–S43, 2016.
[14] J. Weerasinghe and R. Greenstadt, "Feature Vector Difference based Neural Network and Logistic Regression Models for Authorship Verification Notebook for PAN at CLEF 2020," no. September, pp. 22–25, 2020.
[15] E. Araujo-Pino, H. Gómez-Adorno, and G. Fuentes-Pineda, "Siamese Network applied to Authorship Verification Notebook for PAN at CLEF 2020."
[16] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group Web forum messages," IEEE Intell. Syst., no. October, pp. 67–75, 2005.
[17] B. Boenninghoff, J. Rupp, R. M. Nickel, and D. Kolossa, "Deep bayes factor scoring for authorship verification," arXiv, no. September, pp. 22–25, 2020.
[18] C. Ikae, "UniNE at PAN-CLEF 2020: Author Verification Notebook for PAN at CLEF 2020," 2020.
[19] M. Koppel and Y. Winter, "Determining if two documents are written by the same author," J. Am. Soc. Inf. Sci. Technol., vol. 65, no. 1, pp. 178–187, 2014.
[20] A. Jamak, A. Savatić, and M. Can, "Principal component analysis for authorship attribution," Proc. 11th Int. Symp. Oper. Res. Slov. SOR 2011, no. September 2012, pp. 189–196, 2011.
[21] E. Akcapinar Sezer, H. Sever, and P. Canbay, "Deep Combination of Stylometry Features in Forensic Authorship Analysis," Int. J. Inf. Secur. Sci., vol. 9, no. 3, pp. 154–163, 2020.
[22] M. Potthast, T. Gollub, M. Wiegmann, and B. Stein, "TIRA Integrated Research Architecture," in Information Retrieval Evaluation in a Changing World, Springer, 2019.