Character-based Convolutional Neural Network for Style Change Detection

Notebook for PAN at CLEF 2018

Nils Schaetti

University of Neuchâtel
rue Emile Argand 11
2000 Neuchâtel, Switzerland
nils.schaetti@unine.ch

Abstract. This paper describes and evaluates a model for style change detection using character-based Convolutional Neural Networks (CNN). We applied this model to the style change detection task of the PAN18 challenge and show that its architecture allows this model to be applied to any language. This CNN based on a character-embedding layer, 25 filters and a temporal max-pooling layer reaches a classification accuracy of 62.13%. The evaluation is based on a collections of text gathered from various sites of the StackExchange network, covering different topics. Regarding accuracy, our model arrives last out of the five participants but second in terms of runtime. (PAN STYLE CHANGE DETECTION task at CLEF 2018).

1 Introduction

Today, the use of the work of an author without its authorisation, known as textural plagiarism, is a major problem in fields such as education and research. The field of automatic plagiarism detection raises new questions : how to find if a text has been written by one or more authors? The increasing access to the Word Wide Web make millions of textual resources easily accessible and providing and enormous amount of sources for potential plagiarism. Therefore, technology and methods to automatically detect plagiarism has received increasing intention in the software industry and in the academia.

There are two kinds of tasks in plagiarism analysis, external plagiarism detection and *intrinsic plagiarism detection*. The first refers to the use of a given reference corpus to identify pairs of very similar passages in a suspicious document. In the second, no reference corpus is given and we must rely on the detection of irregularities, inconsistencies or anomalies within a document. The second is more ambitious since no reference corpus is given, but the first one is the target of most studies.

To face this challenge, the main line of research known as 'stylometry' attempted to quantify the writing style using a variety of measures, representing kind of stylistic information, such as lexical features (word frequencies, word n-grams) or syntactic feature features (part-of-speech) and some studies have demonstrated the effectiveness of character n-grams.

As this year PAN18 challenge propose a style change detection task, we decided to evaluate a character-based CNN (Deep-Learning) model on this task. This paper is organised as follow. Section 2 introduces the dataset used for training, validation and testing, as well as the measures and methodology used to evaluate our approach. Section 3 explains the proposed character-based Convolutional Neural Network (CNN) model used to classify the texts. In section 4, we evaluate the strategy we created and compare results on the test collections. In the last section, we draw conclusions on the main findings and possible future improvements.

2 Corpus and Methodology

To compare different experimental results on the style change detection task with different models, we need a common ground composed of the same datasets and evaluation measures. In order to create this common ground, and to allow the large study in the domain of intrinsic plagiarism detection, the PAN CLEF evaluation campaign was launched [5]. Multiple research groups with different backgrounds from around the world have proposed a detection algorithm to be evaluated in the PAN CLEF 2018 campaign [6, 1] with the same methodology [3].

All teams have used the *TIRA* platform to evaluate their strategy. This platform can be used to automatically deploy and evaluate a software [2]. The algorithms are evaluated on a common test dataset and with the same measures, but also on the base of the time need to produce the response. The access to this test dataset is restricted so that there is no data leakage to the participants during a software run. For the PAN CLEF 2018 evaluation campaign, a collection of texts was created. Based on this collection, the problem to address was to predict if the text is the work of one or more author [4].

The training and validation data were collected from various site of the StackExchange network. The texts come from the same language. For each text, there is a two-class label we can predict which can take the value *True* (stylistic change(s) in the text) or *False* (no stylistic change(s)). The test sets are also texts collected from the StackExchange network and the task is therefore to predict the *changes* label for each text in the test data.

The training collection is composed of 2'980 text, 1'490 for each class. To allow our classification model to reach higher accuracy, we extended the training collections by switching the parts written by different authors to create new examples. For example, if an example has three authors, create six (3!) new examples combining the different parts. As we produce more examples with changes that without, the resulting dataset is

Corpus	Document	Changes	No changes
Training	2980	1490	1490
Validation	1492	746	746
Extended training	18913	13007	5906

Table 1: Training, validation and extended training collections

biased towards document containing changes. We left the creation of an extended and not biased dataset for future research as it could improve the performance. This result in a final training set of 18'913 texts, 13'007 for the class of multi-authored document and 5'906 for the class single authored documents. An overview of these collections is depicted in table 1. The number of documents from each collection is given under the label "Documents" and the total number of document per class in the collection are indicated respectively under the labels "Changes" and "No changes". The training and validation data set are well balanced as for each collection, there is the same number of documents for each class.

A similar test set will be used to compare the participants' strategies of the PAN CLEF 2018 campaign, and we don't have information about its size due to the *TIRA* system. The response for the changes is a binary choice (*false* / *true*). The overall performance of the system is the classification accuracy. The accuracy is the number of documents where the class is correctly predicted divided by the number of documents in the collection.

3 Character-based Convolutional Neural Network (CNN)

In machine learning, a *Convolutional Neural Network* (or CNN) is a kind of feed-forward artificial neural network, in which the patterns of connection between the neurons are inspired from the visual cortex.

In our system, we applied a character based CNN to each text in a collection. A text is fed into the model as an array of character with a fixed size of 12'000. If the document is shorter than 12'000, the additional space is felt with zeros as each character is represented by an index. For each document, we passed it to lower cases and transformed it into a list of character. Each character are transformed to indexes with a vocabulary V

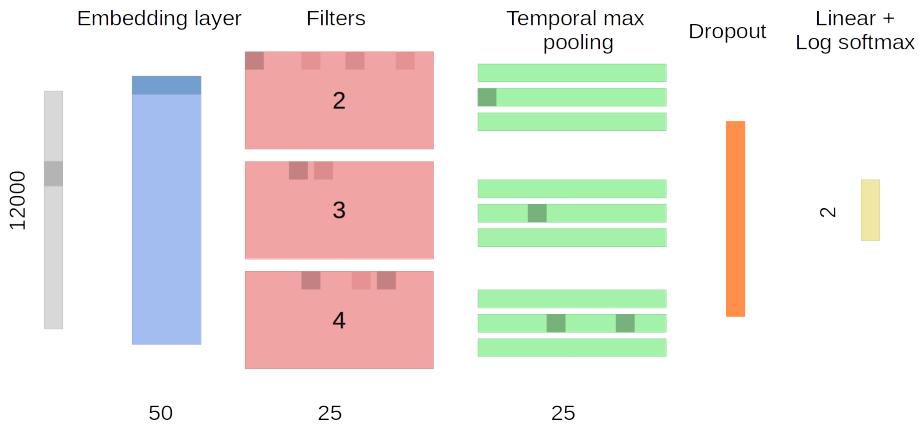

Fig. 1: Structure of the character based *Convolutional Neural Network* with the following layers : embedding layer (dim=50), three convolutional layers (kernel size 2, 3 and 4), three max pooling layers of size 700 and stride 350, a drop-out layer and a final linear layer of size 2 with log softmax outputs.

Corpus	10-Fold CV	**Random**
Validation	0.6340	0.5000
Test	0.6213	0.5000

Table 2: Evaluation for the three collections

constructed during the training phase.

The first layer is a *embedding layer* with a size equal to the vocabulary size $|V|$ and a dimension of 50 for each character. This layer has two purposes, first to reduce the dimensionality of the inputs to 50, compared to $|V|$ for one-hot encoded vectors, and secondly, to encode similarities between character into a multi-dimensional space where two character appearing in similar context are near each others. The second layer is composed of three different *convolutional layers* with kernel sizes of 2, 3 an 4. Theses layers encode patterns of 2, 3 or 4 consecutive character 2-grams and each layer has 25 filters and 75 patterns can thus be represented. During the training phase, our model will then find the 75 most effective patterns of character to encode the irregularities.

The third layer is composed of three *max pooling layers* with size 700 and stride 350, one for each preceding convolutional layers. These layers encode the pattern matching for each part of size 700 of the texts. We pass the output through a *ReLU* non-linearity and a dropout layer. The final layer is a *linear layer* of size 2, one output per class. The training phase consists of using the whole extended training dataset. We used the *stochastic gradient descent algorithm* to train our model with a *learning rate* of 0.0005, a *momentum* of 0.9 and *cross-entropy* as *loss function*. At the end of the training phase, we choose the CNN which obtained the best accuracy on the validation dataset.

To implement our model, we used TorchLanguage[1], a package based on pyTorch designed for Natural Language Processing.

4 Evaluation

To evaluate our two models we tested their accuracy on the extended training corpus. The table 2 shows the results of accuracy on the validation set. Our model attained an accuracy of 63.40% compared to 50% for a random classifier. The table 2 shows also the results on the test collection obtained on the *TIRA* platform. Our model attained an accuracy of 62.13% compared to 50% for a random classifier, not far from the results previously obtained on the training corpus.

The table 3 shows the ranking evaluation for the style change detection task. The best model achieves an accuracy of 89.30% and our model arrives last with 62.10%, but second in terms of runtime with only 3 minutes and 36 seconds, compared to more that one and a half hour for the best model.

[1] https://github.com/nschaetti/TorchLanguage

User	Accuracy	Runtime
zlatkova18	0.893	01:35:25
hosseinia18	0.825	10:12:28
ogaltsov18	0.803	00:05:15
khan18	0.643	00:01:10
schaetti18	0.621	00:03:36

Table 3: PAN18 style change detection, evaluation results

5 Conclusion

This paper evaluated a Deep-Learning model for style change detection based on documents gathered from the StackExchange network. Based on the hypothesis that textual documents posted on websites can be used to detect stylistic changes, we introduced a CNN classifier for document classification that can predict this characteristics but with a very limited capacity probably due to a very small data set. The character 2-grams based CNN shows a high over-fitting even with an extended data set and with the use of a dropout layer.

The CNN model achieves its best performance on the validation dataset with 60.3% accuracy after 30 iterations. On the test dataset, the CNN model achieves 62.13% accuracy. The biggest challenge of this task for the kind of model we evaluated is the lack of data to achieve a good approximation of the network's parameters. Regarding accuracy, our model arrives last out of the five participants but second in terms of runtime.

References

1. Kestemont, M., Tschugnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)
2. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
3. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Jones, G., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, Septembre 11-14, 2017, Proceedings. Springer, Berlin Heidelberg New York (Sep 2017)

4. Rangel, F., Rosso, P., Potthast, M., Stein, B.: In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Labs Working Notes
5. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. Working Notes Papers of the CLEF (2016)
6. Stamatatos, E., Rangel, F., Tschuggnall, M., Kestemont, M., Rosso, P., Stein, B., Potthast, M.: Overview of PAN-2018: Author Identification, Author Profiling, and Author Obfuscation. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J., Soulier, L., Sanjuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 9th International Conference of the CLEF Initiative (CLEF 18). Springer, Berlin Heidelberg New York (Sep 2018)