BertAA: BERT fine-tuning for Authorship Attribution

Maël Fabien[1,2], **Esaú Villatoro-Tello**[1,3], **Petr Motlicek**[1], and **Shantipriya Parida**[1]

[1]Idiap Research Institute, Martigny, Switzerland.
{firstname.lastname}@idiap.ch
[2]Ecole Polytechnique Fédérale de Lausanne, Switzerland.
mael.fabien@epfl.ch
[3]Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Mexico City, Mexico.
evillatoro@correo.cua.uam.mx

Abstract

Identifying the author of a given text can be useful in historical literature, plagiarism detection, or police investigations. Authorship Attribution (AA) has been well studied and mostly relies on a large feature engineering work. More recently, deep learning-based approaches have been explored for Authorship Attribution (AA). In this paper, we introduce BertAA, a fine-tuning of a pre-trained BERT language model with an additional dense layer and a softmax activation to perform authorship classification. This approach reaches competitive performances on Enron Email, Blog Authorship, and IMDb (and IMDb62) datasets, up to 5.3% (relative) above current state-of-the-art approaches. We performed an exhaustive analysis allowing to identify the strengths and weaknesses of the proposed method. In addition, we evaluate the impact of including additional features (e.g. stylometric and hybrid features) in an ensemble approach, improving the macro-averaged F1-Score by 2.7% (relative) on average.

1 Introduction

Authorship Analysis is the field of Natural Language Processing that studies the characteristics of a text and extracts information on its author. It is made of 3 sub-tasks, which include author profiling, i.e. detecting sociolinguistic attributes such as gender or age, authorship verification which identifies the degree of similarity of texts, and authorship attribution (El et al.). Authorship Attribution (AA) is the process of attributing a text to the correct author among of closed set of potential writers. AA is widely used in plagiarism detection or attribution of historical literature (Li). This classification task is also well known in forensic investigations (Yang and Chow, 2014).

AA has been studied on short texts (Aborisade and Anwar, 2018), such as Tweets as well as longer texts, such as judgments of a few thousand words on average (Sari et al., 2018). The main challenge in AA is the extraction of relevant features characterising the author's identity. Majority of approaches proposed in the past relied on a large amount of feature engineering, in order to reflect both the content and the style of the author (Madigan et al., 2005; Aborisade and Anwar, 2018; Seroussi et al., 2014; Bozkurt et al., 2007).

In this paper, we propose a method, BertAA, that relies on the fine-tuning of a pre-trained BERT language model, to which we add a dense layer and a softmax activation for authorship classification, trained for a few epochs. This is one of the very first attempts to analyze the performances of pre-trained language model fine-tuning for in-domain AA, especially for a large number of authors (up to 100). As most Deep-Learning methods for AA, BertAA does not require text preprocessing nor feature engineering. Our method offers state-of-the-art (SOTA) performances on well-known corpora, with a relative accuracy improvement of up to 5.3%. We also illustrate the strengths and weaknesses of such a system. We also show that building an ensemble architecture, which also incorporates stylometric and hybrid features tends to improve the macro-averaged F1-score. Finally, we set a benchmark for the full IMDb corpus (Seroussi et al., 2014) for 5, 10, 25, 50, 75, and 100 authors, which, to the best of our knowledge, has never been studied in its full format for AA.

The next section discusses the relevant approaches developed in the literature. Section 3 presents the corpora used as well as a brief exploration of each of the sources. Section 4 details the architectures of BertAA, while Section 5 describes the results we obtained, and Section 6 discusses the strengths and weaknesses of our method, as well as future work directions. Finally, Section 7 depicts our conclusions.

2 Related work

Traditionally, AA largely relies on the process of extracting features related to content or style of an author (Stamatatos, 2009). More recently, some approaches propose to use deep learning methods for AA tasks, whether relying on a previous feature extraction step or not. The following sections briefly describe these various methods.

2.1 Traditional methods

Term Frequency - Inverse Document Frequency (TF-IDF) is used in AA at the word or the word or character N-gram level. It captures the words, the stems, or the combinations of words or letters that an author uses. Some recent works combine the votes of several classifiers on several levels of N-grams (Muttenthaler et al.). Such methods are referred to in the literature as being content-related classifiers (Sari et al., 2018).

In addition, stylometric features reflect the style of the author (Sari et al., 2018). The main hypothesis behind this feature extraction is that each author has its own writing style (e.g use of punctuation, average word length, sentence length, number of upper cases...). Features reflecting the style are used as an input for a LR usually, as seen in (Madigan et al., 2005; Aborisade and Anwar, 2018; Madigan et al.). An optional step of text pre-processing is often added (Allison and Guthrie), and more specifically through stop-words removal and stemming. Sari et al. (2018) reached an accuracy of 95.9% on the IMDb62 dataset (Seroussi et al., 2014), 1.1% (absolute) above a character N-gram classifier, by including stylometric features in a classifier. Soler-Company and Wanner (2017) also showed that including syntactic and discourse features can help achieve SOTA performances in author and gender identification.

To combine the numerous sources of input features, AA is also performed using ensemble learners, made for example of several SVM classifiers (Bacciu et al., 2020). Each classifier is trained on certain features related to distinct concepts, such as style, content, author profiling, etc.

2.2 Deep Learning based methods

While overcoming the burden of feature engineering, deep learning-based methods have reached SOTA results, whether through the use of Long Short-Term Memory (LSTM) (Qian et al.) at both the sentence and article-level or using multi-headed Recurrent Neural Network (RNN) (Bagnall, 2016) for on short multi-lingual texts. Convolutional Neural Networks (CNN) have also been widely explored for AA and can extract information from raw signals in speech processing or computer vision.

Ruder et al. (2016) explored CNNs at the word and character level for AA and found that CNNs at the character level tend to outperform other simple approaches based on SVMs for example, while CNNs at the N-gram level have been shown to perform competitively (Shrestha et al., 2017). Zhang et al. (2018) proposed a Syntax-augmented CNN model which outperforms other approaches on the Blog authorship and the IMDb62 datasets.

Siamese networks are well known in computer vision, e.g. for facial recognition tasks (Wu et al., 2017). Saedi and Dras (2019), used Convolutional Siamese Networks to perform AA. They compared their approach with a BERT fine-tuning over 3 epochs and showed that Siamese Networks are more robust over large-scale AA tasks ($N > 50$). This type of approach has the advantage of being able to evaluate the similarity between texts, as shown in (Qian et al.).

In 2020, Barlas and Stamatatos (2020) leveraged pre-trained language models (BERT, ELMo, ULMFiT, GPT-2) for the specific case of cross-topic and cross-domain AA on the CMCC dataset (Goldstein-Stewart et al., 2009), on a subset of 21 authors. The authors used a multi-headed classifier with a demultiplexer. In an N-authors classification task (N typically < 100), N classifiers would be trained, each of them seeing predominantly data from one author. In prediction, the text to classify is passed through all classifiers, and after normalization, the scores are compared. This work shows that BERT seems to work best on large vocabularies, and outperforms multi-headed RNNs.

Contrary to previous work, in this paper we perform an exhaustive analysis on the performance of pre-trained language models, we identify the advantages and limitations on three well-known benchmark datasets. In addition, we evaluate the impact of incorporating stylometric and hybrid features through ensemble techniques.

3 Authorship Attribution Corpora

Several corpora have been studied for the task of AA. In this section, we briefly describe each corpus we used as well as its key features. The AA task

we performed focuses on identifying the author of a text among a list of the top N authors for whom we collected the largest number of texts.

3.1 Enron Email corpus

Enron Email corpus has been widely studied over the previous decade since the bankruptcy of Enron. 517'401 emails from around 160 employees were made public, and data preparation for email classification was then done by Klimt and Yang (2004). The emails mainly contain conversations of managers at Enron, and given the fraudulent nature of the emails, it is commonly used as a study case for criminal network investigations (Aven, 2015).

Emails were collected from the "Sent" folder of each of the 160 employees. Since around 13% of the emails contained the name of the sender, as a signature or side information in a forwarded message, we dropped these observations. We also removed all messages of less than 10 tokens to apply the same processing as Ruder et al. (2016). Our end corpus contained 130'000 emails. Emails are on average 150 tokens long, and the median length is 61 tokens.

Enron Email corpus has already been studied for several Authorship Analysis tasks, including Authorship Verification (Halvani et al., 2020; Brocardo et al., 2013), as well as for AA tasks (Neumann and Schnurrenberger; Li; Allison and Guthrie). Gender identification and sentiment analysis were also studied by Clough et al. (2011).

3.2 IMDb Authorship Attribution Corpus

The IMDb Authorship Attribution corpus was introduced by Seroussi et al. (2014). 271'000 movie reviews were produced by 22'116 distinct authors, with an average of 12.3 texts per author. Texts are on average 121 tokens long. No preprocessing or filtering was applied to the corpus.

Most of the works that we have found referred to the IMDb62 dataset, a truncated version of the IMDb Authorship Attribution Corpus with 62 authors and 1'000 texts per author. We chose to benchmark our solution on the IMDb62 against other approaches, but also to evaluate the performance of our model on the full version of the corpus since it contains a class imbalance (closer to a real-life scenario) and has more data, with an average of 3'900 texts per author for the top 5 authors. The full version of the corpus has, to the best of our knowledge, never been studied for AA for a various number of authors. Hence, our approach sets a benchmark.

3.3 Blog Authorship Attribution Corpus

The Blog Authorship Attribution corpus is a corpus of blog articles from 2004 and before, collected from blogger.com. It was introduced by Schler et al. as part of a study on the effects of age and gender on blogging. More than 680'000 posts are available, from more than 19'000 authors. An average of 35 posts was collected per author. No preprocessing or filtering was applied to the corpus. Although it might seem surprising, it is worth mentioning that this dataset is the one containing the shortest texts on average (79 tokens for the top 5 authors, vs 190 for Enron). Many of the blog posts collected were replies to existing blog posts or short articles.

For our experiments, we considered the top 5, 10, 25, 50, 75, and 100 authors with the largest number of texts. Table 1 presents the summary statistics of the length and number of documents per author, in the various configurations considered, for each dataset. As a summary, Enron has rather long texts, a large number of texts per author with a large associated standard deviation. IMDb reviews are shorter and the number of texts per author is lower than for Enron. Finally, for the Blog dataset, the texts are short, and the number of texts per author is smaller than for Enron, with fewer variability than for IMDb.

Dataset	N	Avg. Num. Tokens	Avg. Nb. Texts
Enron	5	190 (± 375)	11205 (± 2324)
	10	201 (± 419)	8745 (± 3052)
	25	185 (± 375)	5626 (± 3230)
	50	183 (± 361)	3685 (± 3014)
	75	194 (± 386)	2774 (± 2779)
	100	208 (± 717)	2259 (± 2567)
IMDb	5	106 (± 184)	3900 (± 2197)
	10	127 (± 185)	2817 (± 1895)
	25	110 (± 167)	1873 (± 1434)
	50	104 (± 152)	1324 (± 1155)
	75	102 (± 158)	1080 (± 1005)
	100	102 (± 157)	932 (± 907)
IMDb62	62	341 (± 223)	1000 (± 0)
Blog	5	79 (± 191)	2659 (± 780)
	10	91 (± 184)	2350 (± 639)
	25	99 (± 174)	1832 (± 599)
	50	98 (± 167)	1466 (± 562)
	75	120 (± 209)	1270 (± 538)
	100	126 (± 228)	1122 (± 533)

Table 1: Descriptive statistics for the 4 datasets. N: number of authors, Avg. Num. Tokens: average number of tokens per text, Avg. Nb. Texts: average number of texts. Standard deviation in parenthesis.

4 BertAA : BERT-based Authorship Attribution

Content-related features in AA take into account the topics and the semantics of the text. Recent works on language representation models have however shown that transformers such as BERT (Devlin et al., 2019) reach SOTA performances for various tasks, hence improving GLUE score as well as several other metrics. It has been extensively used for text classification tasks (Sun et al., 2020), and BERT is known to be well-performing at extracting semantic and syntactic information.

To the best of our knowledge, no systematic review of the performance of fine-tuned pre-trained language models for AA has been reported yet, and such classifier has never been combined with a stylometric and hybrid features in an ensemble model. Hereby, we introduce BertAA, a fine-tuning of BERT with a dense layer and a softmax activation, trained for a few epochs for AA. The output dimension of the dense layer corresponds to the number of authors in the corpus.

BERT is made of 12 Transformer blocks and 12 self-attention heads. The input size, i.e. the maximum length of tokens is 512, and the hidden layer representation dimension is 768 (Vaswani et al., 2017). As described by Sun et al. (2020), to use BERT as a classifier, a simple dense layer with softmax activation is added on top of the final hidden state h of the first token [CLS], through a weight matrix W, and we predict the probability of label c the following way:

$$p(c \mid \mathbf{h}) = \text{softmax}(W\mathbf{h}). \quad (1)$$

Then, all weights, including BERT's ones and W, are adapted, in order to maximize the log-probability of the correct label. The training is done using a Cross-Entropy loss function. We used a pre-trained BERT available from the Transformer library (Wolf et al., 2020), trained on large corpora. The fine-tuning of BERT for the AA task was done on a Tesla P100-PCIE-16GB.

Additionally, we incorporate stylometric and hybrid features to BertAA, in 2 models called BertAA + Style and BertAA + Style + Hybrid through a LR. Thus, our system is able to account for content, stylometric, and hybrid features. The architecture of BertAA + Style + Hybrid is presented in Figure 1.

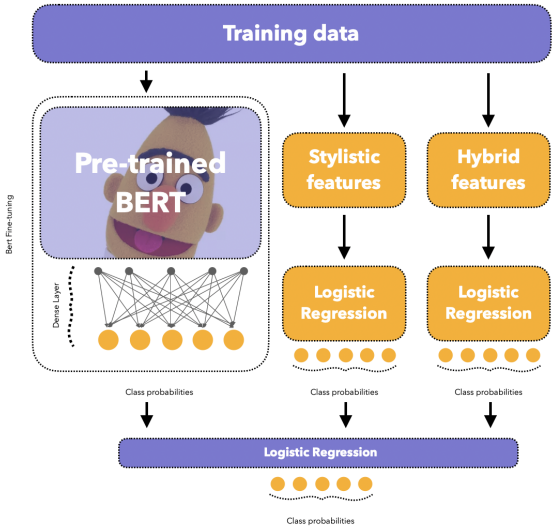

Figure 1: BertAA + Style + Hybrid architecture.

The stylometric classifier first extracts the lexical stylometric features as proposed by Sari et al. (2018). The features extracted are the length of text, the number of words, the average length of words, the number of short words, the proportion of digits and capital letters, individual letters and digits frequencies, hapax-legomena, a measure of text richness, and the frequency of 12 punctuation marks. A LR is trained on these features. The hybrid features we extract are the frequencies of the 100 most frequent character-level bi-grams and tri-grams. Classification is then done using a LR. Finally, the output probabilities of Bert classifier, the stylometric, and the hybrid ones are concatenated and classified using an additional LR.

5 Results

Parameters we chose for our architectures are presented in Table 2.

Model	Parameter	Value
Hybrid feat.	Char. N-grams	(2,3)
LR	Penalty	l2
	Tolerance	0.0001
	C	1.0
	Max Iterations	100
	Intercept	True
BERT	Config	bert-base-cased
	Epochs	1 to 5
	Input token length	512

Table 2: Parameters of the experiments.

We ran the experiments on 5, 10, 25, 50, 75, and 100 authors for the full IMDb, the Blog, and Enron datasets presented above. Our model was trained

on 5 epochs for each experiment unless specified otherwise. The results are presented in Table 3. We picked the top N authors with the largest amount of texts, for each of the datasets, and kept 20% of test data using a stratified approach, meaning that the proportions of each class are kept equal in the training and testing set. We report the results of BertAA, BertAA + Style and BertAA + Style + Hybrid. We compare our approach with a word-level TF-IDF - LR model with stemming and stop-words removal. We also add as a benchmark the performance of a LR trained only on stylometric features, and an additional LR trained on the character-level N-gram hybrid features.

BertAA outperforms the TF-IDF and LR benchmark on all experiments, with an average relative accuracy gain of 14.3%. It reaches a competitive performance on 5 authors on the Enron dataset, since only 2 samples were not classified correctly out of 4104, hence leading to an accuracy of 99.95%.

Comparing results on Enron to other approaches in literature is not trivial since it largely depends on the data preparation that was done. We decided to remove short emails, and remove utterances containing the name of the sender (as a signature for example), but not all papers involving Enron data for AA precisely describe their data preparation. Furthermore, we found no results in the literature on IMDb full-corpus for the top N authors. Hence, our results set a benchmark on the full IMDb, on average 8.2% above a word-level TF-IDF. Next, we compare our results with current SOTA on the IMDb62 and the Blog Authorship datasets.

5.1 How does the performance compare to SOTA?

In Table 4, we report the accuracy of our best systems (no additional features, 5 epochs) on the Blog Authorship corpus against the performances of several CNN-based architectures, including the character-level CNN presented in (Ruder et al., 2016) and current SOTA Syntax-enriched CNN(Zhang et al., 2018). We report results over 10 and 50 authors. For 10 authors, the accuracy of our best BertAA system (no additional features, 5 training epochs) reaches 65.4% which is, to the best of our knowledge, the current SOTA on the Blog Authorship Corpus, and represents a relative improvement of 2% over the Syntax CNN. When the number of authors increases, our system displays an accuracy of 59.7%, which represents a relative improvement of 5.3% accuracy compared to the previous SOTA. The main characteristics of the Blog Corpus are that texts are rather short on average (respectively 91 and 98 tokens on average for 10 and 50 authors), while the number of texts per authors remains quite high on average, with a rather small standard deviation, suggesting that BertAA is well suited for datasets with short sentences, and a large but balanced number of texts per author.

Approach	10	50
Impostors (Koppel and Winter, 2014)	35.4	22.6
SCAP (Frantzeskou et al., 2006)	48.6	41.6
LDAH-S (El et al.)	52.5	18.3
CNN (Ruder et al., 2016)	61.2	49.4
Continuous N-gram (Sari et al., 2017)	61.3	52.8
N-gram CNN (Zhang et al., 2018)	63.7	53.1
Syntax CNN (Zhang et al., 2018)	64.1	56.7
BertAA	**65.4**	**59.7**

Table 4: Accuracy on Blog Authorship

5.2 Are external features useful?

In order to assess the impact of external features, we compute the accuracy per author on the Blog dataset for 10 authors. We compare the per-author accuracy of a word-level TF-IDF + LR classifier and BertAA, to identify whether TF-IDF outperforms our system on some classes in Figure 2.

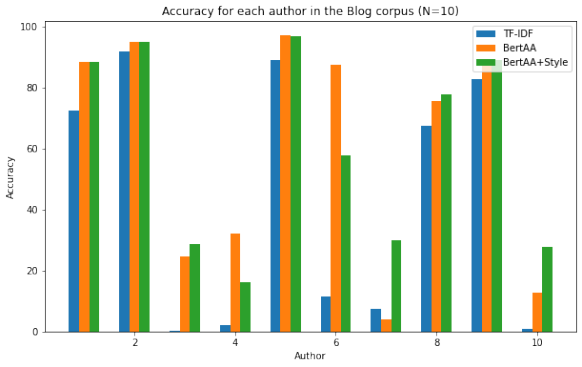

Figure 2: Accuracy per author for TF-IDF and BertAA (+Style) on the Blog Dataset (N=10)

On most authors, BertAA slightly outperforms TF-IDF, although both methods reach good accuracies. However, BertAA brings additional value, especially where TF-IDF performs poorly, e.g. on Author 3 in the figure. In some specific cases, such as for author "7" on the figure, TF-IDF achieves a better performance than BertAA. In such a case, adding the stylometric features improves the per-

Dataset	N-Authors	Baslines			Proposed Method		
		Stylo.	Char N-gram	TF-IDF	BertAA	+ Style	+ Style + Hybrid
Enron	5	75.0	84.4	98.0	**99.95**	**99.95**	**99.95**
	10	54.9	70.5	96.4	**99.1**	**99.1**	**99.1**
	25	35.6	53.2	92.7	**98.7**	**98.7**	**98.7**
	50	20.4	44.8	90.8	98.1	**98.2**	**98.2**
	75	17.3	40.6	90.1	**97.6**	97.5	97.5
	100	15.8	36.9	88.3	97.0	97.0	**97.1**
IMDb	5	65.8	92.1	98.1	**99.6**	**99.6**	**99.6**
	10	44.6	79.2	93.9	98.1	**98.2**	**98.2**
	25	25.5	55.8	84.1	**93.2**	92.9	92.9
	50	17.4	44.2	82.1	**90.7**	90.6	90.6
	75	14.7	37.6	79.2	**88.3**	87.8	87.8
	100	11.8	33.6	76.6	**86.1**	85.3	85.4
Blog	5	34.7	40.0	45.7	**61.3**	59.7	59.8
	10	18.9	31.9	45.0	**65.4**	62.4	62.4
	25	9.9	23.4	42.0	**65.3**	64.4	64.4
	50	6.2	15.7	41.4	**59.7**	58.7	58.7
	75	5.0	15.7	42.2	**60.9**	59.0	59.2
	100	4.2	13.8	40.5	**58.8**	57.3	57.6

Table 3: Accuracy on the number of authors for all approaches on the 3 datasets.

formance of our model on this author. But what is the overall impact of additional features on the model performance?

Adding stylometric and hybrid features in the first experiment on the Blog corpus, with 10 authors, the accuracy decreases from 65.4 to 62.4%. However, the macro-averaged F1-score we report using these features is higher, at 61.4% instead of 56.7% when no features are added.

This behavior of BertAA is illustrated in the confusion matrices in Figure 3, in which we report the accuracy per class (i.e. per author). Surprisingly, BertAA is stuck at 0% accuracy on certain authors, as it tends to allocate all the texts to a sub-set of authors, which can lead to a good accuracy but a lower macro-averaged F1-score. On the other hand, adding other features (stylometric and hybrid) improves the macro-averaged F1-score, but reduces the accuracy in that specific case.

According to our experiments, as illustrated in Figure 4, the F1-score on the Blog Authorship corpus for 5, 10, 25, 50, 75, and 100 authors improves by 2.70% (relative) when stylometric features are added to BertAA, and by 2.73% (relative) when including hybrid features.

In the blog corpus, more than 2'300 texts are collected per author, for the top 10 authors, which offers a sufficient quantity of training data, and a limited number of authors. But this does not guarantee that our model behaves well under a smaller set of training data and a wider classification task, such as on the IMDb62 dataset.

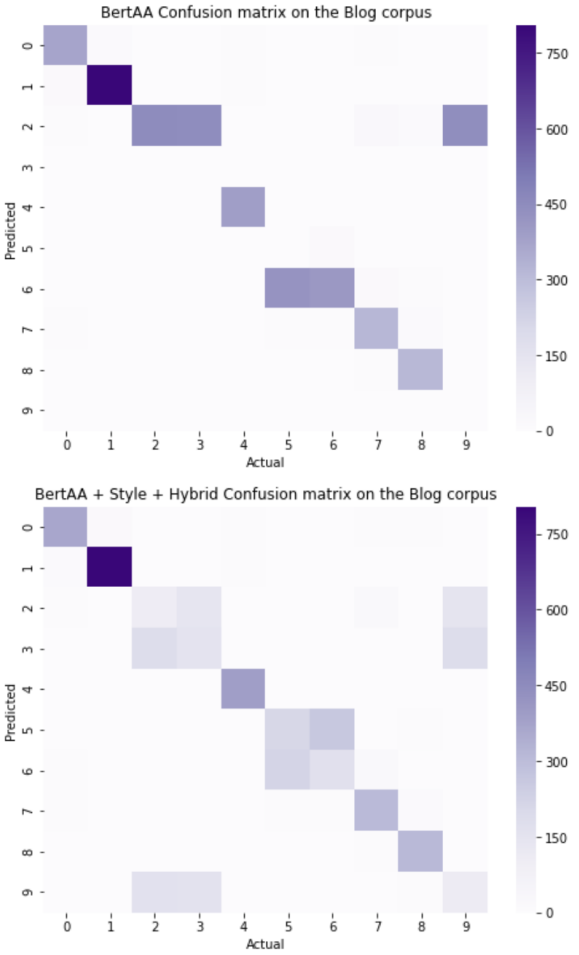

Figure 3: Confusion matrix of BertAA and BertAA + Style + Hybrid on the Blog corpus

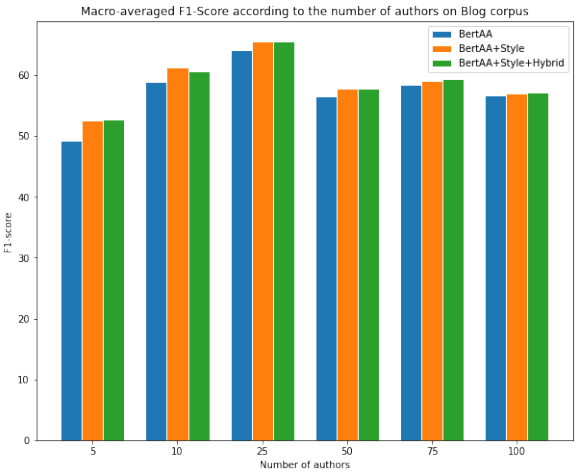

Figure 4: Macro-averaged F1-score when including stylometric and hybrid features according to number of authors.

5.3 More authors, less data

We ran additional experiments on the IMDb62 dataset with a larger number of authors (62), and fewer training samples per author (1'000). To replicate the setup of most methods presented in Table 5, 20% of the data were used as a test sample. The split is made randomly, since no standardized training and testing corpus exists for all these datasets. In Table 5, we report the performance of the various BertAA architectures and compare our approaches to various methods including current SOTA.

Approach	Accuracy
LDA+Hellinger (El et al.)	82
Word Level TF-IDF	91.4
CNN-Char (Ruder et al., 2016)	91.7
Comp.Att.+Sep.Rec. (Song et al., 2019)	91.8
Token-SVM (Seroussi et al., 2014)	92.52
SCAP (Frantzeskou et al., 2006)	94.8
Cont. N-gram Char (Sari et al., 2017)	94.8
(C+W+POS)/LM (Kamps et al., 2017)	95.9
N-gram + Style (Sari et al., 2018)	95.9
Syntax CNN(Zhang et al., 2018)	**96.2**
BertAA + Style + Hybrid - 1 epoch	88.7
BertAA + Style - 3 epochs	91.1
BertAA + Style + Hybrid - 5 epochs	92.3
BertAA + Style + Hybrid - 10 epochs	93.0

Table 5: Accuracy of various approaches on IMDb62

The Syntax-enriched CNN presented in (Zhang et al., 2018) reached an accuracy of 96.2%. Most other approaches lie between 91 and 94%. Considering that IMDb62 offers 1'000 training samples per author, the training of BertAA over a single epoch did not perform well. We then increased the number of training epochs and reached 92.3% at 5 epochs, and up to 93.0% at 10 epochs. This highlights the limitations of our model in situations with less training data and more authors.

Figure 5 plots the relative accuracy of BertAA over the number of authors for all three datasets. The starting point at 100 represents the accuracy reached by the model at 5 authors. A decreasing trend would therefore illustrate that the model accuracy is negatively impacted by a larger number of authors. On Enron and the Blog, the decrease in accuracy is limited, since 95 to 97% of the performance on 5 authors is maintained at 100 authors. The largest decrease occurs for IMDb dataset, at around 87% of the accuracy at 5 authors for 100 authors. This can likely be explained by the fact that IMDb comments are published publicly on the IMDb website, and that many authors might read comments of a movie before publishing theirs. Words, topics, punctuation, or phrases might therefore be re-used by some authors when publishing their comments.

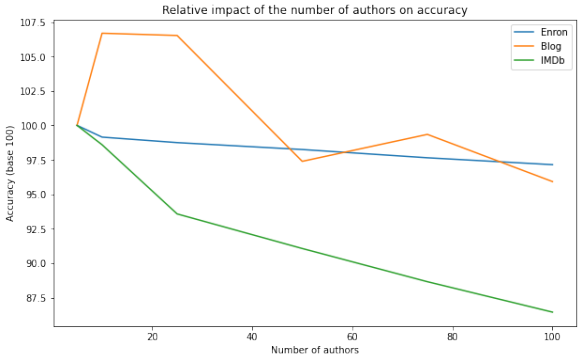

Figure 5: Relative impact of the number of authors on accuracy, base being the case with 5 authors.

Since the accuracy increases over the number of epochs on IMDb 62, in the next section, we further explore the impact of the number of training epochs on the model performance.

5.4 How much fine-tuning is too much?

In literature, the recommended number of epochs for BERT has been set between 2 and 4 (Sun et al., 2020), 3 being a common choice. In order to explore the effect of the number of training epochs on the model's accuracy, we report in Figure 6 the accuracy for the IMDb62 dataset using several models (BertAA, BertAA + Style, BertAA + Style + Hybrid), and compare it to the baseline TF-IDF. BertAA + Style appears to be the best performing model and starts to offer better performances

than TF-IDF after 4 epochs. However, no peak performance is reached, and the accuracy is still improving after 10 epochs, although to a lesser extent. The impact of the training epochs on a dataset with 62 authors is higher since BertAA is not performing as well, but we can suppose that the impact of the number of epochs is reduced on a smaller set of authors. We have chosen to train our models on 5 epochs for most of our experiments since it offers a good tradeoff between the training time and accuracy.

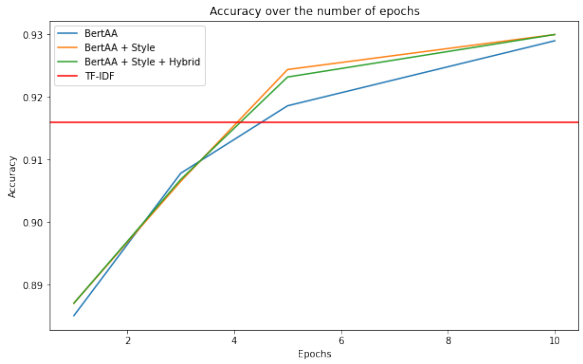

Figure 6: Accuracy over the number of epochs

6 Discussion

Our approach can be summarized as an extrapolation of BERT's general outstanding scores, for AA. We show that reaching SOTA results can be achieved using only a single dense layer and a softmax activation on top of a pre-trained BERT with a few training epochs. We highlighted that BertAA performs well on rather short texts, few imbalances in the number of texts per author, and a large number of texts per author.

Previous works (Sari et al., 2018) have shown that using stylometric and hybrid features improves the accuracy of AA tasks. We also show that adding such features when leveraging pre-trained language models can improve the macro-averaged F1-Score by 2.7% (relative) on average, although impacting the accuracy.

The use of BertAA should be limited to cases where BERT is itself a good candidate, i.e. when there are sufficient training data per author. This condition might be hard to reach in real applications for police investigations. Short texts and few imbalances have also been identified as requirements for better model performances. Our model is also currently unable to perform text similarity evaluation in the context of Authorship Verification.

There are many possible extensions to this work. According to our experiments on Enron, Blog Authorship, and IMDb corpora, AA can successfully leverage transformers-based language representation models. So far, we have not performed further pre-training of BERT on the target domain, which could also help BertAA. We have not tried yet to use another pre-trained language model. Future works should also explore other model architectures like RoBERTa(Liu et al., 2019), or try to extract additional stylometric, hybrid, profiling, or content-related features. Including the computation of similarity metrics on embeddings learned through the BERT fine-tuning would also be a way to compare the similarity between texts for Authorship Verification tasks. We will also explore BertAA in AA tasks on ASR transcripts, where punctuation and capital letters are not present for example.

7 Conclusion

With the rise of Deep Learning and Transformers in Natural Language Processing, feature engineering and text preprocessing are less needed.

In this work, we presented an approach based on fine-tuning of a pre-trained BERT for author classification. This is one of the very first attempts to analyze the performances of pre-trained language model fine-tuning for in-domain AA. We showed that our approach, which leverages BERT, reaches competitive performances on three well-known benchmark datasets, even on a large number of authors. The model best performs when sufficient training data per author are available, there is no large class imbalance, and texts remain rather short.

We also show that in a large scale AA task, adding stylometric and hybrid features to BertAA in an ensemble model can improve the macro-averaged F1-score by 2.7% (relative) on average. Finally, we set a new benchmark on the full IMDb Authorship Attribution Corpus for 5, 10, 25, 50, 75, and 100 authors. Future works will explore adding features to BertAA, further pre-training BERT on target-domain, exploring other pre-trained language models, and extending our approach to Authorship Verification. [1]

[1] Code and datasets are available here

8 Acknowledgment

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022). The second author, Esaú Villatoro-Tello, was supported partially by Idiap, SNI-CONACyT, CONACyT project grant CB-2015-01-258588, and UAM-C Mexico during the elaboration of this work.

References

Opeyemi Aborisade and Mohd Anwar. 2018. Classification for Authorship of Tweets by Comparing Logistic Regression and Naive Bayes Classifiers. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 269–276.

Ben Allison and Louise Guthrie. Authorship Attribution of E-Mail: Comparing Classifiers Over a New Corpus for Evaluation. page 5.

Brandy L. Aven. 2015. The Paradox of Corrupt Networks: An Analysis of Organizational Crime at Enron. *Organization Science*, 26(4):980–996. Publisher: INFORMS.

Andrea Bacciu, Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, and Julinda Stefa. 2020. Cross-Domain Authorship Attribution Combining Instance-Based and Profile-Based Features. page 14.

Douglas Bagnall. 2016. Author Identification using Multi-headed Recurrent Neural Networks. *arXiv:1506.04891 [cs]*. ArXiv: 1506.04891.

Georgios Barlas and Efstathios Stamatatos. 2020. Cross-Domain Authorship Attribution Using Pre-trained Language Models. In Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis, editors, *Artificial Intelligence Applications and Innovations*, volume 583, pages 255–266. Springer International Publishing, Cham. Series Title: IFIP Advances in Information and Communication Technology.

İlker Nadi Bozkurt, Özgür Bağlıoğlu, and Erkan Uyar. 2007. Authorship attribution: performance of various features and classification methods. In *22nd International Symposium on Computer and Information Sciences, ISCIS 2007 - Proceedings*, pages 158–162. IEEE. Accepted: 2016-02-08T11:41:59Z.

Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, and Isaac Woungang. 2013. Authorship verification for short messages using stylometry. In *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–6, Athens, Greece. IEEE.

Paul Clough, Colum Foley, Cathal Gurrin, Gareth Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Murdock, editors. 2011. *Advances in Information Retrieval: 33rd European Conference on IR Resarch, ECIR 2011, Dublin, Ireland, April 18-21, 2011, Proceedings*. Information Systems and Applications, incl. Internet/Web, and HCI. Springer-Verlag, Berlin Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Sara El, Manar El Bouanani, Ensias Mohammed, Mohammed Ben, Abdallah Regragui, and Madinat Al. *General Terms Authorship analysis*.

Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas. 2006. Source Code Author Identification Based on N-gram Author Profiles. In *Artificial Intelligence Applications and Innovations*, IFIP International Federation for Information Processing, pages 508–515, Boston, MA. Springer US.

Jade Goldstein-Stewart, Ransom Winder, and Roberta Sabin. 2009. Person Identification from Text and Speech Genre Samples. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 336–344, Athens, Greece. Association for Computational Linguistics.

Oren Halvani, Lukas Graner, Roey Regev, and Philipp Marquardt. 2020. An Improved Topic Masking Technique for Authorship Analysis. *arXiv:2005.06605 [cs]*. ArXiv: 2005.06605.

Jaap Kamps, Giannis Tsakonas, Yannis Manolopoulos, Lazaros Iliadis, and Ioannis Karydis. 2017. *Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings*. Springer. Google-Books-ID: it0zDwAAQBAJ.

Bryan Klimt and Yiming Yang. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Machine Learning: ECML 2004*, volume 3201, pages 217–226. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author: Determining If Two Documents Are Written by the

Same Author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.

Xuan Li. Authorship Attribution on the Enron Email Corpus. page 27.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

David Madigan, Alexander Genkin, David D Lewis, Shlomo Argamon, Dmitriy Fradkin, Li Ye, and David D Lewis Consulting. Author Identification on the Large Scale. page 20.

David Madigan, Alexander Genkin, David D. Lewis, and Dmitriy Fradkin. 2005. Bayesian Multinomial Logistic Regression for Author Identification. *AIP Conference Proceedings*, 803(1):509–516. Publisher: American Institute of Physics.

Lukas Muttenthaler, Gordon Lucas, and Janek Amann. Authorship Attribution in Fan-Fictional Texts given variable length Character and Word N-Grams. page 9.

Hendrik Neumann and Martin Schnurrenberger. *E-Mail Authorship Attribution applied to the Extended Enron Authorship Corpus (XEAC)*.

Chen Qian, Tianchang He, and Rao Zhang. Deep Learning based Authorship Identification. page 9.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. Character-level and Multi-channel Convolutional Neural Networks for Large-scale Authorship Attribution. *arXiv:1609.06686 [cs]*. ArXiv: 1609.06686.

Chakaveh Saedi and Mark Dras. 2019. Siamese Networks for Large-Scale Author Identification. *arXiv:1912.10616 [cs]*. ArXiv: 1912.10616 version: 1.

Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or Style? Exploring the Most Useful Features for Authorship Attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yunita Sari, Andreas Vlachos, and Mark Stevenson. 2017. Continuous N-gram Representations for Authorship Attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 267–273, Valencia, Spain. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. Effects of Age and Gender on Blogging. page 6.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship Attribution with Topic Models. *Computational Linguistics*, 40(2):269–310.

Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional Neural Networks for Authorship Attribution of Short Texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain. Association for Computational Linguistics.

Juan Soler-Company and Leo Wanner. 2017. On the Relevance of Syntactic and Discourse Features for Author Profiling and Identification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 681–687, Valencia, Spain. Association for Computational Linguistics.

Wei Song, Chen Zhao, and Lizhen Liu. 2019. Multi-Task Learning for Authorship Attribution via Topic Approximation and Competitive Attention. *IEEE Access*, 7:177114–177121. Conference Name: IEEE Access.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to Fine-Tune BERT for Text Classification? *arXiv:1905.05583 [cs]*. ArXiv: 1905.05583.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*. ArXiv: 1706.03762.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*. ArXiv: 1910.03771.

Haoran Wu, Zhiyong Xu, Jianlin Zhang, Wei Yan, and Xiao Ma. 2017. Face recognition based on convolution siamese networks. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5.

Min Yang and Kam-Pui Chow. 2014. Authorship Attribution for Forensic Investigation with Thousands of Authors. In *ICT Systems Security and Privacy Protection*, IFIP Advances in Information and Communication Technology, pages 339–350, Berlin, Heidelberg. Springer.

Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. 2018. Syntax Encoding with Application in Authorship Attribution. In *Proceedings of the 2018*

Conference on Empirical Methods in Natural Language Processing, pages 2742–2753, Brussels, Belgium. Association for Computational Linguistics.