Reduce & Attribute: Two-Step Authorship Attribution for Large-Scale Problems

Michael Tschuggnall
Universität Innsbruck
michael.tschuggnall@uibk.ac.at

Benjamin Murauer
Universität Innsbruck
b.murauer@posteo.de

Günther Specht
Universität Innsbruck
guenther.specht@uibk.ac.at

Abstract

Authorship attribution is an active research area which has been actively studied for many decades. Nevertheless, the majority of approaches consider problem sizes of a few candidate authors only, making them difficult to apply to recent scenarios incorporating thousands of authors emerging due to the manifold means to digitally share text. In this study, we focus on such large scale problems and propose to effectively reduce the number of candidate authors before applying common attribution techniques. By utilizing document embeddings, we show on a novel, comprehensive dataset collection that the set of candidate authors can be reduced with high accuracy. Moreover, we show that common authorship attribution methods substantially benefit from a preliminary reduction if thousands of authors are involved.

1 Introduction

Correctly determining the author of an anonymous text has been researched for several decades. Undoubtedly the most groundbreaking work in this area has been conducted in 1964 by Mosteller and Wallace, who attempted to automatically assign the authors of the Federalist Papers by utilizing a simple, but yet efficient statistical approach operating on function words (Mosteller and Wallace, 1964). By showing that writers can indeed be distinguished by their writing style, many approaches have been published in the following years, proposing enhancements by incorporating a variety of so-called stylometric features, methods and learning techniques (Stamatatos, 2009). By categorizing the problem of authorship attribution as a special form of text categorization (Sebastiani, 2002), also the respective methods in terms of different machine learning algorithms are effectively in use. With the advent of deep learning, also several approaches have been proposed recently which utilize comprehensive neural networks. Nevertheless, a recent comparison of all submitted approaches to the cross-domain authorship attribution task at PAN[1] indicates that deep learning is currently not able to surpass traditional methods (Kestemont et al., 2018).

Regardless of the features and methods used, the efficacy of an approach can only be measured by using appropriate datasets. Thereby, the majority of existing approaches focus only on a small number of candidate authors (up to 20, most of the times ten or less, e.g., (Stamatatos, 2009; Juola, 2012)). Only a few studies have examined the performance of authorship attribution approaches on larger amounts of possible authors, e.g., 114 (Madigan et al., 2005), 145 (Luyckx and Daelemans, 2008), 808 (Hitschler et al., 2017) or 1,000 (Shrestha et al., 2017) candidates. To the best of our knowledge, only three studies dive into the multiple thousands of authors: 10,000 (Koppel et al., 2006, 2011) and even 100,000 (Narayanan et al., 2012). While both studies agree that authorship attribution at large scale is substantially more difficult, they still show the potential of performing identification with acceptable accuracy. To be precise, both approaches report good performances only in scenarios where either an *"I don't know"* answer is also accepted (Koppel et al., 2011) or the attribution is not precise, i.e., the statement that the correct author is among the top k ones is sufficient (Narayanan et al., 2012).

Motivated by previous findings, which showed that direct authorship attribution is not feasible in a large scale scenario, we contribute to this field by proposing a two-step approach in this study. Specifically, we propose at first to reduce the number of candidate authors while keeping the correct

[1] PAN is an internationally renowned initiave in the field of digital text forensics and stylometry, https://pan.webis.de

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 951–960
Hong Kong, China, November 3-4, 2019.

author in the reduced set with reasonable accuracy. Incorporating the promising results reported by using embedding representations ((Posadas-Durán et al., 2017)), we also find that a vector space based on document embeddings (Le and Mikolov, 2014) in combination with cosine similarity yields the best results for reducing candidate authors in the large scale. As authorship attribution generally heavily depends on the datasets used (e.g., the text type or the number of training and test documents (Luyckx and Daelemans, 2011; Potthast et al., 2016)) and the datasets used in the mentioned studies are not available,[2] we created a collection of 179 individual, novel datasets using a large question-and-answer (Q&A) network on which we extensively test our models. Using these datasets, we finally also show that a preliminary reduction of candidates substantially improves the overall accuracy of finding the correct author in large settings.

At a glance, our contributions are as follows: (1) We evaluate document embeddings in combination with cosine similarity and show that they outperform n-grams (which have proven to be among the most discriminating features, e.g., Stamatatos, 2013; Kestemont et al., 2018) with respect to the task of reducing the number of candidate authors in large scenarios. Thereby we show that neither n-grams used with similarity measures nor support vector machines are able to keep up with document embeddings. (2) We show that eliminating candidate authors using our approach in large settings—prior to performing direct attribution—substantially improves the accuracy of commonly used attribution methods. (3) We created a novel dataset collection based on a large Q&A network, consisting of 179 sub-datasets, each of which features six up to nearly 20,000 authors. To ensure reproducibility and to encourage further research, we make the dataset publicly available to the research community.

The remainder of this paper is organized as follows: At first, Section 2 summarizes related work and subsequently Section 3 presents the dataset. The proposed approach to reduce candidates using document embeddings and its evaluation is presented in Section 4, while Section 5 shows its impact on direct authorship attribution. Finally, Section 6 concludes and discusses future work.

[2]The dataset used by (Koppel et al., 2011) is partly available (see Section 4.2)

2 Related Work

Features and Methods

In the last decades many different features have been proposed for stylometry problems, which can basically be categorized into lexical, syntactic, structural and other specialized features (Stamatatos, 2009; Stein et al., 2011). For the specific task of authorship attribution, lexical metrics are predominant. Thereby, features are utilized on the character- and word-level, including character/word frequencies (Zheng et al., 2006), average word- and sentence lengths (Grieve, 2007), function word frequencies (Argamon et al., 2003; Zhao and Zobel, 2005), bag-of-words (BOW, Agun and Yilmazel, 2017) or especially character/word n-grams (Sapkota et al., 2015; Stamatatos, 2013; Schwartz et al., 2013 and variants thereof (Stamatatos, 2017). Moreover, derived features such as different readability measures (Tweedie and Baayen, 1998) or compression ratios (Marton et al., 2005) have also been investigated.

Syntactic features include the analysis of (n-grams of) Part-of-Speech (POS) tags (Zhao and Zobel, 2007) or the analysis of the parse tree of sentences (Luyckx and Daelemans, 2008; Tschuggnall and Specht, 2014), whereas structural features analyze indicators like the average paragraph length or the use of indentation (Zheng et al., 2006). In addition, various additional metrics have been proposed, e.g., the analysis of spelling and grammatical errors present in a text (Koppel and Schler, 2003).

From a methodical view, a wide range of machine learning techniques is in use, including Bayesian models, logistic regression, support vector machines (SVM) or decision trees. In most cases, the studies apply multiple classifiers and compare their results (e.g., see the surveys of Stamatatos, 2009; Juola and Stamatatos, 2013; Potthast et al., 2016; Kestemont et al., 2018). Recently, deep learning techniques have also been applied to authorship attribution problems. Thereby, various approaches have been proposed which use convolutional neural networks (CNN, Rhodes, 2015; Shrestha et al., 2017). With respect to input features, embeddings on different levels are heavily utilized, e.g., on words (*word2vec*, Mikolov et al., 2013), documents (*doc2vec*, Le and Mikolov, 2014), or n-grams of characters (Shrestha et al., 2017) or POS-tags (Hitschler et al., 2017). In addition, studies have reported that

embeddings are also highly efficient when fed into common machine learning techniques like SVMs or logistic regression (Agun and Yilmazel, 2017; Posadas-Durán et al., 2017).

In general, it has been shown that especially character n-grams and variants thereof are among the most discriminating features, which perform very well with common machine learning techniques such as out-of-the-box SVMs (e.g., Stamatatos, 2013; Kestemont et al., 2018), ensembles (e.g., Custódio and Paraboni, 2018) as well as with recent deep learning methods (e.g., Shrestha et al., 2017; Rhodes, 2015). Due to this success of n-grams, we chose to rely on them as a reference as is shown in Sections 4 and 5.

Large-Scale Authorship Attribution

As mentioned earlier, the majority of authorship attribution approaches target a relatively small number of candidate authors (up to at most 20). The few studies considering more than a hundred authors utilize various lexical features such as character n-grams together with syntactic features, and achieve accuracies ranging from 50-80% (Madigan et al., 2005; Luyckx and Daelemans, 2008). For about 800 authors, Hitschler et al. (2017) achieve 13% accuracy with a CNN, and Shrestha et al. (2017) also utilize a CNN to attribute the correct author out of 1,000 candidates with an accuracy of 36%.

Koppel et al. (2006, 2011) conducted two experiments on blogs with 10,000 authors. First, they achieve about 35% by using inverse-document-frequencies of stylistic features, represented in a vector space and compared using cosine similarity. In a second study, aiming for precision rather than recall (i.e., to rather output *don't know* than to guess), they use *space-free character 4-grams* with cosine similarity, and enhance their approach by iteratively evaluating randomized subsets of features. By doing so, they report a precision of 93% for the cases an answer is given.

Finally, the most comprehensive study with respect to number of candidate authors has been conducted by Narayanan et al. (2012), who evaluate different features with several machine learning techniques on a dataset consisting of 100,000 authors of blogs. In their study, the main focus is laid on security concerns, i.e., that the correct author can be identified in an attack. The authors show that a combination of a simple nearest neighbor approach with a regularized least squares classifier is able to detect the correct author of a blog in 20% of the cases and that the correct author is in the top-20 ranked candidates in 35% of the cases. Moreover and along the lines of Luyckx and Daelemans (2011) or Eder (2010), it is shown that the size of available training/test texts substantially influences the performance.

As the studies of (Koppel et al., 2011) and (Narayanan et al., 2012) are the only ones targeting authors in the large scale, we will also use these studies as references throughout this paper (in terms of their methodology and reported results). Nevertheless, a direct comparison is difficult as they either target different aims and/or the underlying datasets are not or only partly available. In contrast to these studies, we propose a novel two-step method in this paper and provide comprehensive large scale studies alongside, which can easily be reproduced in both methods as well as data used.

3 The SE-179 Dataset Collection

For the task of authorship attribution, a suitable dataset has to consist of realistic documents where the authorship of each document can undoubtedly be attributed to a single author. In the case of single-domain or single-topic analyses, it has to additionally be assured that all candidate authors write about the same topics—such that they cannot be exposed by simply looking at specific topic-related content words. Along the lines of Kestemont et al. (2018), who showed that data from Q&A forums can successfully be employed to analyze the writing style, we also used the same Q&A platform, namely *StackExchange*[3], to create our dataset.

The StackExchange network consists of several sites where people answer questions related to specific topics (*Stackoverflow* being the most popular site). In contrast to Kestemont et al. (2018) where only selected posts of selected StackExchange sites were crawled, we use the provided data dump[4] containing all questions and answers of all sites. Because the posts for each site are related to a single topic (e.g., photography), it allows us to create individual datasets from each site. Thereby the procedure for creating a dataset from a site was as follows:

[3] https://stackexchange.com

[4] provided directly by StackExchange at https://archive.org/details/stackexchange

(1.) We collected all questions and answers by all users participating in the site. (2.) We removed all posts that were edited by a person different from the original author (in the StackExchange network, basically everyone can edit anyone's posts). (3.) We cleaned each post, i.e., we removed code snippets, block quotes, bullet lists, embedded images and replaced links with `$URL$`. We then dismissed all posts containing less than ten tokens after cleaning. (4.) We combined all remaining posts of each user into a single document and removed all users with less than 500 tokens. Subsequently, we divided each document into a training and a test document. Thereby we assured that each training and test document contains at least 500 tokens, and in case this was not possible (because there were less than 1,000 tokens available), we only kept the training document to increase the number of candidates. I.e., there exist several training documents where there exists no corresponding test document. Note that it is a common procedure to fix training and test documents in order to ensure reproducibility (Stamatatos et al., 2018).

Consequently, we created a balanced, single-topic dataset from each site, containing different numbers of authors depending on the size of the community of the respective site. Table 1a shows the statistics of the resulting 179 datasets[5], including the average tokens per document (avg t/d) as well as the ratio between number of training to test documents (ttr). We consequently call the overall dataset collection *SE-179*. With respect to languages throughout the collection, the predominant one is English, but also individual problems in different languages are present[6].

For our study, the datasets containing many authors as listed in Table 1b are of high interest, nevertheless we conducted our experiments also on all other datasets as is detailed in Section 4.2. By doing so we can avoid potential biases towards specific datasets. As we are concerned about reproducibility, we make the SE-179 collection publicly available and encourage other researchers to utilize it according to their needs.[7]

[5]We didn't process the *Stackoverflow* site due to computational limitations with respect to its size.

[6]I.e., one for each Chinese, Esperanto, French, German, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish and Ukrainian.

[7]The dataset is available at https://doi.org/10.5281/zenodo.3441861.

authors	datasets	train avg t/d	test avg t/d	ttr
≤ 10	3	757	1143	45%
11–100	31	740	569	47%
101–250	34	698	569	46%
251–500	31	703	577	46%
501–1,000	33	684	574	45%
1,001–5,000	38	669	566	45%
5,001–10,000	7	677	572	44%
$> 10{,}000$	2	613	552	40%

(a) General statistics

site origin	authors	site origin	authors
Superuser.com	19,272	Softwareengineering	6,351
Serverfault.com	16,450	Electronics	6,119
Askubuntu.com	9,830	Unix	5,507
Physics	7,418	Stats	5,307
Mathoverflow.net	6,361	Wordpress	3,898

(b) Top 10 large scale datasets.

Table 1: Statistics of the SE-179 dataset collection including average tokens per document (avg t/d) and the ratio between number of training and test documents (ttr).

4 Reducing Candidate Authors

In this section, we outline our approach of effectively reducing the number of candidate authors by utilizing document embeddings. After describing the applied technique in Section 4.1 as well as the reference implementations used, we show the results in Section 4.2.

4.1 Methods

Document Embeddings

Based on the promising results reported by Posadas-Durán et al. (2017), we also utilize document embeddings with doc2vec (Le and Mikolov, 2014). Considering the two possible representation techniques provided by doc2vec, i.e., distributed memory (DM) and distributed bag of words (DBOW), we evaluated the three basic models (i) DM using concatenation (DM/concat), (ii) DM using average (DM/avg), (iii) DBOW as well as the two combinations (iv) DBOW+DM/concat and (v) DBOW+DM/avg. For each model, we evaluated vector sizes (dimensions) of $d = \{100, 200, 300\}$ (or the double in case of the combined models) and relied on the default/optimal settings found by Posadas-Durán et al. (2017) for the specific model parameters. With respect to the textual input, we at first tokenize the text and then experiment with the following settings to compute the embeddings:

– *type*: we either provide the text as is (*unigram*) or we compute *bigrams* of the words

– *stem*: decides whether stemming should be applied or not

– *windowing*: if a window length (w_l) is set, we traverse the document using a sliding window containing w_l tokens and thereby create new "documents" for each author. The window step w_s defines the number of tokens the window is shifted after each iteration. Additionally, we compute models from the original documents without windowing.

For the final reduction of candidate authors according to a given test document, the procedure is as follows: (1.) According to the previously described settings, we learn models from all available training documents of all authors. During this step, all documents are assigned a vector of dimension d (the model dimension), which form an according vector space. In case windowing is used, each author is represented by several vectors (one per window). (2.) For the given test document, we apply the same preprocessing steps (i.e., input type, stemming and windowing) and make use of the functionality provided by doc2vec to estimate a vector for a document that was not seen during learning. (3.) Similar to Koppel et al. (2011) we then compare the test document's vector with all document vectors in the vector space by computing the cosine similarity. Ordering by this similarity and using the top-k authors finally allows reducing the set of candidates to an arbitrary extent. In case windowing is used, i.e., when there are multiple documents by each author, we use the average of the similarities of all the author's document vectors.

Reference Implementations

To compare the proposed approach, we reimplemented the approach described by Koppel et al. (2011). Specifically, in this approach so-called *space-free* character 4-grams are computed for each document and their normalized frequencies form the basis for a vector space. By repeatedly (k_1 times) selecting k_2% of the feature set randomly, cosine similarity is used to compare the documents. In our reimplementation we used the optimal values as reported, i.e., $k_1 = 100$ and $k_2 = 40\%$. As an additional reference, we used regular n-grams instead of the space-free variants.

authors	**model**	d	stem	w_l	w_s	type
≤ 10	DM/avg	200	yes	–	–	unigram
11–100	DM/avg	100	yes	–	–	unigram
101–250	DM/concat	100	yes	300	50	unigram
251–500	DM/concat	100	yes	300	50	unigram
501–1,000	DM/concat	100	yes	300	50	unigram
1,001–5,000	DM/concat	200	yes	300	50	unigram
5,001–10,000	DM/concat	100	yes	–	–	unigram
$> 10{,}000$	DM/concat	100	yes	–	–	unigram

Table 2: Best doc2vec models with respect to number of authors.

4.2 Estimating Best Reduction Models

Contrary to Narayanan et al. (2012) we find it more suitable to not test whether the correct author is in the top-k results, but to evaluate how often s/he is in the result set after reducing by percentage (e.g., eliminating 90% of the candidates). This makes especially sense as we are dealing with 179 different datasets of different sizes, where a comparison of the top-k results with a fixed k is not meaningful (e.g., it makes a huge difference if the correct author is in the top-5 in a dataset containing 20 authors or in one containing 16,000 authors). Thus, we experimented with the reduction rates 10-90%, 95%, and 99%, and measured the hit rate, i.e., the percentage of how often the correct author is still in the reduced candidate set.

In a first preliminary step, we aimed to find the best models with respect to reduction rate and candidate author size[8]. We evaluated on all 179 datasets using the respective training documents for learning and the test documents for testing. For the larger datasets, we tested on 1,000 randomly selected test documents (as has been done by Koppel et al. (2011)).

After conducting the experiment, we found that the reduction rate doesn't make any difference with respect to the model type and that the best performing models only depend on the number of authors. Table 2 shows the best settings for different number of authors, computed by using the average of all corresponding datasets and regardless of the reduction rate.[9] It can be seen that stemmed unigrams work best in all cases and that windowing is not the preferred option when looking at large (and small) candidate sizes.

Using the best models found (depending on the number of candidate authors) we evaluated their

[8] E.g., what is the best doc2vec-model for reducing an 8,000 author dataset by 70%?

[9] Note that we cannot provide single hit rates for each configuration, as they significantly depend on the reduction rate.

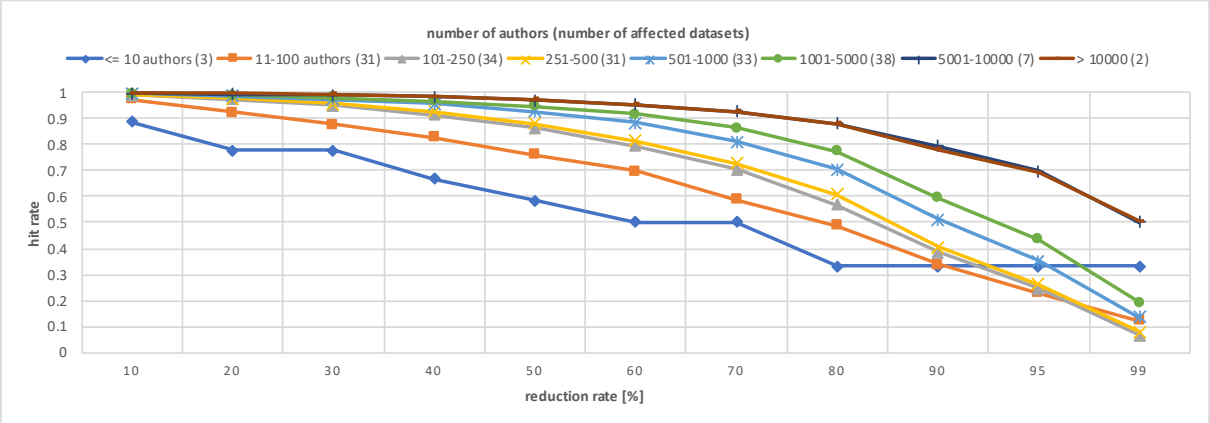
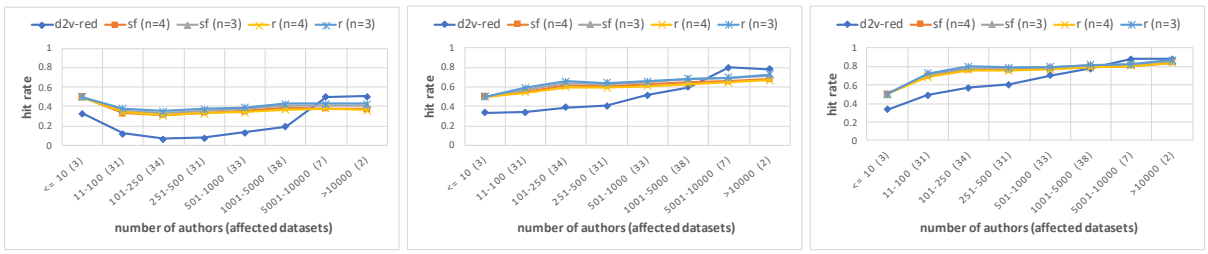


Figure 1: Hit rates for the doc2vec reduction models averaged over all datasets with respect to reduction rate and candidate author size.



(a) reduction rate = 99%

(b) reduction rate = 90%

(c) reduction rate = 80%

Figure 2: Comparison of the doc2vec reduction (d2v-red) with the method proposed by Koppel et al. (2011) using space-free (sf) and regular (r) n-grams. The y-axis shows the average hit rates over all datasets for different reduction rates, grouped by number of candidate authors.

performance with respect to the reduction rate. That is, we reduced the number of candidates by the respective percentage and measured—in terms of hit rate—if the correct author is still in the remaining set. As expected and can be seen in Figure 1, the performance decreases as the reduction rate increases. In general, the more candidate authors, the better the model is able to filter out irrelevant ones: E.g., the datasets having more than 5,000 authors could be reduced by 50% with a hit rate of 0.97, and by 80% with a hit rate of 0.88. When reducing these datasets by 99%, a hit rate of approximately 0.5 remains.

In a further experiment, we compared the reduction results to the reference systems described in Section 4.1, i.e., with regular and space-free n-grams as proposed by Koppel et al. (2011), Figure 2 exemplarily shows the average results over all datasets, grouped by the number of candidate authors for the reduction rates 99%, 90% and 80%, respectively. Regardless of the individual reduction rate, the doc2vec model is inferior to the other models for datasets having less than 5,000 authors, but can significantly[10] exceed them when more authors are involved. For example, when reducing candidates by 90% in a 5,000+ candidate author setting, it is able to keep the correct author with a hit rate of 0.69 in average, whereas the best other model (regular 3-grams) achieves a hit rate of 0.60. Although the superiority of our doc2vec model decreases with lower reduction rates, it is still better than the other models for all reduction rates in scenarios having more than 5,000 candidate authors, as is shown in Table 3.

5 Attribution on Reduced Candidates

In the previous section, we have shown that the number of candidate authors in large authorship attribution problems can effectively be reduced by compiling a document embedding model based on word unigrams. As a follow-up, we wanted to assess the influence of this reduction technique for state-of-the-art authorship attribution methods. The basic idea is to apply a two-step attribution by

[10] We computed a McNemar's test (Dietterich, 1998) and interpreted $p < 0.05$ as significant.

red.	d2v-red	sf (n=4)	sf (n=3)	r (n=4)	r (n=3)
10%	**0.998**	0.994	0.993	0.994	0.993
20%	**0.995**	0.988	0.986	0.988	0.986
30%	**0.991**	0.982	0.978	0.981	0.979
40%	**0.984**	0.973	0.969	0.973	0.971
50%	**0.972**	0.958	0.954	0.958	0.959
60%	**0.954**	0.931	0.932	0.937	0.938
70%	**0.928**	0.890	0.896	0.895	0.904
80%	**0.881**	0.812	0.827	0.811	0.834
90%	**0.792**	0.659	0.695	0.653	0.700
95%	**0.697**	0.551	0.585	0.538	0.593
99%	**0.501**	0.377	0.412	0.376	0.431

Table 3: Comparison of the proposed doc2vec reduction (d2v-red) with the method proposed by Koppel et al. (2011) using space-free (sf) and regular (r) n-grams. The table shows the average hit rates for the respective reduction rates (red.) over all 9 datasets containing more than 5,000 authors.

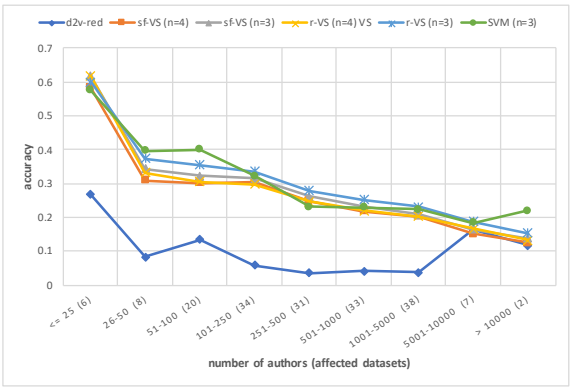


Figure 3: Accuracies for direct authorship attribution averaged over different candidate author sizes. *sf-VS* and *r-VS* represent space-free and regular n-grams, respectively, represented in a vector space as proposed by Koppel et al. (2011), and *SVM* refers to the reference implementation provided by PAN-2018.

transforming large scale problems to normal-scale problems: (1.) reduce the number of candidate authors, (2.) apply regular authorship attribution approaches for the remaining candidates.

5.1 Direct Attribution Baseline

In a first step, we created a baseline by computing the accuracies achieved for direct authorship attribution, i.e., for finding the correct author without any reduction. For this, we utilized the proposed reduction technique, but reduced to exactly one author instead of a set of authors. Similar to Section 4.1, we again utilized the approach of Koppel et al. (2011) with space-free and regular character 3-/4-grams in combination with a vector space and cosine similarity. As an additional reference for comparison, we made use of the reference implementation provided for the author identification task at the PAN 2018 event (Kestemont et al., 2018). It computes character 3-grams and makes classifications using a standard SVM, yet achieving competitive results by applying grid search (Murauer et al., 2018). The results averaged over candidate author sizes are presented in Figure 3, revealing that doc2vec is very imprecise for direct authorship attribution in non-large cases. The other approaches generally perform similarly, except for the largest datasets where the SVM achieved the best results (0.21 for the datasets with more than 10,000 authors).

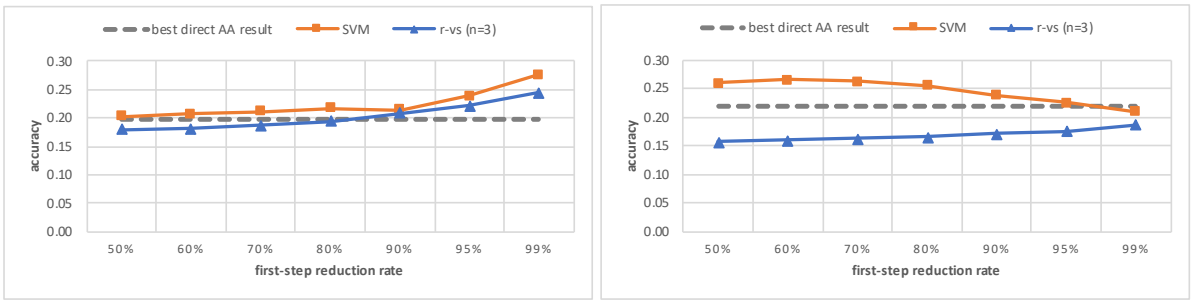
5.2 Two-Step Attribution

To measure the influence of the reduction proposed in Section 4, we conducted an experiment on the largest datasets by at first reducing the number of candidates, and subsequently performing regular authorship attribution. Considering the best results of direct attribution as presented previously, we evaluated the regular 3-gram vector space approach and the SVM implementation of PAN[11]. For each approach, we at first reduced the authors by the respective reduction rate, applied the two approaches and compared it to the best result achieved by direct attribution.

Figure 4 depicts the results for the 5,000-10,000 author datasets and for those having more than 10,000 authors, respectively. It can be seen that in general the accuracy–especially that of the SVM–can be improved, nevertheless, the best first-step reduction rate depends on the problem size: For datasets up to 10,000 candidates, the best option is to reduce the number of authors by 99% before performing attribution. On the contrary, the best accuracy for problems with more than 10,000 candidates could be achieved by using a reduction rate of 60%.

As stated initially in the paper, the evaluation results of authorship attribution techniques is highly dependent on the dataset (Luyckx and Daelemans, 2011; Potthast et al., 2016), and while our datasets within the SE-179 collection are highly heterogeneous with respect to topics and author sizes and also languages, they still belong to the genre of question-answering platforms. We therefore aimed to evaluate the dataset used by Koppel et al. (2011) to gain additional insight into the perfor-

[11]Note that for each test document a corresponding SVM has to be trained on the remaining candidate authors after reduction.



(a) 5k-10k authors datasets

(b) > 10k authors datasets

Figure 4: Two-step attribution with preliminary candidate reduction using doc2vec and cosine similarity.

mance on a different genre, i.e., blogs. Unfortunately, the data is only available in a raw form, i.e., we had to reconstruct the dataset as to the best of our knowledge incorporating the facts given by the authors. Consequently, we ended up with blog entries from 10,000 authors, each containing about 2,000 words, whereby the first 1,500 words were used for training and the last 500 for testing.

Considering the superiority of the SVM with character 3-grams in the previous experiment, we compared the performance of the SVM on all datasets of the SE-179 collection with more than 5,000 authors with the performance on the recreated blog dataset. Figure 5 shows the relative improvements of our proposed two-step attribution compared to the best direct attribution results, which are very similar for both the SE-179 datasets and the blog dataset, i.e., 0.202 and 0.204, respectively. As can be seen, a preliminary reduction of candidates substantially improves the performance, especially for the blog dataset for which the accuracy could be increased by more than 10%.

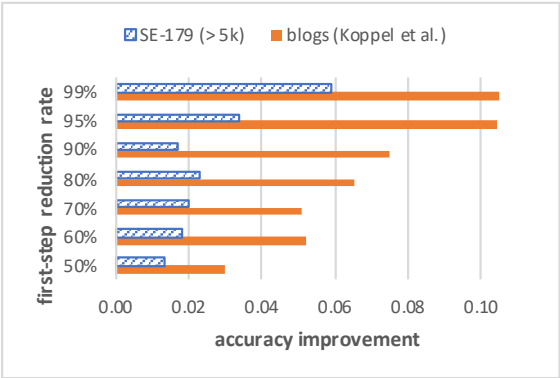


Figure 5: Relative performance improvements of the PAN-2018-SVM baseline implementation on SE-179 datasets with more than 5,000 authors and the reproduced blogs dataset (Koppel et al., 2011). The chart shows the improvements in accuracy that could be gained by applying preliminary candidate reduction.

6 Conclusion and Future Work

In this paper, we tackled the problem of large scale authorship attribution incorporating thousands of authors by first filtering candidate authors before the actual classification step. Extensive evaluations on a novel, publicly available dataset collection reveal that document embeddings in combination with cosine similarity are able to effectively reduce the number of candidate authors for large scale problems. We also outlined that a preliminary reduction increases the overall attribution accuracy in such cases.

As for future work, several open issues should be addressed. In this study, we relied on related work that suggests document embeddings, nevertheless other embedding techniques like *word2vec* (Mikolov et al., 2013), *fastText* (Joulin et al., 2016) or *GloVe* (Pennington et al., 2014) could be evaluated. Moreover, for the computation of similarities between document vectors, we relied on cosine similarity, whereas several other metrics should be evaluated. In the case of authorship attribution, we similarly utilized a common, established technique (SVM). As the reduction of problem sizes additionally enables the utilization of resource-intensive algorithms, more experiments are needed in that direction, especially using deep learning techniques. Finally, it would be worth investigating how this approach performs on cross-domain/-topic scenarios and other text genres like short messages or other social media contents.

References

Hayri Volkan Agun and Ozgur Yilmazel. 2017. Document embedding approach for efficient authorship attribution. In *Knowledge Engineering and Applications (ICKEA), 2017 2nd International Conference on*, pages 194–198. IEEE.

Shlomo Argamon, Marin Šarić, and Sterling S Stein. 2003. Style mining of electronic messages for multiple authorship discrimination: First results. In *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 475–480, Washington, DC, USA. ACM.

José Eleandro Custódio and Ivandré Paraboni. 2018. Each-usp ensemble cross-domain authorship attribution. *Working Notes Papers of the CLEF*.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Maciej Eder. 2010. Does size matter? authorship attribution, small samples, big problem. In *Proceedings of the Digital Humanities Conference*, London, UK. ALLC.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Julian Hitschler, Esther van den Berg, and Ines Rehbein. 2017. Authorship attribution with convolutional neural networks and pos-eliding. In *Proceedings of the Workshop on Stylistic Variation*, pages 53–58.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Patrick Juola. 2012. An overview of the traditional authorship attribution subtask. In *Notebook Papers of the 7th Evaluation Lab on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN)*, Rome, Italy.

Patrick Juola and Efstathios Stamatatos. 2013. Overview of the author identification task at pan 2013. In *Notebook Papers of the 9th Evaluation Lab on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN)*, Valencia, Spain.

Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018. Overview of the author identification task at pan-2018: cross-domain authorship attribution and style change detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al.*, pages 1–25.

Moshe Koppel and Jonathan Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, volume 69, pages 72–80, Acapulco, Mexico.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. 2006. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–660. ACM.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 513–520. Association for Computational Linguistics.

Kim Luyckx and Walter Daelemans. 2011. The effect of author set size and data size in authorship attribution. *Literary and linguistic Computing*, 26(1):35–55.

David Madigan, Alexander Genkin, David D Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. 2005. Author identification on the large scale. In *Proc. of the Meeting of the Classification Society of North America*, volume 13.

Yuval Marton, Ning Wu, and Lisa Hellerstein. 2005. On compression-based text classification. In *Advances in Information Retrieval*, pages 300–314. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

F. Mosteller and D. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley.

Benjamin Murauer, Michael Tschuggnall, and Günther Specht. 2018. Dynamic parameter search for cross-domain authorship attribution. *Working Notes of CLEF*.

Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. 2012. On the feasibility of internet-scale author identification. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 300–314. IEEE.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, Ildar Batyrshin, David Pinto, and Liliana Chanona-Hernández. 2017. Application of the distributed document representation in the authorship attribution task for small corpora. *Soft Computing*, 21(3):627–639.

Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Jakob Köhler, Winfried Lötzsch, Fabian Müller, Maike Elisa Müller, et al. 2016. Who wrote the web? revisiting influential author identification research applicable to information retrieval. In *European Conference on Information Retrieval*, pages 393–407. Springer.

Dylan Rhodes. 2015. Author attribution with cnns. *Avaiable online: https://www. semanticscholar. org/paper/Author-Attribution-with-Cnn-s-Rhodes/0a904f9d6b47dfc574f681f4d3b41bd840871b6f/pdf (accessed on 22 August 2016)*.

Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 93–102.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1880–1891.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *Computing Surveys*, 34(1):1–47.

Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, 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*, volume 2, pages 669–674.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 21(2):421–439.

Efstathios Stamatatos. 2017. Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 1138–1149.

Efstathios Stamatatos, Francisco Rangel, Michael Tschuggnall, Benno Stein, Mike Kestemont, Paolo Rosso, and Martin Potthast. 2018. Overview of pan 2018. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 267–285. Springer.

Benno Stein, Nedim Lipka, and Peter Prettenhofer. 2011. Intrinsic plagiarism analysis. *Language Resources and Evaluation*, 45(1):63–82.

Michael Tschuggnall and Günther Specht. 2014. Enhancing authorship attribution by utilizing syntax tree profiles. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), volume 2: Short Papers*, pages 195–199, Gothenburg, Sweden. Association for Computational Linguistics.

Fiona J Tweedie and R Harald Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352.

Ying Zhao and Justin Zobel. 2005. Effective and scalable authorship attribution using function words. In *Proceedings of the Information Retrieval Technology: Second Asia Information Retrieval Symposium (AIRS)*, pages 174–189. Springer.

Ying Zhao and Justin Zobel. 2007. Searching with style: Authorship attribution in classic literature. In *Proceedings of the thirtieth Australasian conference on Computer science-Volume 62*, pages 59–68. Australian Computer Society, Inc.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393.