Developing a Benchmark for Reducing Data Bias in Authorship Attribution

Benjamin Murauer
Universität Innsbruck
b.murauer@posteo.de

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

Abstract

Authorship attribution is the task of assigning an unknown document to an author from a set of candidates. In the past, studies in this field use various evaluation datasets to demonstrate the effectiveness of preprocessing steps, features, and models. However, only a small fraction of works use more than one dataset to prove claims. In this paper, we present a collection of highly diverse authorship attribution datasets, which better generalizes evaluation results from authorship attribution research. Furthermore, we implement a wide variety of previously used machine learning models and show that many approaches show vastly different performances when applied to different datasets. We include pre-trained language models, for the first time testing them in this field in a systematic way. Finally, we propose a set of aggregated scores to evaluate different aspects of the dataset collection.

1 Introduction

In authorship attribution, various machine learning techniques are used to predict who has written a specific document, given a set of candidate authors. This means that a dataset used for such experiments must be well-controlled for many aspects like topic, length, etc. to ensure that the model detects the writing *style* of an author rather than something else like the topic of the content (Grieve, 2007; Stamatatos, 2009). For example, a model may find it easy to detect the author if each author writes about a single specific topic, and therefore the model would detect topic rather than style. Therefore, a well-controlled dataset should cover only one topic and one genre so that the only difference between the authors can be attributed to their writing style. Consequently, this makes the results of experiments using these well-controlled datasets prone to data bias, and they become difficult to generalize.

One approach to mitigate that bias is to distinguish the style of authors from the content, which allows to use cross-topic datasets. In this subset of tasks, the documents used for training the models are deliberately different from the texts used for validation thereafter. For example, given a set of journalists that write articles in multiple sections, when news articles about politics are used for training and articles about sports written by the same authors are used for testing, the overlap of topical content can be reduced and the model can only detect stylistic features. Similarly, cross-genre datasets take this one step further and require different genres of documents for training and testing (e.g., text messages and scientific essays). These datasets reduce the amount of stylistic information that can be used for each author to a subset that is expressive in both genres. In cross-language datasets, the training and testing data are written in different languages, further reducing this overlap.

Even when using cross-domain (topic, genre, etc.) datasets, the difficulty of generalizing assumptions regarding the authors' writing style remains, as any conclusions can only be stated for the concrete authors in that dataset. This may be sufficient for some applications, in which the style of *specific* authors or well-controlled groups of authors is analyzed. However, statements that claim to hold up more generally require evaluation with multiple and diverse datasets.

In this paper, we present a collection of datasets for authorship attribution which cover a wide variety of aspects, fulfilling these needs. We include datasets with few and many candidate authors, with different numbers of documents per author, with differently sized documents, and cross-topic, cross-genre and cross-language datasets. We provide detailed suggestions on train/test splits and perform evaluation experiments with a wide variety of models to demonstrate how much the choice of the dataset can impact classification results.

We provide exemplary attribution results for all datasets for a wide variety of machine learning

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 179–188

https://doi.org/10.26615/978-954-452-056-4_018

models. We specifically include several pre-trained language models, as they have shown great success in different NLP fields over the last years, but a systematic analysis of their performance in the authorship attribution field has not yet been performed to the best of our knowledge.

Lastly, to evaluate how well a model performs for each aspect of the collection, we provide a set of aggregated scores that combines results from different datasets.

Our contributions in this paper are therefore threefold: (1) we present a collection of selected, highly diverse datasets for authorship attribution that are able to better generalize evaluation results, (2) we benchmark several pre-trained language models on these datasets, providing a previously unavailable baseline in the field of authorship attribution, and (3) we provide a set of scores that evaluate a model based on the different aspects of the datasets.

To ensure reproducibility and foster future research, we publish all code online[1]. Thereby, we focus on providing tooling that minimizes the efforts required to expand both the dataset collection as well as the evaluation scores.

2 Related Work

While many previous studies use either only one dataset or don't specifically increase the diversity of the datasets used, they often fail to address this implicit data bias. Even foundational work in this field trying to categorize features in this field in a fundamental way can be prone to this issue. For example, Grieve (2007) measure the effectiveness of 39 different feature types for attribution. They address the importance of the dataset being representative for a language and explicitly explain the characteristics of the texts and authors in great detail, but consequently, by using a single dataset, their findings of feature performances are restricted to those very characteristics. Nevertheless, findings of such fundamental work are often referenced for research that uses completely different datasets.

One idea to mitigate any bias on the content of a dataset is to focus on the separation between *style* and *content*. This can be achieved by explicitly modelling the topic (Sari et al., 2018) or by using cross-topic or cross-domain datasets, where the training data and the test data have a different genre or contain texts about different topics (Stamatatos, 2013; Sapkota et al., 2015; Kestemont et al., 2018). For the latter, the key idea is that by minimizing the topic or genre-specific content contained in the overlap of training and testing data, any performances measured must conclude from the stylistic information from the authors. Nevertheless, for both approaches, the bias towards those authors remains in the evaluation.

Even from within a dataset, the choice of training and testing data can have a large impact on the outcome and additionally varies across languages (Eder and Rybicki, 2012). Additionally, Eder (2013) demonstrated that the amount of text required to reliably attribute an author also depends on the language, and suspects that this result may be depending on the genre of text as well. Similarly, Luyckx and Daelemans (2011) show that while some feature types are more robust to the size of the dataset, the performance of others varies greatly depending on the number of documents per author and the number of authors.

In this paper, we want to showcase a collection of diverse authorship attribution datasets and perform attribution experiments with several widely used exemplary machine learning models. The higher goal of our work is that it should be easy to make evaluation results of authorship attribution research easily comparable and also generalizable.

Using pre-trained language models for authorship attribution has not been researched in great detail. Some approaches use them as part of a larger ensemble model (Fabien et al., 2020) or as a feature extraction step in front of the actual classifier (Barlas and Stamatatos, 2020). However, when it comes to the performance of the unaltered models that are readily available, no overview for comparisons using widely used authorship attribution datasets are available to the best of our knowledge.

3 Datasets

For this benchmark, we have selected a wide variety of authorship attribution corpora. Thereby, we focussed on multiple aspects of datasets that may influence the classification process and try to provide a diverse but controlled set of these aspects:

- Genres: social media comments, business news, novels, reviews, etc.
- Topics: different news article topics, different fan fiction domains, etc.

[1] https://git.uibk.ac.at/csak8736/authbench

Dataset	Text type	$\times_T$	$\times_G$	$\times_L$	A	D	Words	$\|d\|$	D/A	Imb
CCAT50	financial/industrial news				50	2,500	1,254.4K	502	50	0.0
CL-Novels	prose	✓		✓	6	144	1,199.1K	8,354	24	11.7
CMCC	multiple	✓	✓		21	630	378.4K	601	30	0.0
Guardian	book reviews, opinions	✓	✓		13	264	276.0K	1,043	20	4.3
IMDb62	movie reviews				62	49,572	16,904.4K	341	800	1.0
PAN18-FF	prose	✓			20 (12)	88	69.7K	796	7	0.0
Reddit	social media comments			✓	45 (28)	2,366	1,259.7K	532	94	83.5

Table 1: Datasets used in this paper. $\times_{T,G,L}$ denote whether the datasets are cross-topic, cross-genre or cross-language, respectively. A denotes the total number of authors. The Reddit and PAN18-FF datasets have sub-problems with disjunct authors, and the number in parenthesis denotes the mean number of authors per sub-problem. D denotes the total number of documents. $|d|$ is the mean length of a document measured in number of words. D/A denotes the average number of documents available for training per author. Imb is the imbalance of the dataset, measured by the standard deviation of the number of documents per author.

- Languages: single-language datasets, multi-language datasets, cross-language datasets
- Dataset sizes: document size, number of documents per author, number of authors

These lead to the selection of seven datasets, which will be described briefly in the following section. An overview of some basic statistics is presented in Table 1.

3.1 Selected Datasets

The CCAT50 dataset (Liu, 2011) is a subset of the Reuters Corpus Volume 1 (Lewis et al., 2004) and contains 5,000 financial news articles from 50 authors, each having 50 training documents and 50 testing documents.

The CL-Novels dataset (Bogdanova and Lazaridou, 2014) contains English novels by 19th-century authors (Jane Austen, Charlotte Brontë, Lewis Carroll, Rudyard Kipling, Robert Louis Stevenson, and Oscar Wilde) and some Spanish (human) translations of their works. Although the novels are split into 500 sentence chunks (as the original authors did), these chunks are still the largest documents in this benchmark (cf. Table 1).

The CMCC dataset (Goldstein-Stewart et al., 2008) contains texts from 21 students about 6 different topics (church, gay marriage, privacy rights, legalization of marijuana, war in Iraq, gender discrimination) in 6 different genres (email, essay, interview transcript, blog article, chat, discussion transcript). This means that depending on how the data is split into train and test parts, it can function as either cross-topic or cross-genre dataset.

The Guardian dataset (Stamatatos, 2013) consists of book reviews and opinion articles written by professional journalists of *The Guardian* newspaper. The documents are categorized into the two genres of book reviews and opinion articles, and the latter is further divided into four topics (politics, society, world, UK). Hence, similar to the CMCC dataset, the choice of the train/test split defines whether this dataset is cross-topic or cross-genre.

The IMDb62 dataset (Seroussi et al., 2010) contains movie reviews written by 62 users of the internet movie database platform[2]. It features by far the most documents per author (1,000).

The PAN18-FF dataset (Kestemont et al., 2018) consists of fan fiction prose texts written by admirers of authors, novels, TV shows, movies, etc. Thereby, the authors invent and create new stories surrounding the original universes, which are called fandoms. The dataset contains authors that have written fiction in multiple fandoms, making it a cross-domain. Furthermore, the dataset is divided into 10 explicit sub-problems, 2 for each of 5 languages (English, Spanish, French, Italian, and Polish). For each problem, training and testing documents are predefined. Note that these problems are single-language problems and the authors don't overlap across different problems, which means that this is a multilingual, but not a *cross*-lingual dataset.

The Reddit dataset (Murauer and Specht, 2019) consists of comments by multilingual users of the Reddit social media platform. It contains five different language pairs for which users have written comments in both languages (English as well as one of German, Spanish, Portuguese, Dutch, and French). Compared to the other datasets, it is the

[2] www.imdb.com

Dataset	Splits	Description
CCAT50	2	predefined (50%/50%)
CL-Novels	$*15$	leave-one-novel out
CMCC$\times_G$	6	leave-one-genre-out
CMCC$\times_T$	6	leave-one-topic-out
Guardian$\times_G$	2	leave-one-genre-out
Guardian$\times_T$	4	leave-one-topic-out
IMDb62	5	stratified 5-fold
PAN18-FF	10	predefined
Reddit	10	leave-one-language-out

Table 2: Train/test splits for each dataset. $*$ Evaluations with identical training documents were combined.

most unbalanced dataset, as for some authors far more documents are available than for others (cf. column 'Imb' in Table 1).

3.2 Evaluation Splits

Deciding which parts of a dataset are used for training and testing plays an important role in interpreting the evaluation results, and being able to replicate results. In this section, we explain how these splits are selected for each dataset. Table 2 contains the overview of the train/test splits used in this paper.

The CCAT50 dataset has predefined subsets for training and testing of equal size.

The CL-Novels dataset is evaluated using *leave-one-novel-out*, as suggested by the original authors: Let D be the set of all novels, l_n the language of novel n, and t_n the original English title of both Spanish and English versions of n. Then, for $\forall n \in D$, the model is trained with all novels $\mathbf{m} = \{m \in D | l_m \neq l_n \wedge t_m \neq t_n\}$. The model is then evaluated on n. For example, for the split that has the English version of *Alice in Wonderland* as test data, m consists of training documents that (1) are not English, and (2) are not (a translated version of) *Alice in Wonderland*. Consequently, all n that only appear in one language have the same training documents $\mathbf{m}$. For these splits, the same model has to be trained only once and can be used for evaluation for all of the splits, increasing efficiency.

The CMCC and Guardian datasets contain multiple topics and genres. We adopt both a *leave-one-genre-out* as well as a *leave-one-topic-out* strategy, which is in line with related studies using these resources. Hence, in the experiments, these datasets are listed twice: once as a cross-genre dataset, and once as a cross-topic dataset.

For the homogeneous IMDb62 dataset, we use a stratified 5-fold cross-validation scheme.

The PAN18-FF dataset is divided into 10 sub-problems, each with a predefined training and testing part.

Finally, for the Reddit dataset, we use *leave-one-language-out* splits, where all documents of language l_1 are used for testing, and all documents of the respective other language l_2 are used for training. This is repeated for l_1 and l_2 swapped, and for each language pair (sub-problem) in the dataset.

3.3 Availability

For the IMDb62, PAN18-FF, CMCC, and Guardian datasets, permission to use them is required from the original authors. The CCAT50[3] and Reddit[4] datasets are freely available online. We reconstructed the CL-Novels dataset from the information of the original paper (Bogdanova and Lazaridou, 2014) by downloading the appropriate novels from the Project Gutenberg[5]. We removed introduction texts by the hosting platform (which always include the name of the author) and any appendices and notes from translators. The novels are in the public domain and we make the resulting cleaned dataset available for download online[6].

4 Experiment Setup and Models

The purpose of the evaluation experiments in this paper is to show that the performance of different models varies greatly across different datasets. We therefore perform classification experiments with several classification models to provide an impression of how the choice of a dataset influences the evaluation, but don't claim to provide the best possible configurations of those models for the analyzed datasets.

We select several features in combination with a linear support vector machine, as well as several solutions based on pre-trained language models. Important parameters for the models are listed in Table 3.

[3] https://archive.ics.uci.edu/ml/datasets/Reuter_50_50

[4] https://github.com/bmurauer/reddit_corpora

[5] https://www.gutenberg.org/

[6] https://git.uibk.ac.at/csak8736/authbench

Parameter	Value
TF/IDF	
normalization	$L2$
number of features	all
Doc2Vec	
learn rate	0.02
epochs	20
vector size	100
SVM	
C	1.0
penalty	$L2$
loss	squared hinge
multiclass strategy	one-versus-rest
max. iterations	1,000
tolerance	$1e^{-4}$
Language Model *	
BERT	bert-base-cased
DistilBERT	distilbert-base-cased
RoBERTa	roberta-base

Table 3: Parameters of the models used in the experiments. * Models from https://huggingface.co/models, accessed July 2021.

4.1 Features with Linear SVM

As a simple baseline, we use tf/idf-normalized frequencies of character 3-grams. They have been shown to be effective in authorship attribution and are capable of capturing both content-related information as well as author-specific stylistic nuances (Sapkota et al., 2015; Stamatatos, 2013, 2017).

We further adopt two syntax-based features. Firstly, we use part-of-speech (POS) tag n-grams. These abstract from the content of the text and focus on the grammatical structure, which has been shown to identify authors in similar settings (Kaster et al., 2005; Bogdanova and Lazaridou, 2014). Secondly, we utilize the DT-grams feature by Murauer and Specht (2021), which uses POS tags, but additionally incorporates dependency grammar information. *Universal* POS tags (Nivre et al., 2016) are a mapping of language-specific tags into a universal, language-independent space, and we utilize them for the experiments on the cross-language datasets. For both syntax-based features, we use language-specific POS tags for the mono-language datasets, and universal POS tags for the cross-language datasets.

Document embeddings have been shown to be effective for authorship attribution (Gómez-Adorno et al., 2018), and we experiment with both character 3-grams and words as tokens.

4.2 Pre-Trained Language Models

We test three transformer-based models: BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), and RoBERTa (Liu et al., 2019). Previous works use this family of models in combination with an ensemble (Fabien et al., 2020) or as a feature extraction stage for further processing (Barlas and Stamatatos, 2020), but no comprehensive analysis has been performed which uses these models without further modifications to the best of our knowledge. We use the parameters suggested by the respective original authors and use a sequence length of 256 tokens. As many documents are longer than that, we use a sliding window approach that extracts samples from the documents that fit into the maximum sequence length of the models.

5 Evaluation Results

For all models presented in Section 4, we perform authorship attribution experiments for all datasets covered in Section 3 (using the train/test splits as discussed in Section 3.2). In Table 4, the results of these classifications is shown, where each score (measured in macro-averaged F1) represents the mean score of the respective model and dataset for all train/test splits for that dataset. For example, the PAN18-FF dataset has 10 explicit subproblems, so each score in the PAN18 column of Table 4 represents the average score of the model for those 10 problems. As described in Section 3.2, the CMCC and Guardian datasets have two different ways of splitting the data (cross-topic and cross-genre).

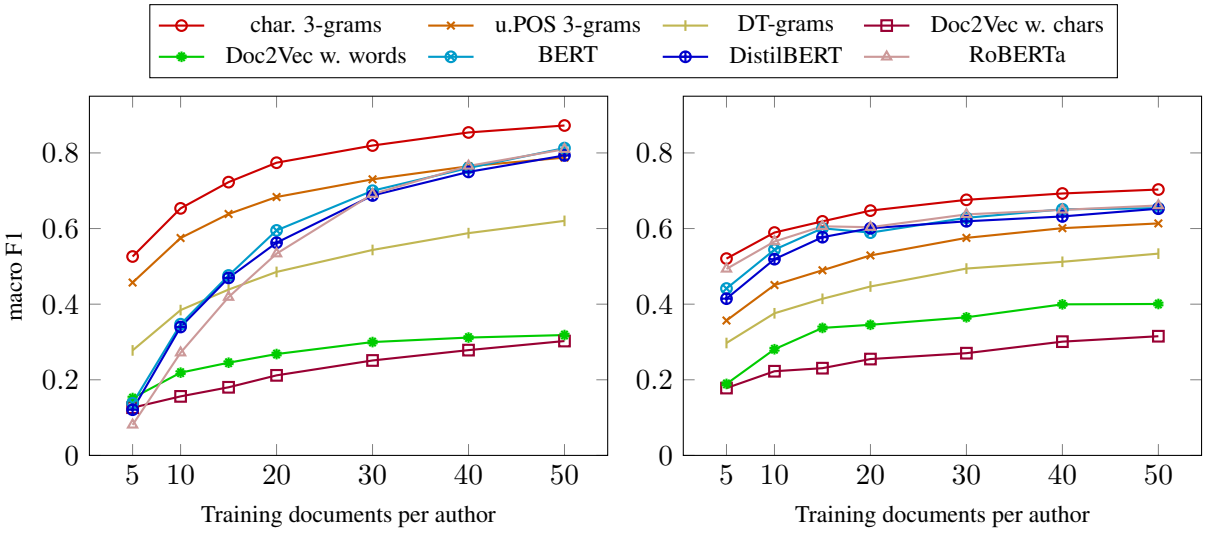
From these exemplary experiments, different conclusions can be drawn depending on which subsets of the results are analyzed. In the remainder of this section, we focus on the different aspects that the selected datasets feature.

5.1 Sensitivity to Dataset Size

The IMDb62 and CCAT50 datasets are large enough to extract differently sized subsets to analyze the performance of the models. While the Reddit dataset has comparably many documents, we refrained from including it in this experiment due to the imbalance of the dataset. Therefore, we restricted the number of training documents to 5, 10, 15, 20, 30, 40, and 50 documents for each

Model	CCAT50	CL-Novels	CMCC $\times_T$	CMCC $\times_G$	Guardian $\times_T$	Guardian $\times_G$	IMDb62	PAN18	Reddit
char. 3-grams	**0.703**	0.184	**0.614**	**0.685**	0.819	0.534	0.970	**0.493**	0.049
univ. POS 3-grams	0.614	**0.187**	0.544	0.452	0.813	**0.547**	0.922	0.381	0.131
DT-grams	0.534	0.178	0.394	0.295	0.648	0.305	0.874	0.281	0.166
Doc2Vec (char. 3-grams)	0.303	0.094	0.152	0.207	0.410	0.292	0.189	0.375	0.035
Doc2Vec (word 1-grams)	0.406	0.144	0.254	0.283	0.523	0.392	0.560	0.224	0.044
BERT	0.662	0.116	0.535	0.316	**0.847**	0.441	0.979	0.335	0.185
DistilBERT	0.665	0.107	0.482	0.293	0.814	0.442	0.976	0.391	0.126
RoBERTa	0.659	0.157	0.604	0.289	0.835	0.450	**0.979**	0.417	**0.268**

Table 4: Mean macro-averaged F1 scores of all splits for each dataset and tested model. $\times_G$ denotes cross-genre splitting, while $\times_T$ denotes cross-topic splitting of the CMCC and Guardian datasets.



(a) IMDb62

(b) CCAT50

Figure 1: F1 score of subsets of the IMDb62 (left) and CCAT50 (right) datasets with controlled number of documents per author. Note that although the number of authors and document sizes are comparable (cf. Table 1), the performances of the transformer-based models differ significantly, especially when few training samples are available for each author.

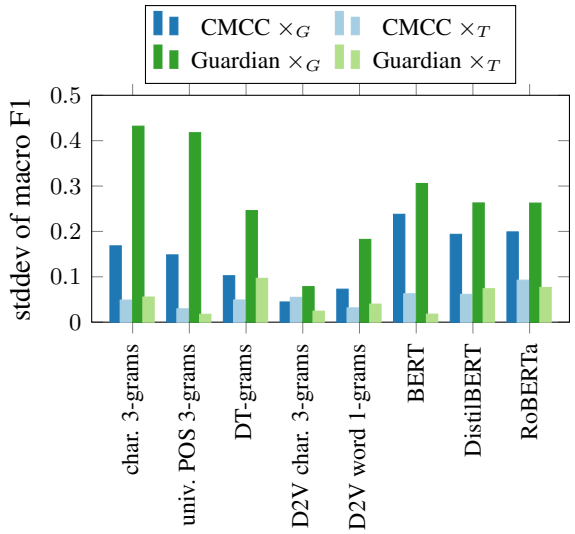


Figure 2: Sensitivity of tested models to cross-genre ($\times_G$) and cross-topic ($\times_T$) splits. The y-axis shows the standard deviation of the F1 score for all splits, high values indicate that the model performed well on some topics/genres and bad on others.

author, while not changing the number of test documents or the number of authors. The sampling of the documents was random, and all experiments were repeated 5 times to mitigate bias. In Figure 1, the F1 score of these sizes is displayed. It can be seen that while the performance for all models rises gradually in the CCAT50 dataset, the transformer-based models have much more trouble with the IMDb62 dataset when provided with fewer training samples, but quickly catch up to the character 3-grams with more training data.

While the reason for this discrepancy remains unanswered by this experiment, it shows that the two datasets exhibit different behaviors when used in combination with transformer-based classifiers. It is likely that other small-scaled datasets also display such incoherences, and it is therefore important to apply any model to multiple datasets to increase the meaningfulness of the evaluation.

In particular, studies along the lines of Luyckx and Daelemans (2011) analyzing comparable problems with varying dataset sizes should also be performed on as many datasets as possible.

5.2 Sensitivity to Genre and Topic

From Table 4, several conclusions can be drawn from the results of the cross-topic/genre datasets. Firstly, we can confirm that cross-genre classification is in general harder than cross-topic classification. Where explicit previous assumptions in this regard use single models and datasets (Stamatatos, 2013; Barlas and Stamatatos, 2020), we affirm this finding with multiple models and datasets. As a single exception, the character 3-grams show a higher performance on the cross-genre version of the CMCC dataset compared to the cross-topic variant.

The table also clearly reflects the difficulty that cross-genre situations impose on the pre-trained language models, which otherwise excel in the cross-topic splits.

Figure 2 shows the standard deviation of the F1 score across the different topics and genres in the CMCC and Guardian datasets. Hence, high values mean that the models perform differently for the topics or genres in the dataset. The figure displays that most models are more sensitive to the genre of the text than they are to the topic, consistently over both CMCC and Guardian datasets. The Doc2Vec model with character 3-grams has a low overall prediction score (cf. Table 4), and shows this effect to a smaller degree.

Note that this does not hold for the average performance over all splits (cf. Table 4): in general, the tested models are performing better on the cross-topic datasets, and do so more consistently for all topics compared to the cross-genre datasets. This result can't be seen from Table 4, and it means that for some cross-genre splits, some models may perform better than the average winner.

5.3 Sensitivity to Language

A surprising overall result for the cross-language datasets in Table 4 is the relatively high efficiency of the pre-trained language models for the Reddit dataset, as they have not been pre-trained using multilingual texts. This performance is not displayed in the other cross-language dataset containing 19th-century novels, which suggest that his behavior could stem from the genre of texts (social media comments), which are more likely to contain words common in multiple languages than documents from the 19th century. However, we suggest that even more datasets are required to answer this specific question.

Cross-language classification problems are defined by two different choices regarding the candidate languages: Firstly, which languages are considered in the classification problem at all, and secondly, which of those languages are used for training and which are used for testing. Figure 3

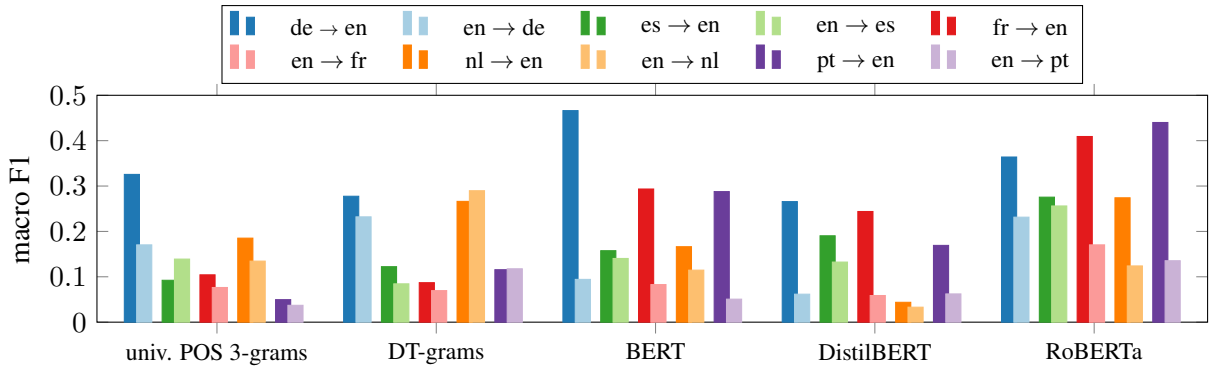


Figure 3: Sensitivity of selected models to the direction of the train/test split for each language pair. de → en denotes the score of the model that was trained with German and tested with English documents.

shows the macro F1 score of two cross-language models (univ. POS tag 3-grams and DT-grams) and the pre-trained language models for the Reddit dataset. The different colors represent the different language pairs of the Reddit dataset, and the two columns of each color represent the classification score (in macro F1) of both train/test directions used for the experiment (thereby, de → en denotes that the model was trained using German documents and tested on English texts).

The performance of the models generally differs across different pairs, which suggests that any cross-language classification approach should use as many language pairs as possible to generalize well. However, cross-language datasets are difficult to compile, as authors writing in more than one language are sparse.

In general, but especially for the language models, the figure also displays that the models perform better when they are fine-tuned using the non-English documents. This suggests that the choice of which language is used for training is an important choice that must be considered and reported by cross-language attribution studies.

6 Scores

The various datasets allow aggregation of the results according to the different aspects described in the previous sections, for which we formulate scores that are listed in Table 5. Each score is calculated by averaging the results from all splits in the respective datasets, weighted by the inverse number of splits in each dataset. Table 6 shows these scores for all models tested in our experiments, while Table 7 shows the standard deviation of each model across the different splits of the respective score. The aim of this separation is to quickly provide an overview of the strengths and weaknesses that a model shows for specifics aspects of the datasets. For example, for the models presented in this paper, it is now more clearly detectable that the character 3-gram features are a very strong baseline, but fail at the cross-language tasks.

Score	Description	Datasets Used
mono	one lang./topic/genre	IMDb62, CCAT50
sm	10 training texts/auth.	IMDb62, CCAT50
ml	mixed languages	PAN18
$\times_T$	cross-topic	CMCC$\times_T$, Guardian$\times_T$
$\times_G$	cross-genre	CMCC$\times_G$, Guardian$\times_G$
$\times_L$	cross-language	Reddit, CL-Novels
avg	mean of all	

Table 5: Scores used to reflect the models performance on the different aspects of the datasets.

The pre-trained language models show promising results for authorship attribution in summary, especially in the unexpected case of cross-language classification. Higher standard deviations indicate that these models are more prone to overfitting to specific fits.

We want to emphasize once more that the aim of this paper is not to provide the best possible results for the tested models, but show how a more expressive evaluation result can be achieved by incorporating multiple datasets into the evaluation process.

7 Limitations and Future Work

The collection of datasets presented in this paper is by no means exhaustive in terms of covered dataset aspects, but it should provide a solid foundation for this purpose. For example, authorship attribution on a larger scale (Narayanan et al., 2012; Tschug-

Model	mono	sm	ml	$\times_T$	$\times_G$	$\times_L$	avg
char. 3-grams	**0.84**	**0.62**	**0.49**	0.70	**0.65**	0.07	**0.56**
u.POS 3-grams	0.77	0.51	0.38	0.65	0.48	0.14	0.49
DT-grams	0.70	0.38	0.28	0.50	0.30	0.17	0.39
Doc2Vec char	0.25	0.19	0.38	0.26	0.23	0.04	0.22
Doc2Vec word	0.48	0.25	0.22	0.36	0.31	0.06	0.28
BERT	0.82	0.45	0.33	0.66	0.35	0.17	0.46
DistilBERT	0.82	0.43	0.39	0.61	0.33	0.12	0.45
RoBERTa	0.82	0.42	0.42	**0.70**	0.33	**0.25**	0.49

Table 6: Aggregated F1 scores reached by the models tested in our experiments.

Model	mono	sm	ml	$\times_T$	$\times_G$	$\times_L$	avg
char. 3-grams	0.19	0.05	0.11	0.12	0.23	0.06	0.13
u.POS 3-grams	0.22	0.09	**0.29**	0.14	0.21	0.09	0.17
DT-grams	**0.24**	0.01	0.16	0.15	0.13	0.08	0.13
Doc2Vec char	0.08	0.05	0.23	0.14	0.06	0.04	0.10
Doc2Vec word	0.11	0.04	0.10	0.14	0.11	0.05	0.09
BERT	0.22	0.14	0.19	0.17	**0.24**	**0.13**	**0.18**
DistilBERT	0.22	0.13	0.19	**0.18**	0.20	0.09	0.17
RoBERTa	0.23	**0.21**	0.13	0.14	0.21	0.11	0.17

Table 7: Aggregated standard deviations of F1 scores reached by the models tested in our experiments.

gnall et al., 2019) requires datasets far beyond the sizes of the presented material, and in general, also requires different methods for evaluation and solving strategies.

From a multilingual standpoint, the dataset collection thus far only contains several European languages, and those are among the smallest datasets in the benchmark. In the long term, our future plans involve including more datasets from as many languages as possible, and ideally also increase the number of cross-language datasets.

As not all models and methods are intended to work with all types of text, we envision well-defined subsets of the benchmark covering the possible application areas for many models. For example, even when a model is only targeted to classify social media text, we aim to provide multiple datasets fulfilling this requirement.

To ensure the continued attribution to this collection, we publish a set of tools[7] which minimize the effort required to add additional datasets to this collection. These tools make it easy to (1) bring the dataset to a common format, (2) define train/test splits that the dataset should be used with, and (3) specify which of these splits contribute to the scores presented in the previous section. Thereby, contributions can be made to existing scores by providing more datasets to reassure them, or add additional scores to the collection. We hope to timely contribute more multilingual datasets to the ml score and expand on different dataset sizes beyond the few thresholds presented by the sm score.

[7] https://git.uibk.ac.at/csak8736/authbench

8 Conclusion

In this paper, we present a collection of datasets aimed to increase the expressiveness and generalizability of authorship attribution experiments. The datasets are carefully chosen to include many different aspects of the text, such as document size, number of documents per author, language, genre, or topic. We choose several well-established text classification models and perform attribution experiments on all datasets, for the first time showing results systematically for pre-trained language models in this field. Thereby, we demonstrate the importance of including multiple datasets in any evaluation by showing differences in the classification score for similar datasets and train/test splits. We conclude the paper by suggesting an aggregated score for each of the presented aspects to easily distinguish the strengths and weaknesses of different models.

References

Georgios Barlas and Efstathios Stamatatos. 2020. Cross-domain authorship attribution using pre-trained language models. In *IFIP Advances in Information and Communication Technology*, pages 255–266.

Dasha Bogdanova and Angeliki Lazaridou. 2014. Cross-language authorship attribution. In *Ninth International Conference on Language Resources and Evaluation (LREC'2014)*, pages 2015–2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Maciej Eder. 2013. Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2):167–182.

Maciej Eder and Jan Rybicki. 2012. Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Literary and Linguistic Computing*, 28(2):229–236.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA: BERT fine-tuning for Authorship Attribution. In *Proceedings of the*

17th International Conference on Natural Language Processing.

Jade Goldstein-Stewart, Kerri Goodwin, Roberta Sabin, and Ransom Winder. 2008. Creating and using a correlated corpus to glean communicative commonalities. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA).

Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov, and David Pinto. 2018. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, 100(7):741–756.

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

Andreas Kaster, Stefan Siersdorfer, and Gerhard Weikum. 2005. Combining Text and Linguistic Document Representations for Authorship Attribution. In *Working Notes of the 28th Conference on Research and Development in Information Retrieval (SIGIR'2005): Stylistic Analysis of Text for Information Access*, pages 27–35.

Mike Kestemont, Michael Tschugnall, 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*.

David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

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.

Zhi Liu. 2011. Reuter 50-50 Dataset. National Engineering Research Center for E-Learning Technology China.

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.

Benjamin Murauer and Günther Specht. 2019. Generating cross-domain text classification corpora from social media comments. In *Proceedings of the 20th International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF'2019)*, pages 114–125.

Benjamin Murauer and Günther Specht. 2021. DTgrams: Structured Dependency Grammar Stylometry for Cross-Language Authorship Attribution.

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 *2012 IEEE Symposium on Security and Privacy*. IEEE.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th Int. Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

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.

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. Association for Computational Linguistics.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2010. Collaborative Inference of Sentiments from Texts. In *Proceedings of the 18th International Conference on User Modeling, Adaptation and Personalization (UMOD'2010)*, pages 195–206.

E. 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 & Policy*, pages 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 (EACL'2017)*, pages 1138–1149. Association for Computational Linguistics.

Michael Tschuggnall, Benjamin Murauer, and Günther Specht. 2019. Reduce & attribute: Two-step authorship attribution for large-scale problems. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 951–960. Association for Computational Linguistics.