


Please cite the Published Version

Silva, Kanishka, Can, Burcu, Blain, Frédéric, Sarwar, Raheem , Ugolini, Laura and Mitkov, Ruslan (2023) Authorship attribution of late 19th century novels using GAN-BERT. In: 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop, 10 July 2023 - 12 July 2023, Toronto, Canada.

DOI: <https://doi.org/10.18653/v1/2023.acl-srw.44>

Publisher: Association for Computational Linguistics

Version: Published Version

Downloaded from: <https://e-space.mmu.ac.uk/632634/>

Usage rights:  [Creative Commons: Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)

Additional Information: This is an Open Access paper originally presented at: 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop

Enquiries:

If you have questions about this document, contact rsl@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

Authorship Attribution of Late 19th Century Novels using GAN-BERT

Kanishka Silva

University of Wolverhampton
United Kingdom
a.k.silva@wlv.ac.uk

Burcu Can

University of Stirling
United Kingdom
burcu.can@stir.ac.uk

Frédéric Blain

Tilburg University
The Netherlands
f.l.g.blain@tilburguniversity.edu

Raheem Sarwar

Manchester Metropolitan University
United Kingdom
r.sarwar@mmu.ac.uk

Laura Ugolini

University of Wolverhampton
United Kingdom
l.ugolini@wlv.ac.uk

Ruslan Mitkov

Lancaster University
United Kingdom
ruslanmitkov@gmail.com

Abstract

Authorship attribution aims to identify the author of an anonymous text. The task becomes even more worthwhile when it comes to literary works. For example, pen names were commonly used by female authors in the 19th century resulting in some literary works being incorrectly attributed or claimed. With this motivation, we collated a dataset of late 19th-century novels in English. Due to the imbalance in the dataset and the unavailability of enough data per author, we employed the GAN-BERT model along with data sampling strategies to fine-tune a transformer-based model for authorship attribution. Differently from the earlier studies on the GAN-BERT model, we conducted transfer learning on comparatively smaller author subsets to train more focused author-specific models yielding performance over 0.88 accuracy and F1 scores. Furthermore, we observed that increasing the sample size has a negative impact on the model's performance. Our research mainly contributes to the ongoing authorship attribution research using GAN-BERT architecture, especially in attributing disputed novelists in the late 19th century.

1 Introduction

Authorship attribution identifies authors of a given set of unknown documents (Hu et al., 2020; Neal et al., 2018; Stamatatos, 2009). Conventional techniques and neural networks are the two main authorship attribution methods. The studies on the conventional approaches typically focus on feature engineering and stylometry. The deep learning approaches have been gaining popularity recently due to the superior results compared to the conventional

approaches. Furthermore, authorship attribution can be tackled in two ways: closed-set and open-set attribution. In closed-set attribution, an author is selected from a set of candidate authors, whereas in open-set attribution, the target author may not be included in the candidate authors' list.

Applications of authorship attribution are employed in various domains, such as digital forensics (Abbasi and Chen, 2005; Sun et al., 2012), social media analysis (Junior et al., 2016; Duman et al., 2016; Brocardo et al., 2017) and digital humanities Juola (2021). In historical texts, the authorship styles may contain socio-linguistic characteristics due to the century in which the author lived, idea movements inspired by the author, and language-specific attributes. Also, in written texts, the genre and topics are crucial in defining the author's style. Several pieces of research have been undertaken in the literature and historical domains, for instance, identifying anonymous or disputed texts (Koppel et al., 2007; Kestemont et al., 2016; Tuccinardi, 2017). The work presented by Fung (2003) analyses the Federalist Papers, which involves 85 articles and essays written by Alexander Hamilton, James Madison and John Jay. Another application of authorship attribution in literature is resolving doubted authorships. For instance, Thompson and Rasp (2016) investigate whether C.S. Lewis wrote *The Dark Towers*. The Shakespearean Authorship Dispute was addressed by Fox and Ehmoda (2012). Furthermore, attributing the author is one of many variations in authorship applications, as research directions are in different domains, such as attributing to the publication year and identifying the literary genre and the topic. One such example is

Tausz (2011) which predicts the date of authorship in historical texts.

This research proposes a GAN-BERT-based model to enhance transformer-based authorship attribution in late 19th-century novels. To our knowledge, this is the first attempt to ensemble GAN and BERT models and, precisely, the GAN-BERT model to address authorship attribution in literary texts. In some of the recent works on authorship attribution, the models were trained in a controlled setting and had less elaboration on the data preparation stage, resulting in the poor reproducibility and generalisation of these models. Here, we present an end-to-end process from domain selection to dataset collection with insights to experiment planning.

An authorship attribution model highly depends on the number of authors represented in the training dataset and the text available per each author. Most of the related works emphasise controlled training environments. To improve the model’s generalisation and ability to perform well on robust scenarios, it should be identified how much the model depends on the number of authors in the training dataset and the amount of text by each author. We use a normalised dataset of 20 novels per author to avoid dataset imbalance. Therefore, to identify how much data provides better model performance, we control the text data sample size drawn from the book text. Therefore, the research questions in this study are as follows:

RQ 1: How to effectively utilise the GAN-BERT model for authorship attribution?

RQ 2: How does the number of authors in the dataset impact the GAN-BERT performance for authorship attribution?

RQ 3: How does the amount of text data (i.e. sample size) drawn from each novel affect the GAN-BERT performance for authorship attribution?

The remainder of the paper is organised into several sections: Section 2 demonstrates a brief literature survey. Then Section 3 describes the proposed model’s architecture, and Section 4 presents the dataset collection and preparation. Section 5 elaborates on the experiment design, focusing on the research questions, Section 6 summarises the results and findings obtained, and finally, Section 7 involves the concluding remarks and future directions.

2 Related Work

Texts vary in terms of topic, sentiment and style. According to Stamatatos (2009), information about the authors can be extracted from the style of their written documents. The task involves identifying the author from unknown documents, known as authorship attribution, which breaks into two major tasks: Authorship Identification and Authorship Verification. Authorship Identification is identifying a document’s author by comparing a set of candidate authors (Stamatatos, 2009). Authorship Identification can be interpreted as a binary classification problem, whereas authorship attribution is a multi-class classification problem. Authorship Verification is a fundamental problem in authorship attribution which focuses on finding whether the considered person wrote one or more documents or not. Authorship Verification is comparatively challenging with less data (Koppel et al., 2011; Luyckx and Daelemans, 2008).

With the popularity of deep neural networks for NLP applications, recent authorship attribution research shares a similar trend. The works of Bagnall (2015a); Hosseinia and Mukherjee (2018); Boumber et al. (2018) are examples of neural network-based models in authorship attribution. Additionally, transfer learning also proved to have astonishing results. Zhang et al. (2021) introduce a Deep Authorship Verification using new metrics: DV-distance and DV-projection, which utilise pre-trained language models. Their work highlights the utilisation of pre-trained language models in our approach. Character and n-gram-based CNN (Ruder et al., 2016), Syntax-augmented CNN (Zhang et al., 2018), and Convolutional Siamese Networks (Saedi and Dras, 2021) are some other authorship attribution models which utilise deep learning techniques. These deep learning-based applications provide valuable insights for our approach to utilising the GAN-BERT model for authorship attribution tasks.

Language Models (LM) used in the authorship tasks can be categorised as n-gram-based and neural network-based (Fourkoti et al., 2019). Ge et al. (2016) used a neural network-based language model. The works of Bagnall (2015b) present a character-level RNN-based LM combining a multi-headed classifier. To address the cross-domain problem, Barlas and Stamatatos (2020) extended Bagnall (2015b)’s works for closed-set authorship attribution by combining a multi-headed LM with

a pre-trained LM. According to [Barlas and Stamatatos \(2020\)](#), having a normalised corpus is crucial for the performance of cross-domain authorship attribution. BertAA ([Fabien et al., 2020](#)) is the recent fine-tuned form of the pre-trained BERT model for the authorship attribution task, which presents extensive experiments on various datasets: Enron Email ([Klimt and Yang, 2004](#)), Blog Authorship ([Schler et al., 2006](#)) and IMDb ([Seroussi et al., 2014](#)). Although pre-trained models have gained popularity and promising results in some authorship tasks, the performance of such models highly depends on the training set.

Generative Adversarial Networks (GAN) are used in authorship-related tasks to prevent adversarial attacks, mainly in the Authorship Obfuscation problem where one’s writing style is masked. [Ou et al. \(2022\)](#) introduce source code authorship verification using GAN models and multi-head attention. A^4NT ([Shetty et al., 2018](#)) is a GAN-based style transformation to perform authorship obfuscation learned from data via adversarial training and sequence-to-sequence LMs. [Kazlouski \(2019\)](#) presents an LSTM-GAN classifier to recognise imitations generated by the A^4NT ([Shetty et al., 2018](#)) model. [Tang et al. \(2019\)](#) presents a data augmentation approach to authorship attribution in Weibo text using Wasserstein-GAN to generate samples of the positive class.

The class imbalance problem is hard to avoid in real-world scenarios, particularly in authorship attribution. [Stamatatos \(2018\)](#) introduced a novel strategy to produce synthetic data for the authorship identification task. The approach that [Stamatatos \(2018\)](#) mentioned is segmenting the training texts into text samples, considering the training size of the class. The works of [Eder \(2015\)](#) highlight how much data is required to identify authors across different languages and genres. The findings in [Eder \(2015\)](#) show that the minimum sample range is 2500-5000, representing the two ends for Latin, English, German, Polish, and Hungarian datasets. Further experiments by [Eder \(2017\)](#) attempt to identify the minimum sample size by removing text one by one from the training set, which yields that 2000 words sample size is appropriate. Also, [Eder \(2017\)](#) emphasises that this finding depends strongly on the authors. [Hadjadj and Sayoud \(2021\)](#) propose a hybrid PCA and SMOTE approach of oversampling, which reports outperforming the state-of-the-art accuracies. The Stylometric Set Similarity (S3)

method presents the authorship attribution task as a set similarity problem by considering 3000 novels from 500 authors curated from Project Gutenberg ([Sarwar et al., 2018](#)). [Granichin et al. \(2015\)](#) present a KNN-resampling approach to authorship identification by simulating samples from 2 texts.

In previous research on authorship attribution, the combination of GAN and transformer models has not yet been explored. Furthermore, to the best of our knowledge, no attempt has been made to use the GAN-BERT model specifically for the task of authorship attribution, especially with sampling strategies for many authors and limited data. The critical literature analysis suggests that deep neural networks in authorship attribution would show promising performance with well-designed sampling strategies. Here, we propose GAN-BERT model for authorship attribution along with various sampling strategies, and analyse how transfer-learning would support the proposed model in literary domain.

3 GAN-BERT Model for Authorship Attribution

Let A be a collection of authors of interest, $A = \{a_1, a_2, \dots, a_N\}$, where N is the total number of authors in A . The document set belonging to each author forms the complete dataset $T = \{t_{a_1}, t_{a_2}, \dots, t_{a_N}\}$ where t_{a_i} is the document set attributed to the author a_i in the dataset. Given a text, t_u of an unknown author u , the proposed model assigns the text to the most likely author from A .

GAN-BERT ([Croce et al., 2020](#)) combines BERT-based models and Semi-Supervised GAN ([Salimans et al., 2016](#)). Figure 1a illustrates the GAN-BERT model architecture, where discriminator D is utilised to classify examples and generator G generates fake examples F . The discriminator takes the vector representations returned via BERT for unlabeled U and labelled L input texts. When training is complete, G is discarded from the model to use the rest of the model for inference.

In contrast to GAN-BERT ([Croce et al., 2020](#)), which utilises a semi-supervised GAN model ([Salimans et al., 2016](#)) with labelled and unlabeled data, we train the GAN-BERT model with labelled data only. The discriminator D is trained over $N+1$ classes to assign the true samples to a class from $\{1, 2, 3, \dots, N\}$. The fake sample generated from the generator G represents the $(N+1)^{th}$ class. The discriminator is suitable for detecting authorship

obfuscation and forgery since it is trained with fake samples similar to the original author-written texts. Figure 1b illustrates the modified GAN model.

The GAN-BERT model generally shows superior results for classification tasks with limited labelled data. Furthermore, the intuition to use GAN-BERT for authorship attribution is that, due to the fake data generated in the generator, it considers not only the real writing styles, but also the possible fake writing styles that are synthesised.

4 Creating the Datasets

4.1 Pre-Screening Authors

We performed pre-screening on the authors before collecting the dataset, which is, to the best of our knowledge, the first attempt to perform a qualitative analysis on the literary domain for authorship attribution. We considered two parameters during the author selection process: distribution and filtering. Distribution parameters ensure that the collected texts span equally among different attributes such as gender, genre and ethnicity. Filtering parameters focus on whether selected works by the distribution parameters should be included or excluded from the dataset. It mainly concerns the novelists' characteristics and the nature of their literary works. A summary of these two parameters is illustrated in Table 1.

4.2 Dataset Collection and Validation

We collected datasets from Project Gutenberg across genres such as novels, short stories, essays, poems and biographies. There is no specific field in Project Gutenberg to indicate genre and year of publication. We manually validated texts to capture the year of publication. We also filtered novels so that all fiction had a word count greater than 10,000. To our knowledge, other researchers using Project Gutenberg have not performed similar data validation to filter novels.

In the master dataset, we have filtered 1232 novels written by 62 authors, which are segmented as follows:

1. Early 19th Century (1800-1835)
2. Mid-19th Century (1836-1870)
3. Late 19th Century (1871-1900)
4. Early 20th Century (1901-1914)

This paper focuses on the late 19th-century segment from the master dataset, which includes 541 novels. We filtered authors based on the number of novels available in the dataset and selected those with at least 20. We narrowed the author selection by selecting the top 20 authors with the most novels from this focused subset. These authors were used to train and test the proposed GAN-BERT model. Therefore the dataset is thus uniformly distributed regarding the number of novels per author. The selected authors are Anthony Trollope, Arthur Conan Doyle, Bret Harte, Fergus Hume, Frances Hodgson Burnett, H.G. Wells, Henry Rider Haggard, Jack London, James Grant, John Kendrick Bangs, Joseph Conrad, Louisa May Alcott, Margaret Oliphant, Marie Corelli, Mark Twain, Mary Elizabeth Braddon, Mrs Henry Wood, Nathaniel Hawthorne, Oliver Optic, and Wilkie Collins.

4.3 Balanced Author Representation

The filtered dataset of late 19th century English novels consists of 400 novels by 20 authors. Especially in deep neural networks, this dataset is insufficient to represent a larger number of authors than 20. Furthermore, as authors have different writing styles, different combinations of authors in the same size dataset have a strong impact on model performance. We observed this problem during the preliminary experiments with manually sampled sets of authors. Therefore, to ensure a balanced representation of authors in the training and validation datasets and to mitigate the effect of different author combinations, we performed random sampling for a considered number, as shown in Figure 2. Different author combinations are denoted by a 'sample set'.

Furthermore, one of the aims of the experiments is to see how increasing the number of authors would affect the model's performance. To do this, we split the dataset to represent different numbers of authors.

4.4 Dataset Splits

We followed the leave-n-out method to split the dataset for manually selected 5 sets. For example, of 20 authors, two were assigned as a 2-author case, while the rest of the 18 were included as an 18-author case. This process is repeated to obtain distinct 5 manually selected author sample sets. The author's case defines how many authors were considered in the train/test datasets. For example, a 2-author case means a focused dataset with only

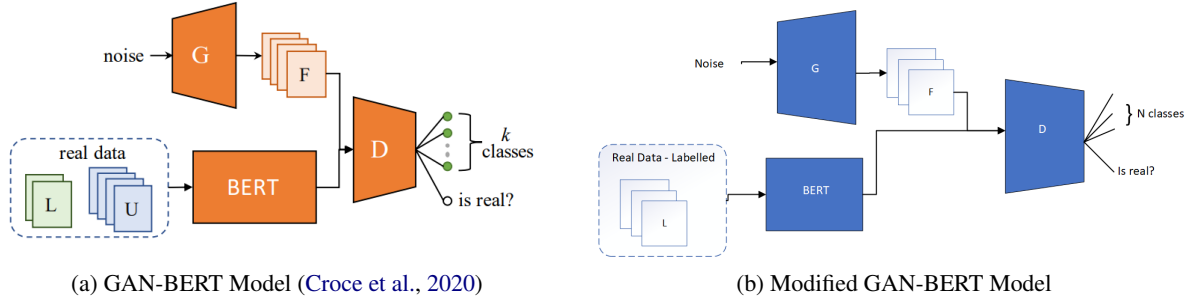


Figure 1: Model Architecture Comparison

Parameter Type	Category	Condition
Distribution Parameters	Genre	Romance, Thrillers, Science Fiction, Realist
	Gender	Male, Female
	Ethnicity	American, British
	Doubted Authorship	Only original works by novelist in the training set
	Readers	Adult, Children
Filtering Parameters	Publication Period	Later 19th Century 1871-1900
	Number of novels during publication period	>3
	Literature Genre	Novels
	The number of total novels	>20
	Written Language	English
	Non-translation	Yes
	Multi-Authors	No
	Digitised work availability	Available on the Project Gutenberg

Table 1: Distribution and Filtering Parameters used for Pre-Screening of Authors

novels by 2 authors. We can define any number of author sample sets to perform experiments in each n-author case. For example, manually selected author sample sets for a 2-author case include 5 different combinations of 2 authors out of 20 can be present. 50 random samples in a 2-authors case mean, out of 20 authors, 50 randomised different 2-author combinations. Random sampling does not cover all combinations of authors in a given author case, but would ensure that the majority of author combinations are considered. The dataset splitting process is illustrated in Figure 2.

We ensured the dataset splits were distinct for all the sample sets per case. The 20-author case was used as the base model to train and perform transfer learning on other models. We used a randomised approach to shuffle and return 50 and 100-author sample sets for a random sample generation.

We split train-test-validation (80:10:10) sets, stratified by author ids, for each sample set considered for the experiments, with one sample set per experimental round. The average results of all sample sets represent a particular n-author case. The base model was trained on all 20 authors in the transfer learning experiments. The stratified split in the train-test-validation ensured a uniform distribution of novels per author, and the test data are

distinct from the training data. In transfer learning, the training set may include evaluation data from the 20-author case.

4.5 Baseline Datasets

To compare the performance of the proposed GAN-BERT model on other baseline datasets, we used the IMDB62 (Seroussi et al., 2014) and Blog Authorship (Schler et al., 2006) datasets. We created a subset of 20 authored content from these datasets to be consistent with the 20-author dataset, which refers to as IMDB20 and Blog20 respectively.

4.6 Dataset Availability

Due to the copyright restrictions explained in Section 7, we do not release the entire dataset. Instead, we release the scripts used for creating and pre-processing the dataset. We also publish the list of the authors, selected novels, and novel indices used to extract the sample sets ¹.

5 Experiment Design

We conducted experiments on different dataset subsets and different model configurations to address the following:

¹<https://github.com/Kaniz92/AA-GAN-Bert/tree/main>

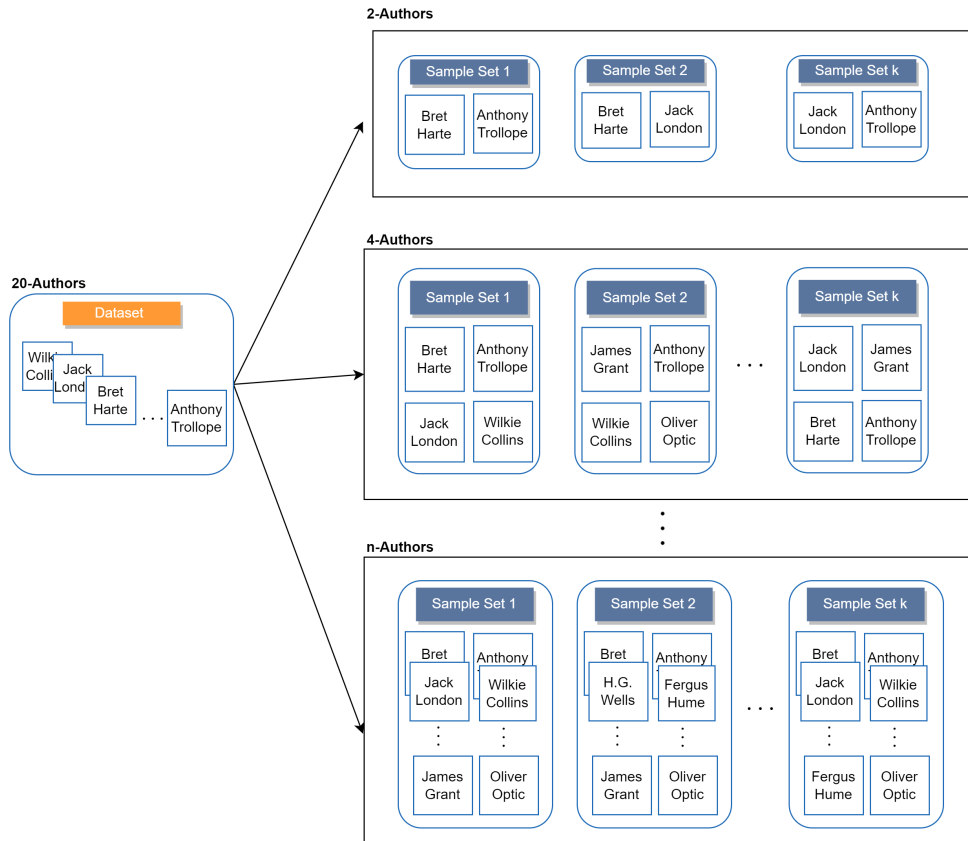


Figure 2: Dataset Splitting Process

1. Random Sampling Author Combinations
2. The Impact of Transfer Learning
3. Number of Authors in Dataset
4. Text Sample Size per Novel

We explored the GAN-BERT model under two dimensions: Random Sampling and Transfer Learning. As illustrated in Figure 2, the 20 novels per each author from the 20-author dataset provide different combinations under different numbers of authors. Therefore, first, we manually selected authors per each n -author case and then randomly sampled 50 and 100 author combinations. In transfer learning experiments, we compared the performance of manually selected sample sets under standalone training and transfer learning from the 20-author dataset to each n -author case.

In a practical scenario of authorship attribution, the number of authors to compare would vary. Therefore, we experimented with the GAN-BERT model response for different numbers of authors in the dataset. Also, the text sample size drawn from a novel can be varied when representing the novel text due to varying text lengths. We used the

manual sampling of authors to identify any trend towards the text sample size drawn from a novel.

In the default setting, unless specified, we used 20 samples per novel drawn sequentially from the book text for training and testing. We first trained the base model on 20-authors for 10 epochs, using Adam optimiser, one hidden layer for both generator and the discriminator, a dropout rate of 0.2, batch size of 8, a warm-up proportion of 0.1, and learning rate of $1e-5$ for both generator and the discriminator. Then the pre-trained 20-author model was used for transfer learning on smaller subsets of each case in $\{2, 4, 6, 8, 10, 12, 14, 16, 18\}$ -author counts and trained further on these sub-sets for 5 epochs.

We compared the proposed GAN-BERT model with different baseline models such as word-level TF-IDF, character n -gram, Stylometric features (Sari et al., 2018) and BertAA (Fabien et al., 2020) on the 20-authors dataset, 18-authors dataset, IMDB, and Blog Authorship datasets. These baseline experiments provide insights into how the created datasets performed with other baseline models and how other datasets would perform with the proposed GAN-BERT model. To be consistent with

the rest of the experiments, we selected 20 samples per each document by an author, but the 20-sample restrictions are not applied to baseline models.

6 Results and Discussion

For each experiment across different sample sets, we reported Accuracy, F1, Precision, and Recall with averaging results sampled manually and randomly.

6.1 Random Sampling Author Combinations

Analysing the model with manually selected author sample sets may fail to describe the results and any trends due to the bias factors. For example, the up-shot performance of the 18-authors model in manually sampled authors as in Figure 3a could be due to biases in generated manual sample sets. Therefore we conducted additional experiments for the 50 and 100 sample sets using random sampling. Rather than selecting books randomly, we focused on arranging authors into different sample sets and then keeping books per each author the same (20 books per author). This experiment explores whether the model could tolerate the robustness of any author combinations. Before deciding on the random sampling limits, we analysed the maximum number of author combinations per each case. To cover all the author cases, the maximum random sampling count is 190, so we decided to experiment on 50 and 100 random samples.

Compared to the manually selected author sample sets, 50 and 100 random sampling achieves a higher accuracy for all the author cases, precisely more than 0.97% of accuracy. Results in Table 2 and Figure 3b show that the model is robust with consistent performance over different author cases.

6.2 The Impact of Transfer Learning

The intuition behind applying transfer learning for the authorship attribution model is that instead of having a model that learns each author’s style and overfits into a particular dataset with a fixed number of authors, it makes the model more practical to use in real-world scenarios if the model learns the authorship attribution task regardless of the number of authors. This also applies to different author styles, regardless of topic, genre or unique author style. Moreover, transfer learning allows the model to transfer knowledge into a limited data set.

Extensive experiments have been carried out to identify how transfer learning has affected the

model’s performance from the 20-author cases to smaller author subsets. We trained standalone and transfer learning models using the same hyperparameters as the base model.

Transfer learning has substantially improved the model’s performance, especially for the increasing number of authors. The best-performing model was observed for the 2-author case, and the worst-performing model was for the 18-author case. Overall, the transfer learning results suggest that it is a promising technique for improving performance, especially for smaller datasets.

6.3 Incremental Number of Authors in the Dataset

We designed the dataset subsets to increment the number of authors by two, ranging from [2, 18], to investigate how the author count would affect the model’s performance. The number of samples per author is uniform across each author sample set and case. We also selected the same 20 books for each author to ensure that the topics or genres do not affect the experiments. One text sample should not exceed 512 words, BERT’s maximum input token size. Therefore we set the one sample size as 512 words and drew 20 sequential text samples from each book, representing one author by 400 (20 x 20) instances before the train-test split.

Both the standalone and transfer learning models for five manually selected author sample sets show a declining trend in performance as the number of authors increases, as illustrated in Table 3 and Table 2. The binary classification shows the best performance overall, while the multi-class classification shows comparatively a lower performance.

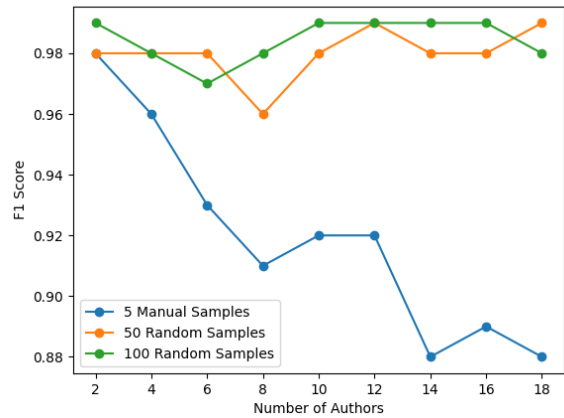
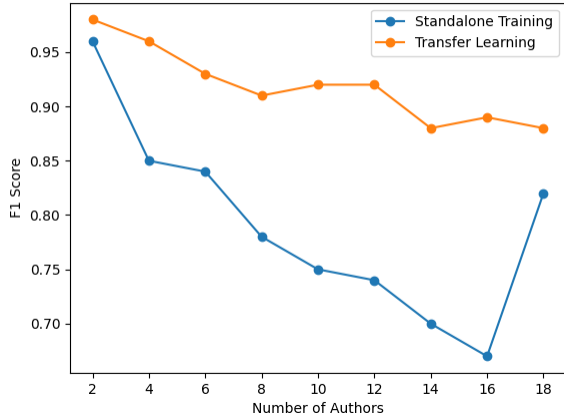
Averaging accuracies for transfer learning for 50 and 100 randomly sampled author sets are illustrated in Table 2. The results do not indicate any clear trend with the author counts, but accuracy and F1 are consistent and higher than manually selected author sample sets.

As illustrated in Figure 3b, manual samples and random samples show clear distinction with increasing the number of authors in the dataset. Therefore, the model performance depends highly on how the sample sets were defined, i.e. different author combinations. Therefore, strategies must be explored to overcome the biases towards different configurations of authors’ sample sets.

n-Authors	5 Manual Samples				50 Random Samples				100 Random Samples			
	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall
2-authors	0.98 [†]	0.98 [†]	0.99 [†]	0.98 [†]	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.98
4-authors	0.96	0.96	0.96	0.96	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
6-authors	0.93	0.93	0.93	0.93	0.98	0.98	0.98	0.98	0.97*	0.97*	0.97*	0.97*
8-authors	0.91	0.91	0.92	0.91	0.96*	0.96*	0.97*	0.96*	0.98	0.98	0.98	0.98
10-authors	0.92	0.92	0.92	0.92	0.98	0.98	0.98	0.98	0.99 [†]	0.99 [†]	0.99 [†]	0.99 [†]
12-authors	0.92	0.92	0.93	0.92	0.99 [†]	0.99 [†]	0.99 [†]	0.99 [†]	0.99 [†]	0.99 [†]	0.99 [†]	0.99 [†]
14-authors	0.88*	0.88*	0.90*	0.88*	0.98	0.98	0.98	0.98	0.99 [†]	0.99 [†]	0.99 [†]	0.99 [†]
16-authors	0.89	0.89	0.90*	0.89	0.98	0.98	0.98	0.98	0.99 [†]	0.99 [†]	0.99 [†]	0.99 [†]
18-authors	0.88*	0.88*	0.90*	0.88*	0.99 [†]	0.99 [†]	0.99 [†]	0.99 [†]	0.98	0.98	0.98	0.98

Table 2: Results of the GAN-BERT Model for Transfer Learning on a 20-Author Dataset

* - mini result across a metric [†] - max value across a metric



(a) F1 Scores between Standalone Training and Transfer Learning

(b) F1 Scores between Manual Sampling and Random Sampling for Transfer Learning

Figure 3: F1 Score Results of the Transfer Learning Approach

n-Authors	Accuracy	F1	Precision	Recall
2-authors	0.95 [†]	0.96 [†]	0.95 [†]	0.95 [†]
4-authors	0.82	0.85	0.82	0.82
6-authors	0.82	0.84	0.82	0.83
8-authors	0.76	0.78	0.76	0.75
10-authors	0.72	0.75	0.72	0.72
12-authors	0.70	0.74	0.70	0.70
14-authors	0.66	0.70	0.66	0.66
16-authors	0.64*	0.67*	0.64*	0.64*
18-authors	0.80	0.82	0.80	0.80

Table 3: Results of the GAN-BERT Model for Standalone Training on Manually Selected Author Sample Sets

* - mini result across a metric [†] - max value across a metric

6.4 Text Sample Size per Novel

To investigate how each novel’s sample size affects the model performance, we selected the 18 authors’ cases and experimented across different text sample sizes ranging from 5 to 35 text chunks per novel. Each sample consists of a text chunk of 512 words

drawn from the book text. For example, a text sample size of 5 means that we selected 5 x 512 text chunks from the book text, which resulting 5 separate instances in the dataset. We performed this experiment using the same 20 books per author.

The results in Table 4 demonstrate that increasing the sample size has a negative impact on the model’s performance across all sample sets for the 18-author model. In this experiment, as the sample size increases, the model is trained on the same novels and 18 authors during training. One of the main findings is that the larger text samples from novels only sometimes lead to better performance. The model may have shown a negative impact in larger text sample sizes due to the high variance in the data or overfitting. Hence, further investigation must be performed to identify the optimal text sample size per novel under different experiment settings.

Sample Size	Accuracy	F1	Precision	Recall
5	0.92 [†]	0.93 [†]	0.92 [†]	0.92 [†]
10	0.91	0.91	0.91	0.91
15	0.89	0.90	0.89	0.89
20	0.80*	0.82*	0.80*	0.80*
25	0.86	0.87	0.86	0.86
30	0.86	0.87	0.86	0.86

Table 4: Effect of Sample Size on Model Performance for 18-Author Classification

* - mini result across a metric † - max value across a metric

6.5 Baseline Experiments

We evaluated various baseline models with different datasets including IMDB20, Blog20, 20-authors and 18-authors. The accuracy results obtained are reported in Table 5. Using stylometric features performed the worst with an accuracy of 0.14 on the IMDB20 dataset. The proposed GAN-BERT model outperforms the stylometric and character n-gram-based models but does not perform as well as the TF-IDF and BertAA models. Our proposed model performs as well as the other models on IMDB20 dataset; however, BERTAA outperforms the others on our dataset. This indicates that further improvements (e.g. including other features such as tf-idf or stylometric features) are needed to enhance the proposed GAN-BERT model performance on specific datasets.

Model	IMDB20	Blog20	20-authors	18-authors
Stylometric (Sari et al., 2018)	0.14*	0.11*	0.14*	0.11*
Character Ngram (Fabien et al., 2020)	0.69	0.23	0.94	0.95
Word level TF-IDF (Fabien et al., 2020)	0.97 [†]	0.47	0.91	0.90
BERTAA (Fabien et al., 2020)	0.97 [†]	0.62 [†]	0.99 [†]	0.99 [†]
Proposed Model	0.96	0.40	0.63	0.80

Table 5: Baseline Experiment Results

* - mini result across a metric † - max value across a metric

7 Conclusion

This research proposes a GAN-BERT-based model for authorship attribution in late-19th-century novels. Our primary focus is identifying how the author counts and the text sample size per book affects the model’s performance. The manually selected five authors’ combinations indicate that the model’s performance degrades when the number of authors increases. The declining trend is the same for transfer-learning models, although the overall performance is better than the standalone models. Additionally, we experimented with how transfer learning has improved the mean accura-

cies over manually selected author sample sets for each n-author case. A future improvement would be an experiment around few-shot and zero-shot tests. Furthermore, it would be interesting to experiment with different GAN and transformer models replaced in this model architecture.

Limitations

While this research provides valuable insights into using the GAN-BERT model for authorship attribution, there are also a few limitations to note. We only focused on a limited number of authors from the late 19th century, which may include shortcomings towards model generalisability. Future research should consider using the whole dataset of long 19th-century novelists to address this limitation. Due to the copyright issues explained in Section 4.6 and Section 7, we do not release the whole dataset, instead, we release scripts to reproduce the datasets. Furthermore, incorporating a rich feature set and comparing performance among different models would be another interesting research direction.

Ethics Statement

The duration 1800-1914 is considered as the out-of-copyright duration in Project Gutenberg, under the categories ‘Rule 1: Works First Published Before 95 Years Ago and Before 1977’ and ‘Rule 10(c) - Works of Treaty Parties and Proclamation Countries First Published Between 1923 and 1977’ (Gutenberg). Although the duration is out-of-copyright regarding literary works, we stored the data securely with restricted access. We do not release the dataset.

References

- A. Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20:67–75.
- Douglas Bagnall. 2015a. Author identification using multi-headed recurrent neural networks. *ArXiv*, abs/1506.04891.
- Douglas Bagnall. 2015b. [Author identification using multi-headed recurrent neural networks](#). In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org.

- Georgios Barlas and Efstathios Stamatatos. 2020. **Cross-domain authorship attribution using pre-trained language models**. In *Artificial Intelligence Applications and Innovations - 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5-7, 2020, Proceedings, Part I*, volume 583 of *IFIP Advances in Information and Communication Technology*, pages 255–266. Springer.
- Dainis Boumber, Yifan Zhang, and Arjun Mukherjee. 2018. Experiments with convolutional neural networks for multi-label authorship attribution. In *LREC*.
- Marcelo Luiz Brocardo, Issa Traoré, Isaac Woungang, and Mohammad S. Obaidat. 2017. Authorship verification using deep belief network systems. *Int. J. Commun. Syst.*, 30.
- Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. **GAN-BERT: generative adversarial learning for robust text classification with a bunch of labeled examples**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2114–2119. Association for Computational Linguistics.
- Sevtap Duman, Kubra Kalkan-Cakmakci, Manuel Egele, William K. Robertson, and Engin Kirda. 2016. Emailprofiler: Spearphishing filtering with header and stylometric features of emails. *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, 1:408–416.
- Maciej Eder. 2015. **Does size matter? authorship attribution, small samples, big problem**. *Digit. Scholarsh. Humanit.*, 30(2):167–182.
- Maciej Eder. 2017. **Short samples in authorship attribution: A new approach**. In *12th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2017, Montréal, Canada, August 8-11, 2017, Conference Abstracts*. Alliance of Digital Humanities Organizations (ADHO).
- Maël Fabien, Esaú Villatoro-Tello, Petr Motlíček, and Shantipriya Parida. 2020. **Bertaa : BERT fine-tuning for authorship attribution**. In *Proceedings of the 17th International Conference on Natural Language Processing, ICON 2020, Indian Institute of Technology Patna, Patna, India, December 18-21, 2020*, pages 127–137. NLP Association of India (NLP AI).
- Olga Fourkioti, Symeon Symeonidis, and Avi Arampatzis. 2019. **Language models and fusion for authorship attribution**. *Inf. Process. Manag.*, 56(6).
- Neal P. Fox and Omran Ehmoda. 2012. Statistical stylometrics and the marlowe-shakespeare authorship debate.
- Glenn Fung. 2003. **The disputed federalist papers: SVM feature selection via concave minimization**. In *Proceedings of the Richard Tapia Celebration of Diversity in Computing Conference 2003, Atlanta, Georgia, USA, October 15-18, 2003*, pages 42–46. ACM.
- Zhenhao Ge, Yufang Sun, and Mark J. T. Smith. 2016. **Authorship attribution using a neural network language model**. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 4212–4213. AAAI Press.
- Oleg Granichin, Lev Klebanov, Dmitry Shalymov, and Zeev Volkovich. 2015. Authorship attribution method based on knn re-sampling approach. In *PROCEEDINGS ELMAR-INTERNATIONAL SYMPOSIUM ELECTRONICS IN MARINE*. Institute of Electrical and Electronics Engineers Inc.
- Project Gutenberg. *Copyright How-To*. <https://www.gutenberg.org/help/copyright.html>.
- Hassina Hadjadj and Halim Sayoud. 2021. **Arabic authorship attribution using synthetic minority over-sampling technique and principal components analysis for imbalanced documents**. *Int. J. Cogn. Informatics Nat. Intell.*, 15(4):1–17.
- Marjan Hosseinia and Arjun Mukherjee. 2018. Experiments with neural networks for small and large scale authorship verification. *ArXiv*, abs/1803.06456.
- Zhiqiang Hu, Roy Ka-Wei Lee, Lei Wang, Ee-Peng Lim, and Bo Dai. 2020. **Deepstyle: User style embedding for authorship attribution of short texts**. In *Web and Big Data - 4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, September 18-20, 2020, Proceedings, Part II*, volume 12318 of *Lecture Notes in Computer Science*, pages 221–229. Springer.
- Sylvio Barbon Junior, Rodrigo Augusto Igawa, and Bruno Bogaz Zarpelão. 2016. Authorship verification applied to detection of compromised accounts on online social networks. *Multimedia Tools and Applications*, 76:3213–3233.
- Patrick Juola. 2021. **Verifying authorship for forensic purposes: A computational protocol and its validation**. *Forensic Science International*, 325:110824.
- Andrei Kazlouski. 2019. **Text style imitation to prevent author identification and profiling**. Master’s thesis, Aalto University. School of Science.
- Mike Kestemont, Justin Anthony Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016. Authenticating the writings of julius caesar. *Expert Syst. Appl.*, 63:86–96.
- Bryan Klimt and Yiming Yang. 2004. **The enron corpus: A new dataset for email classification research**. In *Machine Learning: ECML 2004, 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004, Proceedings*, volume 3201 of *Lecture Notes in Computer Science*, pages 217–226. Springer.
- Moshe Koppel, Jonathan Schler, and Shlomo Engelson Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45:83–94.

- Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *J. Mach. Learn. Res.*, 8:1261–1276.
- Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *COLING*.
- Tempestt J. Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon L. Woodard. 2018. [Surveying stylometry techniques and applications](#). *ACM Comput. Surv.*, 50(6):86:1–86:36.
- Weihan Ou, Steven H.H. Ding, Yuan Tian, and Leo Song. 2022. [Scs-gan: Learning functionality-agnostic stylometric representations for source code authorship verification](#). *IEEE Transactions on Software Engineering*, pages 1–1.
- Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. [Character-level and multi-channel convolutional neural networks for large-scale authorship attribution](#). *CoRR*, abs/1609.06686.
- Chakaveh Saedi and Mark Dras. 2021. [Siamese networks for large-scale author identification](#). *Comput. Speech Lang.*, 70:101241.
- Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. [Improved techniques for training gans](#). In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234.
- 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, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 343–353. Association for Computational Linguistics.
- Raheem Sarwar, Chenyun Yu, Ninad Tungare, Kanatip Chitavisutthivong, Sukrit Sriratanawilai, Yaohai Xu, Dickson Chow, Thanawin Rakthanmanon, and Sarana Nutanong. 2018. [An effective and scalable framework for authorship attribution query processing](#). *IEEE Access*, 6:50030–50048.
- Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. [Effects of age and gender on blogging](#). In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006*, pages 199–205. AAAI.
- Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. [Authorship attribution with topic models](#). *Comput. Linguistics*, 40(2):269–310.
- Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. [A4NT: author attribute anonymity by adversarial training of neural machine translation](#). In *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, pages 1633–1650. USENIX Association.
- Efstathios Stamatatos. 2009. [A survey of modern authorship attribution methods](#). *J. Assoc. Inf. Sci. Technol.*, 60(3):538–556.
- Efstathios Stamatatos. 2018. [Masking topic-related information to enhance authorship attribution](#). *J. Assoc. Inf. Sci. Technol.*, 69(3):461–473.
- Jianwen Sun, Zongkai Yang, Sanya Liu, and Pei Wang. 2012. Applying stylometric analysis techniques to counter anonymity in cyberspace. *J. Networks*, 7:259–266.
- Wanbing Tang, Chunhua Wu, Xiaolong Chen, Yudao Sun, and Chen Li. 2019. [Weibo authorship identification based on wasserstein generative adversarial networks](#). In *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, pages 1–5.
- Andrew Tausz. 2011. Predicting the date of authorship of historical texts.
- Jeffrey R. Thompson and John Rasp. 2016. [Did c. s. lewis write the dark tower?: An examination of the small-sample properties of the thisted-efron tests of authorship](#). *Austrian Journal of Statistics*, 38(2):71–82.
- Enrico Tuccinardi. 2017. An application of a profile-based method for authorship verification: Investigating the authenticity of pliny the younger’s letter to trajan concerning the christians. *Digit. Scholarsh. Humanit.*, 32:435–447.
- 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2742–2753. Association for Computational Linguistics.
- Yifan Zhang, Dainis Bumber, Marjan Hosseinia, Fan Yang, and Arjun Mukherjee. 2021. Improving authorship verification using linguistic divergence. In *ROMCIR@ECIR*.