Supervised GAN-BERT for Authorship Attribution on Archaic Greek Poetry

Olivia McCauley
UC Berkeley
olive.ray.mccauley@berkeley.edu

Toufiq Hossain
UC Berkeley
thoss0513@berkeley.edu

Austin Ho
UC Berkeley
aicho@berkeley.edu

Abstract

This study presents the application of a supervised BERT-GAN model to authorship attribution tasks on Archaic Greek texts. Due to small text corpora, linguistic obscurity, and limited availability of open-source NLP tools, authorship attribution with traditional neural classifiers in this problem area is challenging. We incorporated adversarial training into an authorship attribution task on Archaic Greek corpora as a means of data augmentation in a low-resource language setting. Using a BERT model fine-tuned on Ancient Greek, we extracted CLS tokens that the GAN's generator could imitate, introducing more training examples without the need for feature extraction and text modification. Our BERT-GAN outperforms our baseline model by a small margin, but through our experiments, we demonstrate that employing sampling techniques for data augmentation significantly optimizes our model's performance. This approach not only pushes the boundaries of traditional literary analysis but also sets a precedent for the use of advanced machine-learning techniques in the study of ancient texts.

1 Introduction

Attributing Ancient Greek texts to Homer has been a long-contested issue in the Classics, and similar matters regarding forgeries have arisen with other ancient texts like the Bible. For example, classicists and linguists long debated whether the Homeric Hymns were attributable to Homer due to strong stylistic similarities with the Iliad and Odyssey, with scholars eventually attributing the Hymns to an author or tradition other than Homer. Many ancient texts are highly formulaic, making them easily forgeable by humans and machines, but frauds are difficult to prove from textual clues alone. Authorship attribution of ancient texts thus has broad applications in archaeology and the humanities to attribute authors to literary artifacts whose provenance is contested.

However, ancient literary corpora attributed to single authors are severely limited in volume and content, making it difficult to utilize computational methods that distinguish one author's "stylistic fingerprint" from another's for effective classification. As a result, balancing classes with statistical sampling techniques may not introduce sufficient variance to the corpus, leading conventional text classification methods to underperform. The problem, thus, is twofold. With limited examples, it is challenging for humans or machines to identify the author of a genuine artifact, and potentially even more difficult to identify fakes.

This paper investigates the application of BERT-augmented generative adversarial networks (GANs) to authorship attribution tasks on low-resource ancient literary corpora. We pose the following authorship attribution problem: distinguish between the corpora of three Archaic Greek authors: Homer, Hesiod, and the Homeric Hymns. The GAN-BERT architecture was chosen because it has been shown to perform well on other authorship classification tasks with limited labeled data. Thus, we sought to build a GAN-BERT with a discriminator model that outperforms a comparable neural classifier on our authorship attribution task, even after employing data augmentation techniques. The research questions in this study are as follows:

R1: Does the GAN-BERT model's discriminator outperform traditional CLS token classification methods on a low-resource and class-imbalanced dataset?

R2: Can the GAN-BERT model's generator act as a form of "data-augmentation" in low-resource language problems, by mimicking the labeled data to introduce variance and volume into the training dataset?

R3: Can the GAN-BERT effectively distinguish forgeries from authentic texts?

2 Background

2.1 Authorship Attribution in the Classics

Problems of quantitative authorship attribution (AA) trace their origins to 19th-century stylometry, although debates surrounding the authorship of Archaic Greek poetry thrived for centuries prior. Beginning with Milman Parry's "Oral-Formulaic Theory" of Homeric texts, stylistic features such as poetic meter, line length, hapax legomena, and other syntactic features were used alongside literary analysis to attribute and date Classical texts (Beller and Spicer). As computational methods advanced, standard machine learning methods such as SVMs, KNN, and Decision Trees have been used for authorship attribution on labeled datasets with reasonable success between 60 - 80% (Beller and Spicer).

2.2 Low-Resource Language Challenges

Despite the success of previous investigations, extracting and interpreting syntactic and lexical features from original Ancient Greek texts for machine learning requires domain knowledge and language proficiency, which has limited research in an already low-resource language setting. Ancient Greek has complex syntax and grammatical forms. As a result, much of the computational linguistics research in this area has required collaboration from Classicists as co-authors or human evaluators.

Yet another challenge for authorship attribution in this area is thematic similarity between authors. Works from the Archaic Greek period were recorded from an oral storytelling tradition that often involved imitating preceding poets' styles. Two of the corpora considered for this project, the works of Hesiod and the Homeric Hymns, contain works that are stylistically close to Homer, contributing to this classification problem's difficulty. Data augmentation techniques are also hard to implement due to the lack of NLP tools and data. Even basic methods like word swapping can be inappropriate for the period or alter the poetic meter, damaging data quality (Yusuf et al., 2022).

2.3 GAN-BERTs for Low-Resource Classification Tasks

Adversarial learning techniques have shown promising performance on similar low-resource classification tasks. Saliman et al showed that training a semi-supervised adversarial network (SS-GAN) for multi-class (rather than binary) image classification was effective on small amounts of labeled data (Salimans et al., 2016). Elaborating on this research, Croce et al demonstrated empirically that in natural language topic classification problems, SS-GANs "significantly reduced the need for labeled data, and that even with less than 200 annotated examples it is possible to obtain results comparable with a fully supervised setting (Croce et al., 2020). Yusef et al hypothesized that SS-GANs were highly effective on small datasets because "SS-GANs ... can act as an additional source of information in a semi-supervised setting [by capturing] the characteristics of the training examples and [generating] similar examples that are nearly indistinguishable from the real training examples," and demonstrated that adding an SS-GAN module to BERT-based Arabic dialect classification models improved model performance despite limited labeled data (Yusuf et al., 2022).

BERT-augmented GAN architectures are particularly desirable for our problem space because language-specific BERT models effectively encode information about style and content without the need for further language knowledge. Wang et al evaluated the effectiveness of various GAN-BERT models for Chinese news classification, achieving far superior classification results with adversarial learning than with conventional neural methods (Wang and Zhang, 2022). Silva et al investigated the efficacy of supervised BERT-GAN models for multi-class authorship attribution on 19th century novels, which demonstrated high accuracy scores in the 80%-90% range across multiple corpora with 2 to 18 authors (Silva et al., 2023).

2.4 Project Motivation and Overview

The BERT-GAN's high performance in multi-class authorship attribution, particularly as demonstrated by Silva et al, motivated our use of the architecture for this paper. Our model utilizes the same problem structure as the BERT-GAN training task where "the discriminator D is trained over N+1 classes to assign the true samples to a class from 1,.. N... and the fake sample generated from the generator G represents the (N + 1)th class (Silva et al., 2023). This is hypothesized not only to improve authorship detection, but also to allow the discriminator to detect "obfuscation and forgery since it is trained with fake samples similar to the original author-written texts" (Silva et al., 2023).

To our knowledge, Silva et al's work is the

only BERT-GAN model for authorship attribution and forgery detection (Silva et al., 2023). This makes our research the only attempt to apply supervised GAN-BERT models for authorship attribution on Archaic Greek poetry; a corpora that introduces substantial difficulty for classification due to its constituent texts' formulaic structure, complex grammar, and limited dataset size. We evaluate model success with standard classification metrics, emphasizing accuracy and F1 scores, which are standard for this research area.

3 Data

3.1 Dataset Creation

We collected the dataset from the Perseus Digital Library, an open-access database of natural language texts from the classics. To combat potential leakage issues from rhythmic schemes and literary periods, we selected authors writing poems in dactylic hexameter during the Archaic Greek literary period: Homer, Hesiod, and the Homeric Hymns. We elected to use the original Ancient Greek texts to avoid stylistic distortions introduced during translation. A strong class imbalance was observed, as Homer represented 85% of the examples, while Hesiod and the Hymns each comprised 7% of the data.

3.2 Data Pre-processing

The poetic corpora from each author were parsed from XML form, separated by line, and converted into Unicode. Because the corpora consist of poems, line segments were used instead of sentence segments to divide large texts into examples.

3.3 CLS Token Extraction

A pre-trained Ancient Greek BERT model from HuggingFace was used to tokenize each line(Singh et al., 2021). The tokenized inputs were then processed through the pre-trained model, and the [CLS] token for each line was extracted from the output.

3.4 Class Balancing

Resampling techniques were used to create two new balanced datasets, yielding three datasets in total.

- Imbalanced Dataset: See Dataset Selection.
- Undersampled Dataset: 666 samples were drawn from each class without replacement, for an overall dataset size of 2000.
- 10k Dataset: 10,000 samples were drawn from each class. Since Hesiod and the Homeric Hymns had fewer than 10,000 examples, samples from these classes were drawn with replacement. Since Homer had more than 10,000 line examples, samples from this class were drawn without replacement. The overall dataset size was 30,000.

Each dataset was randomly split into training and testing sets with an 80/20 split.

4 Models

Grid search was employed to identify the best hyperparameters (latent dimensions, batch size, and epochs) for training the GAN. The best hyperparameters for the discriminator were also used to define and train the supervised baseline classifier.

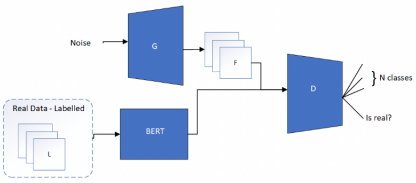

Figure 1: GAN Architecture

4.1 GAN-BERT Model

This model was based on the modified GAN-BERT architecture used by Silva et al. It consists of a discriminator and generator model, which produce class predictions with a softmax activation layer. The generator creates artificial samples from random noise, while the discriminator attributes each token to one of the three authors or the generator. Through adversarial training, the generator creates increasingly realistic samples, and the discriminator becomes more effective at the 4-author classification problem (3 real authors plus the fake generator "author").

```
Model: "generator"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_6 (Dense)             (None, 200)               100200

 dense_7 (Dense)             (None, 100)               20100

 dense_8 (Dense)             (None, 768)               77568

=================================================================
Total params: 197868 (772.92 KB)
Trainable params: 197868 (772.92 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 2: Generator Architecture

Model: "discriminator-fake"

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 768)	0
dense_1 (Dense)	(None, 100)	76,900
leaky_re_lu_1 (LeakyReLU)	(None, 100)	0
dropout (Dropout)	(None, 100)	0
dense_2 (Dense)	(None, 3)	303
dense_3 (Dense)	(None, 4)	16

Total params: 77,219 (301.64 KB)
Trainable params: 77,219 (301.64 KB)
Non-trainable params: 0 (0.00 B)

Figure 3: Discriminator Architecture

4.2 Supervised Classifier Baseline

The baseline classifier architecture was constructed using 100-dimensional Dense layers, Dropout, and Leaky ReLu. The model was trained on the 3-author classification problem using a Softmax layer at the end without fake data.

Model: "supervised-discriminator"

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 768)	0
dense_1 (Dense)	(None, 100)	76,900
leaky_re_lu_1 (LeakyReLU)	(None, 100)	0
dropout (Dropout)	(None, 100)	0
dense_2 (Dense)	(None, 3)	303
activation (Activation)	(None, 3)	0

Total params: 77,203 (301.57 KB)
Trainable params: 77,203 (301.57 KB)
Non-trainable params: 0 (0.00 B)

Figure 4: Discriminator Architecture

5 Results

The model architectures from our experiment are shown below. In order to assess our model, we used the weighted-averaged F1 score, recall, precision, and accuracy, which are standard for classification.

Across all classification metrics on the undersampled and imbalanced datasets, the baseline supervised neural classifier outperformed the GAN discriminator, on both the training and test data. Model performance on the test data is summarized in 1 and 2.

Across every metric except accuracy, discriminator and baseline models performed best on the 10k balanced dataset, with the baseline achieving scores in the 52% range and the discriminator achieving scores of 68%. With the large, balanced 10k dataset, we observe a significant improvement of between 20% - 40% across all metrics. F1, precision, recall, and accuracy are essentially identical within each model on the 10k dataset, indicating that the model performs uniformly across all classes and metrics.

On F1, recall, and precision, the undersampled class does not appear to yield a significant performance improvement for the discriminator or the baseline. For both the baseline and discriminator models, there is at most a 4% difference between the classification metrics on each dataset.

The best accuracy scores for both models are observed in the imbalanced dataset, due to the model's ability to incorrectly over-classify examples as Homer with little penalty due to the class imbalance. This is further reinforced when undersampling to balance class proportions doesn't reduce precision, recall, or f1, but does reduce accuracy substantially.

Supervised Classifier Test Set Performance				
Baseline	Precision	Recall	F1	Accuracy
10k	0.53	0.52	0.52	0.52
Undersampled	0.28	0.34	0.27	0.35
Imbalanced	0.31	0.33	0.31	0.84

Table 1: Weighted-Averaged Classification Metrics on the 3-Author Attribution task

GAN Test Set Performance				
Discriminator	Precision	Recall	F1	Accuracy
10k	0.68	0.68	0.68	0.68
Undersampled	0.19	0.28	0.22	0.26
Imbalanced	0.16	0.25	0.19	0.63

Table 2: Macro-Averaged Classification Metrics on the 3-Author Attribution task

Baseline Testing Set Performance: 10k Balanced Dataset

	precision	recall	f1-score	support
Hesiod	0.56	0.44	0.49	2481
Homer	0.53	0.54	0.54	2503
Homeric Hymns	0.49	0.58	0.53	2516
accuracy			0.52	7500
macro avg	0.53	0.52	0.52	7500
weighted avg	0.53	0.52	0.52	7500

baseline test loss: 0.9704591631889343
baseline test accuracy: 0.5230666399002075

Figure 5: Classification Report for Baseline Test on 10k Dataset

6 Discussion

The results yield substantial insights into the utility of supervised BERT-GAN models in low-resource

```
Discriminator Testing Set Performance: 10k Balanced Dataset
               precision    recall  f1-score   support

       Hesiod       0.57      0.57      0.57      2481
        Homer       0.59      0.66      0.62      2503
Homeric Hymns       0.57      0.51      0.54      2516
Generated CLS       1.00      1.00      1.00      2500

     accuracy                           0.68     10000
    macro avg       0.68      0.68      0.68     10000
 weighted avg       0.68      0.68      0.68     10000


discriminator test loss:  0.8538110256195068
discriminator test accuracy:  0.6819999814033508
```

Figure 6: Classification Report for Discriminator Test on 10k Dataset

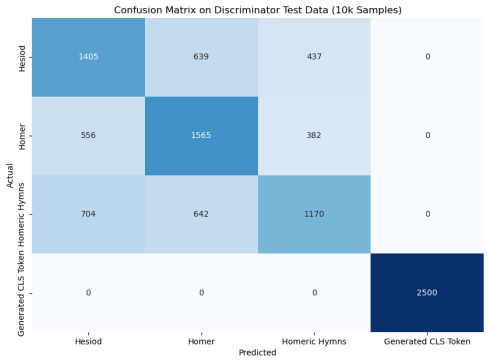

Figure 7: Discriminator Confusion Matrix for Test 10k Data

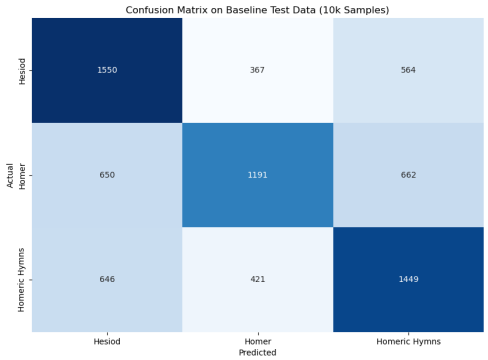

Figure 8: Baseline Confusion Matrix for Test 10k Data

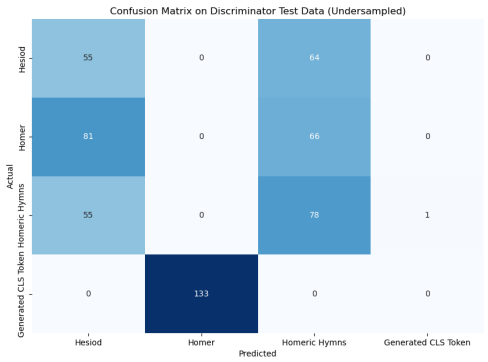

Figure 9: Discriminator Confusion Matrix for Test Undersampled Data

settings. Although previous research showed that adversarial training methods can reduce the need for data, it appears that overall performance on this authorship attribution task improved substantially with additional data. At 68% accuracy, the BERT-GAN model's performance was low compared to the BERT-GAN authorship attribution model for 19th-century literature, which reported results in the 80% - 90% range. This indicates that authorship attribution on our set of authors is potentially more difficult than other tasks posed to BERT-GANs in the literature, perhaps due to lexical and syntactic patterns, or strong stylistic similarity between authors.

Additionally, it appears that the macro-averaged classification metrics for the discriminator were heavily inflated by the generated CLS tokens. It appears that the CLS tokens created by the generator were correctly classified 100% of the time, even though the CLS tokens had been created by a fully trained generator. When comparing the classification metrics of the genuine texts, they are only better than the baseline by <10%. This observation implies that the CLS tokens created by the generator during training on the 10k dataset must have been substantially different than all authors. This could explain the lack of substantial improvement from the baseline – the data quality was not sufficiently similar to improve the classifier's performance.

Inspecting the confusion matrices for the discriminator trained on the undersampled dataset reveals an intriguing difference. Instead of the CLS tokens having been classified perfectly, all CLS tokens were classified as an example of Homer, while no examples of Homer were correctly classified. The difference between generator performance on the 10k dataset versus the undersampled dataset indicates that the fake CLS tokens may have interfered in some way with the model's ability to properly identify Homer, and potentially other classes as well. Moreover, in both the 10k and undersampled dataset, all the CLS tokens were classified in a single category, suggesting that every token produced by the generator is highly similar to the point of not being useful as a form of data augmentation. This also demonstrates the need for further investi-

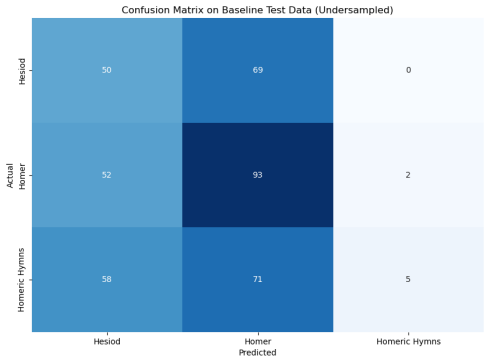

Figure 10: Baseline Confusion Matrix for Test Undersampled Data

gation into the generator's behavior, and additional improvements on this method before it can be used effectively for data augmentation.

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.

At the beginning of this paper, we proposed a GAN-BERT-based model for authorship attribution on Archaic Greek texts. Our primary focus was investigating whether GAN-BERT models could improve the performance of classifiers trained for authorship attribution, and specifically whether the mechanism for that improvement could be attributed to the generator producing false training examples as a form of data augmentation.

Returning to the research questions posed at the beginning of the paper, we make the following observations. With regard to whether the GAN-BERT model discriminator outperforms typical neural classifiers on a low-resource and class-imbalanced dataset, as well as the impact of class balancing methods, we would conclude that the adversarial training structure yields modest increases in model performance only after resampling techniques have been applied. Concerning whether the generator in adversarial training can provide a form of data augmentation, we were unable to make any conclusive determinations. It appears that with further fine-tuning, a generator might be able to produce sufficiently varied training examples, but inconsistent generator behavior across datasets made it hard to arrive at any conclusions. Finally, as to whether the GAN-BERT can effectively identify fakes, it appears that the fakes are internally consistent enough to consistently all be classified into the same authorial corpus, but that category is not always the right one.

Despite the inconclusive results of this study, the authors still believe that adversarial training is a promising method for improving model performance in low-resource language settings, and additional investigation is needed.

8 Further Research

Opportunities for further research are plentiful. The most important extension of this research would be to do more investigation into the generator's behavior, and experiment with the architecture in an attempt to broaden the variety of examples it produces during training time.

Later studies could experiment with adding or substituting different Ancient Greek and Roman authors from different periods, who write in meters besides dactylic hexameter. In terms of experimenting with different architectures, one could create a generator that generates text instead of CLS tokens as a way to obtain high-quality training examples for this and other problems. Data augmentation techniques, while difficult to implement and needing expert supervision, are another avenue researchers could pursue. Additionally, we encountered some issues with reproducibility regarding CLS token quality, so running experiments on the generator or training loop to gather more data and observe statistical trends could be beneficial to this problem area.

References

Sarah Beller and James Spicer. Attribution of Contested and Anonymous Ancient Greek Works.

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*, pages 2114–2119, Online. Association for Computational Linguistics.

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

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. ArXiv:1606.03498 [cs].

Kanishka Silva, Burcu Can, Frédéric Blain, Raheem Sarwar, Laura Ugolini, and Ruslan Mitkov. 2023. Authorship Attribution of Late 19th Century Novels using GAN-BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 310–320, Toronto, Canada. Association for Computational Linguistics.

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021. A pilot study for bert language modelling and morphological analysis for ancient and medieval greek. In *The 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2021)*.

Xiaoning Wang and Yang Zhang. 2022. Research on CGAN-BERT and RGAN-BERT Models for Short Text Classification based on Semi-Supervised Model. In *2022 the 5th International Conference on Information Science and Systems*, pages 118–124, Beijing China. ACM.

Mahmoud Yusuf, Marwan Torki, and Nagwa El-Makky. 2022. Arabic Dialect Identification with a Few Labeled Examples Using Generative Adversarial Networks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 196–204, Online only. Association for Computational Linguistics.