Computational Approaches to Classifying the Platonic Corpus

Michael Gao, Ryo Nagao[1]
Yale University
m.gao@yale.edu,
ryo.nagao@yale.edu

Abstract

This project introduces computational analysis into controversies over ancient pseudo-authors. At least ten Ancient Greek works transmitted to modernity under the name of Plato are considered by many to be spurious, but these have been identified as such only through qualitative and statistical methods. Using a dataset consisting of treatises firmly ascribed to Plato and works by other Ancient Greek authors and splitting it into chunks, we create FNN-, LSTM-, and Transformer-based classifiers that can distinguish between Platonic and non-Platonic works. We then apply this model to the texts in the Platonic Corpus often considered to be spurious so that they can be assigned probabilities scores. We find that the LSTM and the Transformer-based model achieve remarkably high accuracies after training, and their outputs reveal that most of the treatises considered spurious are highly similar to Plato's legitimate works, with the exception of *Definitions* and possibly *Letters*. Although there are reservations such as the scarcity of data and the extent to which computational models can distinguish between similarity and authenticity, we believe that the project opens a novel path for re-evaluating spurious works and combining the fields of computational linguistics and classical philology.

1 Introduction

Controversy over falsely attributed works is commonplace across studies of ancient literature, but the authenticity of the Platonic Corpus has been particularly questioned since the nineteenth century. Although there is no consensus on which works are spurious, *Alcibiades I* and *II*, *Amatores*, *Cleitophon*, *Epinomis*, *Hipparchus*, the thirteen *Letters*, *Minos*, and *Theages* are sufficiently disputed that they cannot be firmly considered legitimately Platonic (Press, 2012; Joyal, 2019). In addition, there is a scholarly consensus that *Definitions* is not a treatise by Plato but a work of little philosophical value that was circulated in the Platonic Academy, perhaps by Speusippus, the successor of Plato as the head of the Academy (Ingenkamp, 1967). However, traditional methods for identifying spurious works have been qualitative or statistical, often stylometric (Brandwood, 2006; Tarrant, 2017). Here, computational methods may offer new insight by locating features of the text beyond the stylistic features that appear on the surface.

The field of classical philology has been revolutionized by the advent of digital tools, but the union of the two still forms a nascent approach to ancient texts (Berti, 2019). Achievements in recent years include the creation of the Classical Language Toolkit for Python, or CLTP (Johnson et al., 2021), BERT models for Latin and Ancient Greek (Bamman and Burns, 2020; Singh et al., 2021; Yamshchikov et al., 2022), PLMs (Riemenschneider and Frank, 2023), and the organization of the First Workshop on Ancient

[1] All code is uploaded to https://github.com/Firebro113/LING-227-Final-Project.

Language Translation in 2023. Nevertheless, the generalization of models still proves to be a difficult task (Kostkan et al., 2023; see also Yousef et al., 2023), and there have been few attempts to apply these approaches to subsections of the extant classical corpora (e.g. Köntges, 2020) or specific authors such as Plato. Although tools are being developed for exploring intertextuality in Platonic reception (Wöckener-Gade and Pöckelmann, 2023), examining the treatises doubtfully ascribed to Plato would open a new direction for the fusion of Classics and NLP.

In this paper, we train a variety of classification models on a dataset consisting of works that are generally agreed to be by Plato and works attributed to other authors from the extant ancient Greek corpus. The aim of this project is to develop models that can accurately distinguish Platonic and non-Platonic works, so that they can then be applied to the spurious treatises to determine whether they should be accepted or rejected based on computational analysis.

2 Methods

2.1 Data

All of the ancient Greek texts used in this project were downloaded from the Perseus Digital Library hosted by Tufts University.[2] One half of the main dataset is the Platonic Corpus excluding the nine spurious works mentioned in the introduction, which totals to about 500,000 words. The other half of the main data consists of roughly the same length of texts by authors other than Plato. Because there is no centralized Ancient Greek dataset from which to pull random texts, these non-Platonic texts were manually chosen and downloaded. A large selection of texts were chosen from a variety of dialects, genres, and time periods ranging from Archaic to Roman to represent the extant Ancient Greek corpus. At the same time, philosophical texts occupy a large portion of our selection because our goal is to create a model that can determine the authenticity of disputed works by Plato, which are philosophical in nature (Irwin, 2006). However, we recognize that the manual selection process involves bias and hope that a future project will create a dataset of the extant Ancient Greek corpus that is appropriate for computational analysis. The works mentioned as spurious above were organized into a separate “dubia” dataset, which totals to about 50,000 words.

The raw data is preprocessed by removing all non-Greek characters except whitespaces and converting the Greek letters into Latin characters using unidecode. The unidecode module is not perfect, however, because it discards diacritics. Diacritics are essential for differentiating numerous words in Ancient Greek: for example, κῆρ (“heart”) with a circumflex accent is not the same as κήρ (“doom”) with an acute accent. As part of this process, unidecode neglects to account for aspirations, which are usually transliterated in classical scholarship: for instance, ὅδος is commonly transliterated as *hodos*, but unidecode returns *odos*. Although this feature may seem unideal for this project, we believe that this is a benefit, as diacritics were added postclassically in manuscript and scholarly traditions and are therefore interpolations in ancient texts (see Allen, 1987; Probert, 2006). The same is true for punctuation, which was also removed from the dataset. We also did not lemmatize the text, although this is possible using the CLTP, because morphology is a relevant characteristic for categorizing texts.

Each text was processed the same way: The text was converted computationally into *chunks*, sliding windows of various sizes: 25, 50, 100, and 200 words, with preceding labels to indicate whether the text was by Plato (labeled 1) or another author (labeled 0). Then, the chunks were split into training, validation, and test sets in a 8:1:1 ratio randomly. Finally, the training, validation, and test sets for each text were combined, and randomized. Similarly, the “dubia” set was also split into chunks of the same sizes, the only difference being there was no label.

Additionally, word embeddings were created for each word in our corpus through the gensim library. The window size was set to 5 and the dimension 100. The entire corpus was used to train the word embeddings. After training, the same embeddings were used for the entire codebase.

2.2 Algorithms

To evaluate the authorship of texts, three different models were developed: a Feed-Forward Neural Network (FNN), a Recurrent Neural Network (RNN) using Long Short-Term Memory (LSTM),

[2] https://www.perseus.tufts.edu/hopper/.

and a Transformer-based model. We chose these architectures for their distinct approaches to using word order and context, which lead to different results when analyzing Plato's authorship. The validation set was used to determine the optimal hyperparameters, which are as follows.

The Feed-Forward Neural Network serves as the baseline model for the task. It is designed to classify text based on word embeddings independent of word order in each chunk. The architecture uses two hidden layers: The first layer contains 256 neurons, and the second layer has 128 neurons, both using the ReLU activation function to introduce non-linearity. Batch normalization is applied after each hidden layer in order to stabilize training. A dropout rate of 0.3 is used after each hidden layer to prevent overfitting. The output Layer consists of a single neuron with a sigmoid function that outputs a probability between 0 and 1, where 1 indicates that the text is attributed to Plato. The input to the FNN is a single vector generated by averaging the embeddings of all words in the chunk.

The Recurrent Neural Network uses an LSTM layer to capture sequential dependencies in text chunks. Unlike the FNN, this model processes the input as a sequence of word embeddings instead of averaging the whole sequence. The architecture includes a single LSTM layer with 128 hidden units, the last hidden layer is fed as input into a dense layer. Batch normalization is applied to standardize the output. A dropout rate of 0.3 is applied after the LSTM layer to prevent overfitting. The dense hidden layer contains 128 neurons with ReLU activation and a dropout of 0.3. The output layer is a single neuron with a sigmoid activation function. The input to the RNN consists of matrices representing each chunk. Each matrix has dimensions corresponding to the sequence length (number of words in the chunk) and the embedding size (word embedding dimensionality).

The Transformer-based model uses self-attention mechanisms to evaluate relationships between words in each chunk. A custom sinusoidal positional encoding function adds pseudo positioning to the input embeddings so the words have relative order. A single encoder block implements multi-head self-attention with 4 attention heads, followed by a feed-forward network with 128 hidden units. The usual layer normalization and dropout of 0.1 are used for regularization. The output from the encoder block is mean pooled to a single vector, which is then passed to a dense layer with 128 neurons (ReLU activation) and then to a final sigmoid neuron for binary classification. The input to the Transformer is identical to the LSTM model.

Model	**Size 25**	**Size 50**	**Size 100**	**Size 200**	**Aver age**
FNN	0.913 121	0.957 869	0.985 235	0.993 508	0.962 433
LSTM	0.999 901	0.999 901	0.999 901	0.999 901	0.999 901
Transf ormer	0.997 812	0.997 812	0.997 812	0.997 812	0.997 812

Table 1: Performance in test set

2.3 Hyperparameters

Due to restricted processing power, many of the models built took hours to run. Thus, hyperparameter tuning was restricted. Testing word embeddings from sizes 50 to 500, We found that size-100 word embeddings gave considerably fair results while optimizing for runtime. Thus, all our models were trained on the word2vec model with dimension 100. Chunk size was another major hyperparameter that we tuned. Specifically, we tested chunk sizes of 25, 50, 100, and 200. In general, we saw increased test set accuracy with higher chunk size, but for both LSTM and Transformer models, there was no discernable difference between the test set accuracy between 50 chunk size and 200 chunk size.

3 Results

3.1 Overview

We tracked two sets of values for each model with each chunk size (25, 50, 100, and 200): one concerns the accuracy of the model measured using the average performance in the test set. This is the value we used to evaluate the efficacy of each model at each chunk size. Meanwhile, the other consists of the probability score returned for each spurious work, calculated using the average of the output returned for each sliding window (1 for Plato, 0 for not Plato). This value is relevant for our primary research question, which is to investigate how the models that we have created will judge the spurious works in the Platonic Corpus.

Although all models achieved remarkably high accuracies in regard to the test set (Table 1), the

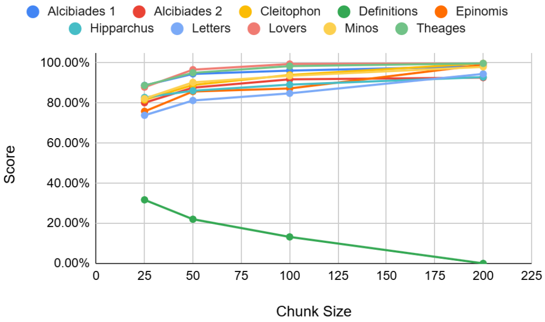

Figure 1: FNN results for dubia

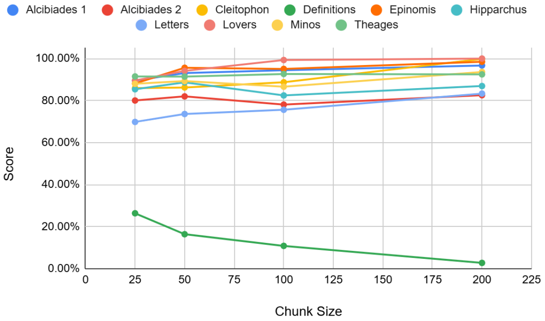

Figure 2: LSTM results for dubia

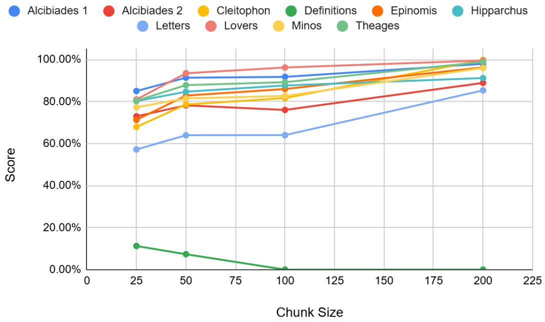

Figure 3: Transformer results for dubia

LSTM had the highest average accuracy of 99.99%. Regarding the spurious works, averages across all texts and chunk sizes were highest for the FNN (83.18%), followed by the LSTM (81.13%), and lastly by the Transformer (76.10%).

3.2 FNN

The FNN achieved high accuracy for the test set, with accuracy rising as the chunk size is increased and an average accuracy of 96.24%. Turning to the spurious works (Fig. 1), *Definitions* scored remarkably lower than any other text, showing negative correlation with chunk size and an average score of 16.79%. Every other spurious text rose in score as the chunk size was increased and generally had high scores. Averages ranged from *Letters*' 93.68% to *Lovers*' 96.00%, giving a range of 2.32%.

3.3 LSTM

The LSTM obtained the highest accuracy for the test set, 99.99% with every chunk size. Concerning the spurious works (Fig. 2), *Definitions* again scored outstandingly lower than the other texts, with a continual decrease and an average of 14.10%. The other spurious texts did not show a consistent increase in score but fluctuated, *Theages* showing a marginal decrease in score from the chunk size of 25 to 50, and 4 texts showing more significant dropoffs from 50 to 100. However, with the exception of *Theages*, a chunk size of 200 returned the highest score. Averages ranged from *Letters*' 75.61% to *Lovers*' 95.75%, giving a range of 20.14%.

3.4 Transformer

The Transformer achieved higher accuracy for the test set than the FNN but lower than the LSTM, consistently showing a 99.78% rate. In the case of the dubia (Fig. 3), as with the other two models, *Definitions* scored conspicuously lower than the other texts, with a continental decrease and an average of 4.69%. Concerning the other spurious texts, they rose in score as the chunk size increased except *Alcibiades 2*, which showed a lower score with a chunk size of 100 than with 50 or 200. Averages ranged from *Letters*' 67.64% to *Lovers*' 92.55%, giving a range of 24.91%.

4 Discussion

We base our analysis of the results primarily on the LSTM and the Transformer for two reasons: for one, the performances of the LSTM and the Transformer were significantly better than that of the FNN. Furthermore, the range between the lowest-scored spurious work and the highest-scored one was significantly higher for the LSTM and the Transformer than the FNN, indicating that the FNN was not effective in distinguishing among the treatises. The models were generally favorable toward the spurious works, with all texts except Definitions achieving scores higher than 50%.

Although those works have average scores that vary about 20% for the LSTM and 25% for the Transformer, it is difficult to draw a firm boundary between what scores can be considered high enough to warrant reconsideration because the scores form a spectrum.

In the case of *Definitions*, every model at every chunk size returned scores that were outstandingly lower than for the other texts, categorizing the work as non-Platonic. Given that these scores agree with the scholarly consensus regarding *Definitions*, this result seems to be an indication that the models are accurate. However, there are several problems with this conclusion. For one, *Definitions* entirely differs from the other spurious works in that it is structured like a dictionary: it is a list of words followed by definitions. This means that most of the text does not contain finite verbs but that the work is a string of noun phrases. It is possible that all three models recognized the absence of finite verbs and other characteristics akin to those of dictionary entries and determined based on form that *Definitions* differs significantly from all of Plato's recognized treatises. Therefore, while the models may be effective in distinguishing the forms of the texts they examine, they do not tell us much about the authorship of *Definitions*.

Meanwhile, *Letters* scores about 5% below the next lowest-scoring treatise in the LSTM and 11% below in the Transformer, which is reason to doubt its authenticity. It may be the case that the same considerations for *Definitions* apply to an extent to *Letters*, which are written in epistolary format and thus diverge from dialectic treatises but are not as different as *Definitions* in that it uses tensed phrases rather than a series of noun phrases. However, since the *Letters* generally lack defenders in authenticity except for letters 7 and 8, the comparatively low accuracy of the *Letters* accords with this qualitative analysis and suggests that the *Letters* may be more spurious than the other works (Forcignanò and Tempesta, 2023).

Turning to other treatises, *Alcibiades I* and *Cleitophon* are the two that have received the most powerful defense in recent scholarship (Joyal, 2019). Whereas *Alcibiades I* consistently produces some of the highest scores, *Cleitophon* is ranked 5th for the LSTM and 7th for the Transformer. In turn, the highest-scoring text across all models is *Lovers*, which is surprising given that its authenticity was called into question even by certain ancient scholars. *Theages* also performs relatively well, occupying 4th rank for the LSTM and 3rd for the Transformer. In this way, the models somewhat diverge from previous scholarly tendencies, which is expected and productive.

In general, the results give us reason to reconsider previous opinions on these supposedly spurious works, since they are found to be very similar to the works authentically ascribed to Plato. This computational method is likely to be more objective than the qualitative or statistical analyses that have primarily been used to identify false attributions. Perhaps more doubt has been placed on much of the Platonic Corpus than is warranted.

Of note is that, interestingly, as the chunk size increases, the more the models tend to ascribe all works besides *Definitions* to Plato. While all these works may be Plato's, this reinforces the idea that perhaps the models have not truly learned Plato's "modus operandi," but instead have just attributed the general writing style of Academic Philosophy to text written by "Plato".

5 Limitations and Future Work

The primary limitation of this work is that of the dataset: we only have about 500,000 words of authentic writing by Plato and are unlikely to find significantly more. Every legitimate treatise is in a similar dialectic format, which makes it difficult for us to account for the possibility that Plato wrote other kinds of works. On the other hand, the non-Platonic texts were collected using a non-random method, which distorts the dataset and highlights the need for the creation of better corpora for ancient languages. One possibility we did not explore was generating texts resembling Plato artificially, for example by using LLMs. The current state of LLMs makes this unviable because they will potentially produce Platonic texts word-for-word in significant portions, which will be difficult for our models to distinguish from completely Platonic texts.

Some linguistic assumptions present across these models are that sliding windows of set sizes give a fair representation of the content of the text, that dialogue can be stripped of character labels, that transliteration preserves the peculiarities of the original text, and that diacritics and punctuation (i.e. textual segments) are not significant (discussed in section 2.1). These changes necessarily distort the results of our study and can be subject to scrutiny in future studies. Especially problematic is the simplification of the dialectic form, which is crucial to Platonic studies and deserves improved representations, though this will inevitably make the dataset more complex to process.

A further consideration is that although we have created models that determine how much a given text is similar to or different from the Platonic and non-Platonic datasets that they are trained on, this does not mean that the results always reveal whether that text was written by Plato or not. For example, in the case of *Definitions*, the models may be using form as the basis of their analysis rather than content. In that case, the results of this project do not shut off the possibility that *Definitions* was written by Plato, though qualitative scholarship has effectively expelled it from the Platonic Corpus. Similar phenomena are more difficult to recognize regarding the other spurious works but may be present.

In terms of future work, an immediate avenue is to perform more experimentation with document segmentation, which would be useful for reaching even higher levels of precision. Also useful would be to separate the individual parts of *Letters*, though we did not attempt this because of the risk of too few documents. Moreover, our analysis did not include the *Appendix Platonica*, consisting of works too spurious to formally include in the Platonic Corpus: *Eryxias*, *Axiochus*, *On Justice*, and *On Virtue*. The only reason we did not examine these texts was that they were not readily available to download from online corpora. Given more resources, we would have liked to include them as part of the dubia dataset, although there is scholarly agreement that they are extraneous and inauthentic.

Additionally, there exists the limitation of processing power. Access to more powerful GPUs opens the door to future work via many routes. For this study, apart from the basic feedforward neural network model, the other two models were only trained on 20% of the available training data. Only 5 epochs were run per model, leaving much more hyperparameter tuning to be done. Perhaps, as the model size increases, we may find that the more sophisticated models, especially the Transformer, can capture the semantic nuances of Plato's writing across the entire corpus. To advance further, more sophisticated models such as BERT models and PLMs may perform better in this task. A logical development of this study would be to create a model that uses the entire Ancient Greek corpus to learn how to identify authors and to place an "unknown" label for texts that do not match any known author. Although such a task was beyond the scope of this project, doing so would allow the model to suggest who the spurious works are by, if they are not by Plato but resemble another author. Such further computational work in the field of philology would open new paths for computational linguists and classicists alike.

Acknowledgments

We would like to acknowledge our course instructor Tom McCoy and his teaching assistant Ayla Karakaş for equipping us with the conceptual and technical skills to complete this project and for their advice throughout.

References

Allen, W. S. (1987). *Vox Graeca: The Pronunciation of Classical Greek*. Cambridge University Press.

Bamman, D., & Burns, P. J. (2020). *Latin BERT: A Contextual Language Model for Classical Philology* (arXiv:2009.10053). arXiv.

Berti, M. (Ed.). (2019). *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*. De Gruyter Saur.

Brandwood, L. (1992). Stylometry and chronology. In R. Kraut (Ed.), *The Cambridge Companion to Plato* (pp. 90–120). Cambridge University Press.

Forcignanò, F., & Tempesta, S. M. (2023). Comparing Corpora, Rethinking Authenticity: Why Are Platonic Letters "Platonic"? In O. Alieva, D. Nails, & H. Tarrant (Eds.), *The Making of the Platonic Corpus* (pp. 203–221). Brill.

Ingenkamp, H. G. (1967). *Untersuchungen zu den pseudoplatonischen Definitionen*. O. Harrassowitz.

Irwin, T. H. (1992). Plato: The intellectual background. In R. Kraut (Ed.), *The Cambridge Companion to Plato* (pp. 51–89). Cambridge University Press.

Johnson, K. P., Burns, P. J., Stewart, J., Cook, T., Besnier, C., & Mattingly, W. J. B. (2021). The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages. In H. Ji, J. C. Park, & R. Xia (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations* (pp. 20–29). Association for Computational Linguistics.

Joyal, M. (2019). What Is Socratic about the Pseudo-Platonica? In C. Moore (Ed.), *Brill's Companion to the Reception of Socrates* (pp. 211–236). Brill.

Köntges, T. (2020). Measuring Philosophy in the First Thousand Years of Greek Literature. *Digital Classics Online*, 1–23.

Kostkan, J., Kardos, M., Mortensen, J. P. B., & Nielbo, K. L. (2023). OdyCy – A general-purpose NLP pipeline for Ancient Greek. In S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, & S. Szpakowicz (Eds.), *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 128–134). Association for Computational Linguistics.

Press, G. A. (Ed.). (2012). *The Continuum Companion to Plato*. Continuum.

Probert, P. (2006). *Ancient Greek Accentuation: Synchronic Patterns, Frequency Effects, and Prehistory*. Oxford University Press.

Riemenschneider, F., & Frank, A. (2023). Exploring Large Language Models for Classical Philology. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15181–15199). Association for Computational Linguistics.

Singh, P., Rutten, G., & Lefever, E. (2021). A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek. In S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, & S. Szpakowicz (Eds.), *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 128–137). Association for Computational Linguistics.

Tarrant, H. (2017). The Socratic Dubia. In A. Stavru & C. Moore (Eds.), *Socrates and the Socratic Dialogue*. Brill.

Wöckener-Gade, E., & Pöckelmann, M. (2023). Innovation in Loops: Developing Tools and Redefining Theories within the Project ‘Digital Plato’ (Digital Plato). In B. Schneider, B. Löffler, T. Mager, & C. Hein (Eds.), *Mixing Methods: Practical Insights from the Humanities in the Digital Age* (pp. 47–60). Bielefeld University Press.

Yamshchikov, I. P., Tikhonov, A., Pantis, Y., Schubert, C., & Jost, J. (2022). BERT in Plutarch’s Shadows. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 6071–6080). Association for Computational Linguistics.

Yousef, T., Palladino, C., & Shamsian, F. (2023). Classical Philology in the Time of AI: Exploring the Potential of Parallel Corpora in Ancient Language. In A. Anderson, S. Gordin, B. Li, Y. Liu, & M. C. Passarotti (Eds.), *Proceedings of the Ancient Language Processing Workshop* (pp. 179–192). INCOMA Ltd., Shoumen, Bulgaria.