|  |
| --- |
| **Instructions for \*ACL Proceedings** |
|  |
| **Michael Gao, Ryo Nagao**  Yale University  [m.gao@yale.edu](mailto:m.gao@yale.edu),  ryo.nagao@yale.edu |
|  |
|  |
|  |
|  |

Abstract

Abstract?

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*, *Clitophon*, *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, primarily stylometric, or statistical, such as those employing multivariate cluster analysis (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 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.

Methods

Data

All of the ancient Greek texts used in this project were downloaded from the Perseus Digital Library hosted by Tufts University.[[1]](#footnote-1) 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). 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.

* 1. Algorithms

abc

1. Results

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 of the test set, while 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).

Although all models achieved remarkably high accuracies in regard to the test set, the 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%).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Size 25** | **Size 50** | **Size 100** | **Size 200** | **Average** |
| FNN | 0.913121 | 0.957869 | 0.985235 | 0.993508 | 0.962433 |
| LSTM | 0.999901 | 0.999901 | 0.999901 | 0.999901 | 0.999901 |
| Transformer | 0.997812 | 0.997812 | 0.997812 | 0.997812 | 0.997812 |

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, *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%.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Text** | **Size 25** | **Size 50** | **Size 100** | **Size 200** | **Average** |
| Alcibiades 1 | 0.888573 | 0.946093 | 0.962266 | 0.984047 | 0.945245 |
| Alcibiades 2 | 0.802056 | 0.877395 | 0.918841 | 0.927316 | 0.881402 |
| Cleitophon | 0.816066 | 0.890534 | 0.940689 | 0.998488 | 0.911444 |
| Definitions | 0.317602 | 0.220795 | 0.132734 | 0.000410 | 0.167885 |
| Epinomis | 0.759075 | 0.858066 | 0.873438 | 0.991314 | 0.870473 |
| Hipparchus | 0.828508 | 0.861841 | 0.892269 | 0.930046 | 0.878166 |
| Letters | 0.739368 | 0.813222 | 0.849064 | 0.945643 | 0.836824 |
| Lovers | 0.879793 | 0.966958 | 0.995310 | 0.998058 | 0.960030 |
| Minos | 0.822500 | 0.903747 | 0.936301 | 0.981055 | 0.910901 |
| Theages | 0.888155 | 0.951387 | 0.984877 | 0.998438 | 0.955714 |
| **Average** | 0.774170 | 0.829004 | 0.848579 | 0.875482 | 0.831808 |

LSTM

The LSTM obtained the highest accuracy for the test set, 99.99% with every chunk size. Among the spurious works, *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%.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Text** | **Size 25** | **Size 50** | **Size 100** | **Size 200** | **Average** |
| Alcibiades 1 | 0.895217 | 0.931002 | 0.944577 | 0.967063 | 0.934465 |
| Alcibiades 2 | 0.800445 | 0.819981 | 0.780917 | 0.825705 | 0.806762 |
| Cleitophon | 0.859376 | 0.862431 | 0.887142 | 0.998191 | 0.901785 |
| Definitions | 0.263781 | 0.164125 | 0.108587 | 0.027610 | 0.141026 |
| Epinomis | 0.881690 | 0.956121 | 0.950774 | 0.984847 | 0.943358 |
| Hipparchus | 0.853926 | 0.886885 | 0.824437 | 0.869262 | 0.858628 |
| Letters | 0.698444 | 0.736132 | 0.756364 | 0.833279 | 0.756055 |
| Lovers | 0.895454 | 0.941334 | 0.993061 | 0.999993 | 0.957461 |
| Minos | 0.879873 | 0.892665 | 0.866075 | 0.936281 | 0.893724 |
| Theages | 0.914935 | 0.914026 | 0.926696 | 0.924735 | 0.920098 |
| **Average** | 0.794314 | 0.810470 | 0.803863 | 0.836697 | 0.811336 |

Transformer

The Transformer achieved higher accuracy for the test set than the FNN but lower than the LSTM, consistently showing a 99.78% rate. 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%.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Text** | **Size 25** | **Size 50** | **Size 100** | **Size 200** | **Average** |
| Alcibiades 1 | 0.850185 | 0.913857 | 0.917753 | 0.979432 | 0.915307 |
| Alcibiades 2 | 0.729873 | 0.782283 | 0.76039 | 0.88924 | 0.790447 |
| Cleitophon | 0.679582 | 0.786041 | 0.817005 | 0.998309 | 0.820234 |
| Definitions | 0.112889 | 0.073735 | 0.000476 | 0.000471 | 0.046893 |
| Epinomis | 0.713127 | 0.828842 | 0.859804 | 0.962932 | 0.841176 |
| Hipparchus | 0.802631 | 0.847109 | 0.876764 | 0.911755 | 0.859565 |
| Letters | 0.572273 | 0.639782 | 0.640313 | 0.853428 | 0.676449 |
| Lovers | 0.809909 | 0.934799 | 0.962090 | 0.995287 | 0.925521 |
| Minos | 0.773213 | 0.815251 | 0.827141 | 0.960195 | 0.843950 |
| Theages | 0.804119 | 0.878313 | 0.892242 | 0.988416 | 0.890773 |
| **Average** | 0.684780 | 0.750001 | 0.755398 | 0.853947 | 0.761031 |

1. 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%. This gives us reason to reconsider previous opinions on these supposedly spurious works, since they are computationally very similar to the works authentically ascribed to Plato.

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*.

The same is to an extent true for *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).

1. 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.

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 its 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 more experimentation with document segmentation would be useful for reaching even higher levels of accuracy. 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, 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. 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.

1. https://www.perseus.tufts.edu/hopper/. [↑](#footnote-ref-1)