Stylometric Classification of Ancient Greek Literary Texts by Genre

Efthimios Tim Gianitsos  
Department of Computer Science  
University of Texas at Austin

Thomas J. Bolt  
Department of Classics  
University of Texas at Austin

Pramit Chaudhuri  
Department of Classics  
University of Texas at Austin

Joseph P. Dexter  
Neukom Institute for Computational Science  
Dartmouth College

Abstract

Classification of texts by genre is an important application of natural language processing to literary corpora but remains understudied for premodern and non-English traditions. We develop a stylometric feature set for ancient Greek that enables identification of texts as prose or verse. The set contains over 20 primarily syntactic features, which are calculated according to custom, language-specific heuristics. Using these features, we classify almost all surviving classical Greek literature as prose or verse with >97% accuracy and F1 score, and further classify a selection of the verse texts into the traditional genres of epic and drama.

1 Introduction

Classification of large corpora of documents into coherent groups is an important application of natural language processing. Research on document organization has led to a variety of successful methods for automatic genre classification (Stamatatos et al., 2000; Santini, 2007). Computational analysis of genre has most often involved material from a single source (e.g., a newspaper corpus, for which the goal is to distinguish between news articles and opinion pieces) or from standard, well-curated test corpora that contain primarily non-literary texts (e.g., the Brown corpus or equivalents in other languages) (Kessler et al., 1997; Petrenz and Webber, 2011; Amasyali and Diri, 2006).

Notions of genre are also of substantial importance to the study of literature. For instance, examination of the distinctive characteristics of various forms of poetry dates to classical Greece and Rome (for instance, by Aristotle and Quintilian) and remains an active area of humanistic research today (Frow, 2015). A number of computational analyses of literary genre have been reported, using both English and non-English corpora such as classical Malay poetry, German novels, and Arabic religious texts (Tizhoosh et al., 2008; Kumar and Minz, 2014; Jamal et al., 2012; Hettinger et al., 2015; Al-Yahya, 2018). However, computational prediction of even relatively coarse generic distinctions (such as between prose and poetry) remains unexplored for classical Greek literature.

Encompassing the epic poems of Homer, the tragedies of Aeschylus, Sophocles, and Euripides, the historical writings of Herodotus, and the philosophy of Plato and Aristotle, the surviving literature of ancient Greece is foundational for the Western literary tradition. Here we report a computational analysis of genre involving the whole of the classical Greek literary tradition. Using a custom set of language-specific stylometric features, we classify texts as prose or verse and, for the verse texts, as epic or drama with >97% accuracy. An important advantage of our approach is that all of the features can be computed without syntactic parsing, which remains in an early phase of development for ancient Greek. As such, our work illustrates how computational modeling of literary texts, where research has concentrated overwhelmingly on modern English literature (Elson et al., 2010; Elsner, 2012; Bamman et al., 2014; Chaturvedi et al., 2016; Wilkens, 2016), can be extended to premodern, non-Anglophone traditions.

2 Stylometric feature set for ancient Greek

The feature set is composed of 23 features covering four broad grammatical and syntactical categories. The majority of the features are function or non-content words, such as pronouns and syntactical markers; a minority concern rhetorical functions, such as questions and uses of superla-
Table 1: Full set of ancient Greek stylometric features.

| Feature                              | Genre | Precision | Recall |
|--------------------------------------|-------|-----------|--------|
| Pronouns and non-content adjectives  | verse | 0.96      | 0.96   |
|                                      | prose | 0.97      | 1      |
| demonstrative pronouns               | verse | 1         | 0.93   |
|                                      | prose | 1         | 1      |
| selected indefinite pronouns         | verse | 0.97      | 0.96   |
|                                      | prose | 1         | 1      |
| personal pronouns                    | verse | 1         | 0.89   |
|                                      | prose | 1         | 1      |
| reflexive pronouns                   | verse | 1         | 0.85   |
|                                      | prose | 1         | 1      |

Table 2: Error analysis of non-exact features. The features are numbered as in Table 1.

in standard studies of English stylistics, such as pronouns, others are specific to ancient Greek. Attention to language-specific features enhances stylometric methods developed for the English language and not directly transferable to languages possessing a different structure (Rybicki and Eder, 2011; Kestemont, 2014). Greek particles, for example, are uninflected adverbs used for a wide range of logical and emotional expressions; in English their equivalent meaning is often expressed by a phrase or, in speech, tone. In order to avoid significant problems arising from dialectical variation, including a large increase in homonyms, we restrict features to the Attic dialect, in which the majority of classical Greek texts were composed. Many features are computed by counting all inflected forms of the appropriate word(s), which can be found in any standard ancient Greek textbook or grammar such as Smyth (1956). A detailed description of the methods for computing the features is given in Appendix A.

Calculation of five features relies on heuristics to disambiguate between words of similar morphology. (All other features can be calculated exactly.) To assess the effectiveness of these heuristics, we hand-annotate the five features in a representative sub-corpus containing three verse (Homer’s Odyssey 6, Quintus of Smyrna’s Posthomerica 12, and Euripides’ Cyclops) and two prose (Lysias 7 and Plutarch’s Caius Gracchus) texts. Table 2 lists the precision and recall of each feature on the aggregated verse and prose texts. In every instance, the precision is > 0.95 and the recall is > 0.85.
3 Experimental setup

3.1 Dataset

We use a corpus of ancient Greek text files, which was assembled by the Perseus Digital Library and further processed by Tesserae Project (Crane, 1996; Coffee et al., 2012). A full list of texts is provided in Appendix B. Each file typically contains either an entire work of literature (e.g., a play or a short philosophical treatise) or one book of a longer work (e.g., Book 1 of Homer’s *Iliad*). 29 files are composites of multiple books included elsewhere in the Tesserae corpus and are omitted from our analysis, leaving 751 files. In total, this corpus contains essentially all surviving classical Greek literature and spans from the 8th century BCE to the 6th century CE.

For our first experiment, we hand-annotate the full set of texts as prose (610 files) or verse (141 files) according to standard conventions (Appendix B). For the second experiment, we hand-annotate the verse texts as epic (82 files) and drama (45 files), setting aside 14 files that contain poems of other genres (Appendix C).

3.2 Feature extraction

All text processing is done using Python 3.6.5. We first tokenize the files from the Tesserae corpus into either words or sentences using the Natural Language Toolkit (NLTK; v. 3.3.0) (Bird et al., 2009). For sentence tokenization, we use the PunktSentenceTokenizer class of NLTK Greek (Kiss and Strunk, 2006). After tokenization, the features are calculated either by tabulating instances of signal n-grams or (for length-based features) counting characters exclusive of whitespace, as described in Appendix A.

3.3 Supervised learning

All supervised learning is done using Python 3.6.5. For each experiment, we use the scikit-learn (v. 0.19.2) implementation of the random forest classifier. A full list of hyperparameters and other settings is given in Appendix D. For each binary classification experiment (prose vs. verse and epic vs. drama), we perform 400 trials of stratified 5-fold cross-validation; each trial has a unique combination of two random seeds, one used to initialize the classifier and the other to initialize the data splitter. Feature rankings are determined by the average Gini importance across the 400 trials.

4 Results

4.1 Prose vs. verse classification

Using the workflow described in Section 3.3, we classify each of the literary texts in the corpus as prose or verse. Table 3 lists the accuracy and weighted F1 score for a sample cross-validation trial, along with the mean for that trial and overall mean across the 400 trials. We find that the texts can be classified as prose or verse with extremely high accuracy using the set of 23 stylometric features and that, despite the small size of the corpus, classifier performance is robust to the choice of cross-validation partition. The five highest-ranked features are given in Table 4. Outside of these five, no other feature has a Gini importance of $>0.05$. All five features predominate in prose rather than poetry, of which three are pronouns or pronominal adjectives. The sustained discussions commonly found in various prose genres may favor the use of pronouns to avoid extensive repetition of nouns and proper names. The high ranking of conjunctions is plausibly connected to the longer sentences characteristic of most prose (mean length 205 characters, compared to 166 characters for poetry).

| Feature       | Gini  | S.D.  |
|---------------|-------|-------|
| ἀυτός         | 0.209 | 0.074 |
| conjunctions  | 0.159 | 0.062 |
| demonstrative pronouns | 0.121 | 0.057 |
| reflexive pronouns | 0.118 | 0.049 |
| μέν          | 0.0623 | 0.029 |

Table 4: Feature rankings for prose vs. verse classifier.
### 4.2 Classification of poems as epic or drama

The genres of epic and drama are in certain respects quite distinct: they differ in length and poetic meter, and the vocabulary of Aristophanes’ comic plays is unlike either epic or tragedy. In other aspects of form and content, however, they have much in common, including passages of direct speech, high register diction, and mythological subject matter. The playwright Aeschylus is even reported to have described his tragedies as “slices from the great banquets of Homer” (Athenaeus, *Deipnosophistae* 8.347E). The similarities between epic and drama thus present an intuitively greater challenge for classification.

Table 5 summarizes the results of the epic vs. drama experiment, for which we achieve performance comparable to that of the prose vs. verse experiment. Table 6 lists the top features, which reflect several important differences between the genres. The most important feature - sentence length - highlights the relatively shorter sentences of drama compared to epic, which can be explained at least in part by the rapid exchanges between speakers that occur throughout both tragedy and comedy. Although sentence length is a feature that can be affected by modern editorial practice, the difference between drama and epic on this score is sufficiently large that it cannot be explained by variations in editorial practice alone (<80 characters/sentence on average across dramatic texts, >150 characters/sentence for epic). The importance of demonstrative pronouns, ranked second, plausibly captures a different side of drama - the habit of characters referring, often indexically, to persons or objects in the plot (e.g., ἐκεῖνος ὁντός ἐμί, ekeinos houtos eimi, “I am that very man,” Euripides, *Cyclops* 105, which uses two demonstrative pronouns in succession). Another typical characteristic of dramatic plot and dialogue accounts for the third highly-ranked feature - interrogative sentences - since both tragedies and comedies often show characters in a state of uncertainty or ignorance, or making inquiries of other characters. Although many of the features in the full set are correlated (e.g., sentence length and various markers of subordinate clauses), none of the top 5 plausibly are, suggesting that the analysis identifies a diverse set of stylistic markers for epic and drama.

### 4.3 Misclassifications

For epic vs. drama, no text is misclassified in more than 12% of the trials. For prose vs. verse, only five texts are misclassified in >50% of the trials (Demades, *On the Twelve Years*; Dionysius of Halicarnassus, *De Antiquis Oratoribus Reliquiae* 2; Plato, *Epistle* 1; Aristotle, *Virtues and Vices*; Sophocles, *Ichneutae*). Most of the common misclassifications result from highly fragmentary or short texts. Almost half the speech of Demades, for example, contains short or incomplete sentences. The misclassified text of Dionysius of Halicarnassus amounts to only a few unconnected sentences; Sophocles’ *Ichneutae* (the only verse text misclassified in over half the trials) is also fragmentary. The third most frequently misclassified text, Plato’s *First Epistle*, in fact highlights the classifier’s effectiveness, as it contains several verse quotations, which (given the short length of the text) plausibly account for the error.

### 5 Conclusion

In this paper, we demonstrate that ancient Greek literature can be classified by genre using a straightforward supervised learning approach and stylometric features calculated without syntactic parsing. Our work suggests a number of natural follow-up analyses, especially extension of the experiments to encompass the full range of tradi-

|                  | Accuracy (%) | Weighted F1 (%) |
|------------------|--------------|-----------------|
| Fold 1           | 92.3         | 92.0            |
| Fold 2           | 100          | 100             |
| Fold 3           | 100          | 100             |
| Fold 4           | 100          | 100             |
| Fold 5           | 100          | 100             |
| Mean             | 98.5         | 98.4            |
| S.D.             | 3.4          | 3.6             |
| Overall          | 99.8         | 99.8            |
| S.D.             | 0.9          | 0.9             |

Table 5: Performance of epic vs. drama classifier for ancient Greek poetry.

| Feature                              | Gini   | S.D.  |
|--------------------------------------|--------|-------|
| mean sentence length                 | 0.186  | 0.12  |
| demonstrative pronouns               | 0.155  | 0.095 |
| interrogative sentences              | 0.127  | 0.12  |
| ὁς                                   | 0.117  | 0.11  |
| variance of sentence length          | 0.0952 | 0.075 |

Table 6: Feature rankings for epic vs. drama classifier.
tional prose genres (such as historiography, philosophy, and oratory) and application of the feature set to other questions in classical literary criticism. In addition, we hope that our heuristic approach will motivate and inform analogous work on other premodern traditions for which natural language processing research remains at an early stage.

Acknowledgments

This work was conducted under the auspices of the Quantitative Criticism Lab (www.qcrit.org), an interdisciplinary group co-directed by P.C. and J.P.D. and supported by a National Endowment for the Humanities Digital Humanities Start-Up Grant (grant number HD-10 248410-16) and an American Council of Learned Societies (ACLS) Digital Extension Grant. T.J.B. was supported by an Engaged Scholar Initiative Fellowship from the Andrew W. Mellon Foundation, P.C. by an ACLS Digital Innovation Fellowship and a Mellon New Directions Fellowship, and J.P.D. by a Neukom Fellowship.

References

Maha Al-Yahya. 2018. Stylometric analysis of classical Arabic texts for genre detection. The Electronic Library, 36:842–855.

M. Fatih Amasyali and Banu Diri. 2006. Automatic Turkish text categorization in terms of author, genre and gender. In Christian Kop, Günther Fliedl, Heinrich C. Mayr, and Elisabeth Mtais, editors, Natural Language Processing and Information Systems, pages 221–226. Springer-Verlag, Berlin.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, pages 370–379.

Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural Language Processing with Python. O’Reilly Media.

Snigdha Chaturvedi, Hal Daumé III, Shashank Srivastava, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In Proceedings of the Thirtyieth AAAI Conference on Artificial Intelligence, pages 2704–2710.

Neil Coffee, J.-P. Koenig, Shakhthi Poornima, Roelant Ossewaarde, Christopher Forstall, and Sarah Jacobson. 2012. Intertextuality in the digital age. Transactions of the American Philological Association, 142:383–422.

Gregory Crane. 1996. Building a digital library: The Perseus Project as a case study in the humanities. In Proceedings of the First ACM International Conference on Digital Libraries, pages 3–10.

Micha Elsner. 2012. Character-based kernels for novelistic plot structure. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 634–644.

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. 2010. Extracting social networks from literary fiction. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 138–147.

John Frow. 2015. Genre. Routledge, London and New York.

Vanessa B. Gorman and Robert J. Gorman. 2016. Approaching questions of text reuse in ancient greek using computational syntactic stylometry. Open Linguistics, 2:500–510.

Lena Hettinger, Martin Becker, Isabella Reger, Fotis Iannidis, and Andreas Hotho. 2015. Genre classification on German novels. In 2015 26th International Workshop on Database and Expert Systems Applications, pages 138–147.

James M. Hughes, Nicholas J. Fotia, David C. Krakauer, and Daniel N. Rockmore. 2012. Quantitative patterns of stylistic influence in the evolution of literature. Proceedings of the National Academy of Sciences USA, 109:7682–7686.

Noraini Jamal, Masnizah Mohd, and Shahrul Azman Noah. 2012. Poetry classification using support vector machines. Journal of Computer Science, 8:1411–1416.

Brett Kessler, Geoffrey Numberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pages 32–38.

Mike Kestemont. 2014. Function words in authorship attribution. From black magic to theory? In Proceedings of the 3rd Workshop on Computational Linguistics for Literature @ EACL 2014, pages 59–66.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. Computational Linguistics, 32:485–525.

Vipin Kumar and Sonajharia Minz. 2014. Poem classification using machine learning approach. In Proceedings of the Second International Conference on Soft Computing for Problem Solving, pages 675–682.

Philipp Petrenz and Bonnie Webber. 2011. Stable classification of text genres. Computational Linguistics, 37:385–393.
A Details of stylometric features for ancient Greek

A.1 Pronouns and non-content adjectives
- ἀλλος (allos, “other”) is computed by counting all inflected forms of ἀλλος, -ης, -ο.
- αὐτος (autos, “self” or “him/her/it”) is computed by counting all inflected forms of αὐτος, -ής, -ός.
- Demonstrative pronouns are computed by counting all inflected forms of the three Greek demonstrative pronouns οὕτως, αὕτη, τοῦτο (houtos, haute, touto, “this”), ὅδε, ἴδε, τώδε (hode, hede, tode, “this”), and ἐκείνος, ἐκείνη, ἐκείνο (ekeinos, ekeine, ekeino, “that”).
- Selected indefinite pronouns are computed by counting all inflected forms of τις, τις, τι (tis, tis, ti, “any”) in non-interrogative sentences. Interrogative sentences are excluded because the Greek interrogative pronoun (τις) is often identical in form to the indefinite pronoun.
- Personal pronouns are computed by counting all inflected forms of the pronouns ἐγώ (ego, “I”) and σύ (su, “you”).
- Reflexive pronouns are computed by counting all inflected forms of ἐμαυτός (emautou, “he himself”).

A.2 Conjunctions and particles
- Conjunctions are computed by counting all instances of the common conjunctions τε, τ’ (te or t, “and”), καί, καί (kai, “and”), ἀλλά, ἀλλα (alla, “but”), καίτοι (kaitoi, “and indeed”), οὐδέ, οὐδέ, οὐ (oude or oud, “and not”), μηδέ, μηδέ, μηδό (mede or med, “and not”), οὔτε, οὔτε (oute or out, “and not”), μήτε, μήτε (mete or met, “and not”), and ἐν (e, “or”).
- μέν (men, “indeed”) is computed by counting all instances of μέν and μεν.
- Particles are computed by counting all instances of ἄν, ἄν (an, a particle used to express uncertainty or possibility), ἀρχα (ara, “then”), γέ, γ’ (ge or g, “at least”), δε, δε (d or de, “but”), δή, δή (de, “indeed”), έως (heos, “until”), χε, πο, κε (k, ke, ken, a particle used to express uncertainty or possibility), μά (ma, used in oaths and affirmations, “by”), μέν, μέν (men, “indeed”), μέντοι (mentoi, “however”), μην, μην (men, “truly”), μον (mon, “surely not”), μή, μή, μή (mete or met, “and not”), μήτε, μήτε (mete or met, “and not”), and τοι (toi, “let me tell you”).

A.3 Subordinate clauses
- Circumstantial markers are computed by counting all instances of ἐπειτα, ἐπειτ’ (epeita or epeit, “then”), ὁμοιος (homos, “all the same”), ὁμοιος (homos, “equally”), καίπερ (kaiper, “although”), and ἀτέ, ἀτ’ (hate or hat, “seeing that”).
- Conditional markers are computed by counting all instances of ei, ei, ei, εἴναι, and εἴναι (ei, ei, ei, ean, ean, all translated “if”).
- ἦναι (hina, an adverb of place often translated “where” or a conjunction indicating purpose often translated “in order that”) is computed by counting all instances of ἦναι and ἦν (hin).
- ὤπος (hopos, an adverb of manner often translated “how” or a conjunction indicating purpose often translated “in order that”) is computed by counting all instances of ὤπος.
• Fraction of sentences with a relative clause is determined by counting sentences that have one or more of the inflected forms of the Greek relative pronouns ὃς, ἣ, ὅ (hos, he, ho, “who” or “which”).

• Temporal and causal markers are computed by counting all instances of μέχρι (mekri, “until”), ἐως (heos, “until”), πρὶν (prin, “before”), ἐπεὶ (pepi, “when”), ἐπειδή (peide, “after” or “since”), ἐπειδάν (peidean, “whenever”), ὃτε (hote, “when”), and ὅταν (hotan, “whenever”).

• ὅστε (hoste, a conjunction used to indicate a result, “so as to”) not preceded by ἤ is calculated by counting all instances of ὅστε not immediately preceded by ἤ. This limitation is imposed to exclude instances in which ὅστε is part of a comparative phrase.

• The mean length of relative clauses is determined by counting the number of characters between each relative pronoun and the next punctuation mark.

A.4 Miscellaneous

• Interrogative sentences are computed by counting all instances of “;” (the Greek question mark).

• Regular superlatives adjectives are computed by counting all instances of -τατος, -τατοι, -τατος, -τατον, -τατοις, -τατης, -τατην, -τατων, -τατοις, -τατης, -τατην, -τατων, -τατοις, -τατης, -τατην, -τατοις, -τατης, -τατην, -τατων, and τατε at word end. One inflected form, -ταται, is excluded so as to avoid confusion with the Homeric third person singular middle/passive indicative verb ending -σαι. This method does not detect certain irregular superlatives, such as ἄριστος (aristos, “best”) or πρῶτος (protos, “first”), which would be significantly harder to disambiguate from non-superlative forms.

• Sentences with ὅ exclamations is determined by identifying sentences that have at least one instance of ὅ (o, “O”), a Greek exclamation.

• ὅς (hos, an adverb of manner often translated “how” or a conjunction often translated as “that,” “so that,” or “since,” among several other possibilities) is computed by counting all instances of ὅς.

• Mean and variance of sentence length is determined by counting the number of characters in each tokenized sentence (see Section 3.2 of main paper).

B List of ancient Greek literary texts

Verse texts: Aeschylus, Agamemnon, Eumenides, Libation Bearers, Persians, Prometheus Bound, Seven Against Thebes, and Suppliant Women; Apollonius, Argonautica; Aristophanes, Acharnians, Birds, Clouds, Ecclesiazusae, Frogs, Knights, Lysistrata, Peace, Plutus, Thesmophoriazusae, and Wasps; Bacchylides, Dithyrambs and Epinicians; Bion of Phlosa, Epitaphius, Epithalamium, and Fragmenta; Callimachus, Epigrams and Hymns; Colluthus, Rape of Helen; Euripides, Alcestis, Andromache, Bacchae, Cyclops, Electra, Hecuba, Helen, Heracleidae, Heracles, Hippolytus, Ion, Iphigenia at Aulis, Iphigenia in Tauris, Medea, Orestes, Phoenixes, Rheus, Suppliant, and Trojan Women; Homer, Iliad and Odyssey; Lucian, Podraga; Lycothron, Alexandra; Nonnus of Panopolis, Dionysiaca; Oppian, Halieutica; Oppian of Apamea, Cynegetica; Pindar, Isthmians, Nemeans, Olympians, and Pythians; Quintus Smyrnaeus, Fall of Troy; Sophocles, Ajax, Antigone, Electra, Ichneutae, Oedipus at Colonus, Oedipus Tyrannus, Philoctetes, and Trachiniae; Theocritus, Epigrams; Tryphiodorus, The Taking of Ilion.

Prose texts: Achilles Tatius, Leucippe et Clitophon; Aelian, De Natura Animalium, Epistulae Rusticae, and Varia Historia; Aelius Aristides, Ars Rhetorica and Orationes; Aeschines, Against Ctesiphen, Against Timarchus, and On the Embassy; Andocides, Against Alcibiades, On His Return, On the Mysteries, and On the Peace; Antiphon, Against the Stepmother for Poisoning, First Tetralogy, Second Tetralogy, Third Tetralogy, On the Murder of Herodes, and On the Choreutes; Apollodorus, Epitome and Library; Appian, Civil Wars; Aretaeus, Curatione Auctorum Morbum and Signorum Auctorum Morbum; Aristotle, Constitution, Economics, Eudemian Ethics, Metaphysics, Nicomachean Ethics, Poetics, Politics, Rhetoric, and Virtues and Vices; Athenaeus, Deipnosophists; Barnabas, Barnabae Epistulae; Basil of Caesarea, De Legendis et Epistulae; Callistatus, Statuarum Descriptiones; Chariton, De
Chaerea; Clement, Exhortation, Protrepticus, and Quis Dis Salvetur; Demades, On the Twelve Years; Demetrius, Elocutiones; Demosthenes, Against Androtion, Against Apaturius, Against Aphobus, Against Aristocrates, Against Aristotian, Against Boeotus, Against Callicles, Against Callippus, Against Conon, Against Dionysodorus, Against Eubulides, Against Evergus and Mnesibulus, Against Locritus, Against Leochares, Against Leptines, Against Macartatus, Against Midias, Against Nausimachus and Xenepothes, Against Neaera, Against Nicostratus, Against Olympiodorus, Against Onetor, Against Pantaenetus, Against Phaenippus, Against Phormio, Against Polycles, Against Spudias, Against Stephanus, Against Theocrines, Against Timocrates, Against Timotheus, Against Zenothemis, Erotic Essay, Exordia, For Phormio, For the Megalopitans, Funeral Speech, Letters, Olynthiac, On Organization, On the Accession of Alexander, On the Chersonese, On the Crown, On the False Embassy, On the Halonnesus, On the Liberty of the Rhodians, On the Navy, On the Peace, On the Trierarchic Crown, Philip, Philipippic, and Reply to Philip; Dinarchus, Against Aristotian, Against Demosthenes, and Against Philocles; Dionysius of Halicarnassus, Ad Ammaeum, Antiquitates Romanae, De Antiquis Oratoribus, De Compositione Verborum, De Demosthene, De Dinarcho, De Isaeo, De Isocrate, De Lysia, De Thucydid, De Thucydivdides Idiomatibus, Epistula ad Pompeium, and Libri Secundi de Antiquis Oratoribus Reliquiae; Epictetus, Discourses, Enchiridion, and Fragments; Euclid, Elements; Eusebius of Caesarea, Historia Ecclesiastica; Flavius Josephus, Antiquitates Judaicae, Contra Apionem, De Bello Judaico, and Vita; Galen, Natural Faculties; Herodorus, Histories; Hippocrates, De Aere Aquis et Locis, De Alimento, De Morbis Popularibus, De Prisca Medicamina, and Jusjurandum; Hyperiades, Against Athenogenes, Against Demosthenes, Against Philopides, Funeral Oration, In Defense of Euxenippus, and In Defense of Lycophron; Isaeus, Speeches; Isocrates, Letters and Speeches; Lucian, Abdicatus, Adversus Indoctum et Libros Multos Ementem, Alexander, Anacharsis, Apologia, Bacchus, Bis Accusatus Sive Tribunalia, Columniae Non Temere Credendum, Catapulas, Contemplantes, De Astrologia, De Domus, De Luctu, De Mercede, De Morte Peregrini, De Parasito Sive Artem Esse Parsiticam, De Sacrificiis, De Saluta-
Lycurgus and Numa, Comparison of Lysander and Sulla, Comparison of Nicias and Crassus, Comparison of Pelopidas and Marcellus, Comparison of Pericles and Fabius Maximus, Comparison of Philopoemen and Titus, Comparison of Sertorius and Eumenes, Comparison of Solon and Publicola, Comparison of Theseus and Romulus, Comparison of Timoleon and Aemilius, Conjugalia Praecepta, Consolatio ad Apollonium, Consolatio ad Uxorem, Crassus, De Alexandri Magni Fortuna aut Virtute, De Amicorum Multitudine, De Amore Prolis, De Animae Procreatione in Timaeo, De Capienda Ex Inimicis Utilitate, De Cohibenda Ira, De Communibus Notitiis Adversus Stoicos, De Cupiditate Divitiarum, De Curiositate, De Defectu Oraculorum, De E Delphos, De Esu Carnium, De Exilio, De Faciaceae Quae in Orbe Lunae Apparet, De Fato, De Fortuna, De Fortuna Romanorum, De Fraterno Amore, De Garrullitate, De Genio Socratis, De Gloria Atheniensium, De Herodoti Malignitate, De Invidia et Odio, De Iside et Osiride, De Liberis Educandis, De Primo Frigido, De Pythiae Oraculis, De Recta Ratione Audendi, De Se Ipsum Citra Invidiam Laudando, De Sera Numinis Vindicta, De Sollertia Animalium, De Stoicorum Repugnantis, De Superstitione, De Tranquillectate Animi, Demetrius, Epitome Argumenti Stoicos, Epitome Libri de Animae Procreatione, Fabius Maximus, Galba, Instituta Laconica, Laecaenarum Apopthegmata, Lucullus, Lycurgus, Marcellus, Marcus Cato, Maxime Cum Principibus Philosopher Esse Diserendum, Mulierum Virtutes, Nicias, Non Posse Suavitier Vivi Secundum Epicurum, Numa, Otho, Parallela Minora, Pelopidas, Pericles, Philopoemen, Phocion, Platonicae Quaestiones, Pompey, Praecepta Geren-dae Reipublicae, Publicola, Pyrrhus, Quaestiones Convivales, Quaestiones Graecae, Quaestiones Naturales, Quaestiones Romanae, Quomodo Adolescentes Poetas Audire Debeat, Quomodo Adulator ab Amico Internoscatur, Quomodo Quis Suos in Vrute Sentiat Profectus, Regum et Imperatorum Apopthegmata, Romulus, Septem Sapien-tium Convivium, Sertorius, Solon, Sulla, Themistocles, Theseus, Tiberius Gracchus, Timoleon, Titus Flamininus, and Vitae Decem Oratorum; Polybius, Histories; Pseudo-Plutarch, De Musica and Placita Philosophorum; Strabo, Geography; Thucydides, Peloponnesian War; Xenophon, Anabasis.

C Genre labels for verse texts

Epic: Apollonius, Argonautica; Colluthus, Rape of Helen; Homer, Iliad and Odyssey; Nonnus of Panopolis, Dionysiacs; Oppian, Halieutica; Oppian of Apamea, Cynegetica; Quintus Smyrnaeus, Fall of Troy; Tryphiodorus, The Taking of Ilios.

D Parameters for random forest models

For all experiments, the parameters for the scikit-learn random forest classifier are set to ‘bootstrap’: True, ‘class_weight’: None, ‘criterion’: ‘gini’, ‘max_depth’: None, ‘max_features’: ‘auto’, ‘max_leaf_nodes’: None, ‘min_impurity_decrease’: 0.0, ‘min_impurity_split’: None, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘min_weight_fraction_leaf’: 0.0, ‘n_estimators’: 10, ‘n_jobs’: 1, ‘oob_score’: False, ‘random_state’: 0, ‘verbose’: 0, ‘warm_start’: False.