Classification for Catharsis:
Predicting the Authorship of Ancient Greek Tragedy

Giovanna Giuliano

Advisor
Professor Isabelle Beaudry

Thesis Committee
Professor Isabelle Beaudry
Professor Timothy Chumley
Professor Ben Gebre-Medhin

A thesis submitted to the Department of Mathematics and Statistics in partial fulfillment of the requirements for the degree of Bachelor of Arts in Statistics.

Department of Mathematics and Statistics
Mount Holyoke College
South Hadley, MA 01075
May 2025

Abstract

The issue of dubious authorship, persistent for centuries in discussions of classical literature, has been enhanced in recent years by the use of machine learning techniques. Statistical classifiers such as naive Bayes, support vector machines, and logistic regression have shown remarkable accuracy in ascribing documents of varying lengths to the right authors and have thus been central in parsing the extent of potential interpolations in ancient Greek literature. This thesis seeks to expand on this recent trend by designing and applying a new algorithm based on generalized linear mixed models, evaluating its performance against standard authorship attribution methodology. In particular, these models will be run on select works by the dramatist Euripides that have not yet been analyzed through this statistical lens.

Acknowledgements

Thank you to my advisor, Professor Isabelle Beaudry, for working with me this year. Her guidance and kindness were invaluable in making something as daunting as a senior thesis approachable and way less scary. I hadn't heard of mixed-effects models before she mentioned them to me in September, so it's hard to imagine what this thesis would have looked like otherwise. Thank you as well to Professors Tim Chumley and Ben Gebre-Medhin for serving on my thesis committee, and to the Department of Mathematics & Statistics as a whole for providing such a supportive learning environment these past four years.

Thank you to Professor Geoffrey Sumi, with whom I first delved into classics my sophomore year. This thesis would not exist without that Ancient Greece class, and its shape was made possible by his insights on a mini version of this work done in his Ancient Rome class. Thank you to Professor Laurie Tupper, who took me on for research last summer. That experience taught me what the digital humanities can look like and about the art of using stats on originally non-quantitative things. It was the spark of inspiration I needed to mix statistics and ancient history in the first place.

Thank you to Luz, with whom I battled nighttime thesis writing sessions; to Rylee, who first put me on Greek tragedy with Euripides' *Medea*; to Eonbi and the greater Posse community for cheering me on and listening to me rant and ramble. And lastly, a special thanks to my family for being ever encouraging from many states away.

Contents

1 Introduction **1**

2 Literature Review **4**

2.1 Authorship Analysis 4

2.1.1 Topic Influence 7

2.2 Authenticity and Authorship of Greek Texts 9

2.2.1 Spurious Texts of Interest 11

3 Methodology **14**

3.1 Text Representation 14

3.1.1 Information Gain 15

3.1.2 Principal Components Analysis 17

3.2 Support Vector Machines 18

3.3 Mixed Models 21

3.3.1 Latent Dirichlet Allocation 24

4 Results **26**

4.1 Authorship Prediction 26

4.2 Selection of the Predictors 30

4.3 Selection of the Topics 32

5 Conclusion **34**

List of Figures

3.1 A hyperplane dividing a data set. The dotted lines indicate the outer edges of the margin, with three support vectors sitting on, and therefore dictating, where the hyperplane lies. Image from James et al. (2021). 19

3.2 On the left, a hyperplane dividing a data set with two parameters with maximal distance from the nearest points. On the right, the addition of a new point within the blue class completely shifts the location of the boundary. Images from James et al. (2021). 20

4.1 Accuracy rates after 10-fold cross validation of SVM using 3-grams, 4-grams, and words, from 100 features to 900. 31

4.2 AIC and BIC values obtained from PCA applied to a fixed logistic regression model, from $M = 1$ to $M = 25$. 32

4.3 Values yielded from the four metrics of `FindTopicsNumber` for $T = \{2, 3, \ldots, 50\}$. The metrics are organized by whether they should be maximized or minimized. A low value for CaoJuan2009 or Arun2010, as well as a high value for Griffiths2004 or Deveaud2014, signifies a distribution of topics such that each topic is as different from the others as possible given their associated words. 33

List of Tables

4.1 True positive and false positive rates for all five models after 10-fold cross validation on the training set. 27

4.2 Predictions for each of the seven test chunks using each model. For SVM, the response is binary: 0 (non-Euripidean) and 1 (Euripidean). For logistic regression, the response is probability of authenticity. 29

1 Introduction

Authorship analysis is the process of ascertaining the author of an unidentified or disputed text based on the author's own writing style and that of their peers, as evidenced by their established work. Such questions may arise in copyright disputes, forensic contexts, or in literary research (Stamatatos 2009). Developments in natural language processing and data science within the past thirty years have led to this type of study becoming easier, more powerful, and more common. One example of where new, "non-traditional" authorship analysis has been suitably applied is in the study of Greco-Roman literature. Centuries of manual copying have allowed for much interpolation in the manuscripts of ancient texts, but analyses of authorship attribution and verification have shown promise in clarifying and offering new perspectives to long-standing debates of authenticity. Of particular attention have been the collection of extant plays by the three Athenian tragedians: Aeschylus, Euripides, and Sophocles. Manousakis (2020) and Stamatatos (2018; 2023; 2024) have done extensive work in authorship attribution on a few Greek tragedies with spurious segments, though much remains to be explored.

Authorship analysis is comprised of two primary components: (1) the descriptive or predictive algorithms and techniques used to identify authorial style and (2) the mode in which the texts at hand have been represented for analysis by the tools in (1). The former is the domain of machine learning

tools such as k-nearest neighbors, support vector machines, neural networks, and linear regression (Koppel et al. 2009). The latter includes decisions of representation: whether the data should contain lexical, grammatical, or semantic features, and how these features should be identified and counted.

A common concern in obtaining features lies in ensuring that the data indeed capture an author's style and not the specific topics discussed in the collected texts. This can happen if the features contain too much content-specific vocabulary. Techniques involve minimizing the inclusion of such specific features entirely or combining them with content-free features in some capacity. The risk posed by including topic-related information lies in the fact that two documents composed by two different authors can be expected to have more commonality in their features if they cover similar topics than if they do not, which can yield misleading predictions. The extant body of Greek tragedy, aforementioned, can have noticeable commonality based on topic. Take, for example, the *Libation Bearers* of Aeschylus and both plays called *Electra* by Euripides and Sophocles, which focus on the same mythological event, though with some variation on the specifics (Hall 2010). Even when the stories told are different, similar characters or themes appear across their plays.

This study seeks to address the issue of unwanted topic influence by proposing the use of a generalized linear mixed-effects regression model for authorship verification of Greek tragedy. This kind of model is suitable for data with a clustered structure where including the cluster-based effects into

the parameters is computationally infeasible; we instead estimate the cluster-based effects as random effects estimated jointly with the fixed predictors. We propose treating the shared topics observed across plays as clusters to which the texts pertain, identified by a separate topic model—something which has not been done in authorship analysis before. We evaluate the performance of this method against the common authorship analysis models, and use them to verify the authorship of two texts attributed to Euripides: *Iphigenia at Aulis* and *The Phoenician Women*. The former contains spurious lines both agreed upon and debated on by classicists, and the latter is filled with many one-to-two line interpolations throughout. Thus we can use one to evaluate how well the models identify known unoriginal text and the other to evaluate how well they identify brief interpolations embedded in original material. In either case we offer some clarity to discussions regarding two plays that have not yet received computational attention.

Section 2 of this paper provides an overview of existing literature regarding the models and texts examined in this study. Section 3 details the methods employed and Section 4 presents and discusses the observed results. Finally, in Section 5, we conclude with a discussion of the limitations of this study and of potential avenues of future work.

2 Literature Review

2.1 Authorship Analysis

Authorship analysis that uses computational tools has existed for quite some time, since Mendenhall (1887) first calculated average word lengths across texts to compare the style of Shakespeare's plays against that of his contemporaries who were theorized to be his allegedly true identity. A few decades later, Mosteller and Wallace (1964), in their analysis of the Federalist Papers, became the first to use "non-traditional" methodology to predict the authorship of anonymous texts. In their case they used word length and word frequencies with an early version of a Naive Bayes classifier to ascertain whether, out of the 77 total papers, the 12 of debatable origin had been written either by Alexander Hamilton or by James Madison. Thus they demonstrated the value of "stylometry," or the linguistic analysis of an author's writing style, for reliably settling the authorship of disputed texts (Modaber Dabagh 2007). This type of analysis is what Stamatatos (2009) calls "similarity-based methods," where the goal is to represent documents with known authorship as points in space, and to assign authorship to a disputed text based on which known texts it is closest to, according to some distance metric. Another key example is Burrows' Delta (2002), which finds the z-scores for select word frequencies in each text and calculates pairwise distances using the mean absolute differences between corresponding z-scores.

In the 1990s authorship analysis was revitalized after the start of the

Internet age stimulated the fields of natural language processing and information retrieval. It became much easier to systematically retrieve and clean information from a wide body of text—something Mosteller and Wallace had to do manually—and new machine learning algorithms were able to handle that data despite their complexity; moreover, research in natural language processing diversified the methods available for representing writing style (Stamatatos 2009). There was more text that could be sifted through and easier ways to do the sifting. Of the methods used, support vector machines have emerged as the best at handling the high dimensionality of textual data for categorization tasks (Koppel et al. 2009).

The type of data adapted from text for analysis, or features, can be lexical, character-based, syntactic, or semantic, of which the first two tend to perform best (Stamatatos 2009; Stamatatos 2018). Lexical features often consist of "function words"—vocabulary like "to," "at," or "and" that are present because the language requires them, regardless of topic. An author might, intentionally or not, use some function words more often than or instead of others, revealing the more subconscious, and thus unique, elements of their writing style (Stamatatos 2009). This is opposed to using "content words," which refer to all other possible words in a document; distinguishing between function or content words requires knowledge of the language of the "corpus," or collection of texts used for study (Stamatatos 2009). Content words have their own place within authorship analysis; for example, one writer may prefer using "fast" where another would use "quick," or "lofty"

instead of "towering." But a model trained heavily on these words risks modeling according to topic matter than to style. The number of words to select from a text, function or content, will depend on the nature of the corpus at hand.

Character-based feature types, the popular alternative to lexical features, center around the use of "n-grams," which are sequences of n consecutive characters (including letters, numbers, white space, and punctuation) from a text. N-grams need not be long: 2-grams have been shown to be effective in the right contexts (Koppel et al. 2009). By their nature n-grams combine both function-based and content-based information, and are often more effective than using whole words (Stamatatos 2018). Both these and lexical features are language-independent in that a computer can decompose a text into a collection of word or n-gram frequencies without needing much information about the linguistic rules of the language at hand. Thus, models that employ them are easily transferable between languages. The choice of n for n-grams is more language-dependent, however; if a language naturally has longer words on average, a larger n may be more appropriate (Stamatatos 2009).

The selection of which words or n-grams to consider in the model is often done by their frequency, such as picking the 1000 most frequent 3-grams across the entire corpus. From there, each text is represented by a vector containing the frequency of each of those 3-grams within the text specifically. Also used is "information gain," which picks features based on how

much information they provide in discriminating between authors (Koppel et al. 2009; see Section 3.1.1). Either way, the data obtained often contain hundreds to thousands of features, potentially more than there are texts available to analyze. Methods like support vector machines operate well in these situations, but dimensionality reduction can be useful regardless. A common option for this is principal components analysis (Stamatatos 2009; see Section 3.1.2).

Syntactic and semantic feature types comprise those that yield more language-dependent model implementations. The former can involve discovery of the frequencies of different syntactic structures, or of words by their parts of speech, and the latter can involve identifying synonym distributions and semantic relationships between words (Koppel et al. 2009; Manousakis and Stamatatos 2023). These tend to be "too noisy and less effective" on their own compared to lexical and character-based features (Stamatatos 2018).

2.1.1 Topic Influence

The question of content word inclusion touches on a wider issue within authorship analysis: how to ensure that the descriptive and predictive analyses are catching an author's true fingerprint in the writing, rather than being influenced by topic similarities between texts. Function words are a good approach to this, but they are "not immune to topic shifts," and neither are character n-grams (Stamatatos 2018). Furthermore, leaving out content words entirely may mean leaving out useful information. Stamatatos (2018)

examines a few possible options that involve replacing content-specific vocabulary (defined in that case as any word not among the top x most common words in the corpus) with different patterns of asterisks so as to maintain the document's original structure. He evaluates these methods only in the context of cross-topic attribution, where the topic of the test texts are not included among the topic(s) of the training documents. This process also means losing some of the information provided by any word choice that is not popular in the corpus.

Conversely, topic modeling methods have shown remarkable success for authorship attribution and verification. These methods seek to identify common topics across a corpus based on which words co-occur the most, and thus obtain for each text a mixture of proportions that show how much of each text belongs to each topic. This greatly reduces dimensionality (Seroussi et al. 2011). The most popular topic model used for this purpose is latent Dirichlet allocation, to great effect (Seroussi et al. 2011; Stamatatos 2018; see Section 3.3.1). These topics do not always correlate with human-understood topics—more than that, they reveal similar word use patterns. This means they can include function or content words. By and large, character-based features remain the most effective choice, especially for Greek drama (Manousakis and Stamatatos 2023). In this work, we evaluate the performance of select authorship analysis methods trained on either character n-grams or a mixture of function and content words, with one method utilizing both words and topics.

2.2 Authenticity and Authorship of Greek Texts

Ancient Greek literature contains many instances of spurious or questionable authorship on account of their transmission. Prior to the invention and proliferation of the printing press in the 15th century these texts had to be copied by hand from a preexisting source, both for publishing during antiquity and for preservation and dissemination through late antiquity and the Middle Ages. This process naturally creates many errors and changes. Sometimes a genuine mistake may be made in a manuscript that a subsequent copyist tries to fix, only to exacerbate the issue further; other times a reader's annotations may be mistakenly incorporated into the main body of the text by someone else (Armstrong 2015). At times intentional edits were made to change or amend the text due to the copyist's preferences. A textual editor creating a new edition for modern study has to collate text from the body of manuscripts that exist of that document, which might contradict each other depending on the textual history and tradition of the original source. The editor makes decisions of what to include based on their knowledge of the history of the text, of the manuscripts, and of the types of errors that may arise within them.

The techniques of authorship analysis have lent themselves well to aiding the discovery, analysis, and discussion of spurious segments of ancient Greek literature. Pavlopoulos and Konstantinidou (2023) modeled the probability distributions of 3-grams from Homer's *Iliad* and *Odyssey*, evaluating (1) if their models correctly identified set-aside excerpts as being from one poem

or the other and (2) if their models correctly distinguished Homer's writing from contemporary poet Hesiod's. Their models performed better than human analysts presented with the same tasks, and furthermore, highlighted which books or fragments of the *Iliad* and the *Odyssey* have the most linguistic proximity to one another and which books stand out as more stylistically distinct. Field (2017) employed naive Bayes, support vector machines, and cosine similarity for 2- through 4-grams to investigate the final chapter of Xenophon's fictionalized biography *Cyropaedia*, not only to assess whether or not the final chapter was truly Xenophonic, but to evaluate how distant it is from its source text compared to the intrabook or intrachapter style variations found in other historical or philosophical texts from Greek antiquity. In the process Field concluded that the final chapter was indeed authentic, with any stylistic differences attributable either to its difference in content from the rest of the work or to natural variation.

Extensive work in this vein has been done by Manousakis and Stamatatos for Greek drama. Manousakis (2020) analyzed the authorship of *Prometheus Bound*, an Aeschylean play doubted even in antiquity, comparing the distributions of function words observed in the text to that of other dramas with principal components analysis (PCA), clustering, and Burrows' Delta. Manousakis also used k-nearest neighbors (k-NN) and support vector machines (SVM) on 3-grams and 4-grams to attribute authorship. Together, Manousakis and Stamatatos have done additional analysis on further plays: using PCA and SVM on 3-grams and 4-grams, they found the suspicious

Rhesus to be, most likely, the product of a fourth century author instead of Euripides (2018); further, they analyzed the spurious conclusion of Aeschylus' *Seven Against Thebes* using k-NN and SVM, again on 3-grams and 4-grams, to find that portions of the conclusion were indeed questionable (2023). In this latter paper they introduced the technique of n-gram tracing, a descriptive tool used previously to analyze an alleged letter written by Abraham Lincoln and designed expressly for short texts, to consider the short disputed lines of *Seven*. They applied this method again to Euripides' *Electra*, finding in that case that the play was true (2024). Their implementations of k-NN and SVM reflect two separate approaches to training models: profile-based versus instance-based. In the former, each text from a candidate author is treated as a unique observation or "instance." In the latter, all the available texts from a candidate author are concatenated together to obtain a cumulative "profile" of the author's trace (Manousakis and Stamatatos 2023). Both approaches appear to offer high accuracy for Greek drama. Given their proven accuracy for both Greek tragedy and other texts, this study uses support vector machines trained on 3-grams and 4-grams as a baseline against which we compare the newly proposed model. All model implementations take an instance-based approach.

2.2.1 Spurious Texts of Interest

This study engages with two Greek tragedies for which some spurious questions remain that have not been treated with non-traditional analysis: *Iphi-*

genia at Aulis and *The Phoenician Women*, both attributed to Euripides.

Iphigenia at Aulis was composed in 405 BCE. Set right before the Greeks depart for the Trojan War, the play centers on the decision by King Agamemnon to sacrifice his daughter Iphigenia to appease the rage of the goddess Artemis. It was published and staged posthumously by his son, who may have edited the text, but the play carries a disputed conclusion aside from that, thought to have been added by a third party such as a theatrical company (Hall 2010). The conclusion includes Iphigenia's second monody (a type of solo lyric poem, contained in lines 1475-1509), a choral song for her departure (1510-1531), and a report delivered by a messenger telling that Iphigenia has been spirited away by Artemis just before the final blow was delivered, and thus providing a tentative relief that she has been spared (1532-1629). The messenger's speech greatly changes (and even undermines) the emotional impact of Iphigenia's departure, and is considered to be spurious by scholars. The monody and choral song preceding the speech, however, are more complicated. The choral song exhibits quite a bit of repetition from the prior monody, drawing suspicion, but David Kovacs has argued the opposite: that the choral song is true, and that Iphigenia was meant to depart sooner, before line 1475 (Weiss 2014). Weiss instead argues that the two segments compliment each other, and that "they respond to each other in ways that suggest ... they were originally intended by Euripides" (2014). Computational analysis of the final 150 lines of the play could provide clarity to the debate.

The Phoenician Women, composed by Euripides in the early 400s BCE, is a story about the royal family of Thebes, which consists of the mythical hero Oedipus and his children. This play deals with the civil war begun by Oedipus' two sons over control of the city. This too is a play with a questionable conclusion, but what complicates conversations regarding this work even more is the overall poor quality of the text, with many short, one to two line interpolations throughout (Hall 2010). Many of these lines are subject to debate, with reasons given being anything from peculiar writing choices that undermine character development to awkward grammar that Euripides would not have employed (Mastronarde 1994). Computational analysis of this play can, as with *Iphigenia at Aulis*, provide some clarity to debate, and also inform how non-traditional techniques respond to interpolations embedded within authentic text.

In this work we apply authorship analysis models to verify the authorship of contested passages from both of these plays.

3 Methodology

3.1 Text Representation

In this section we discuss the mathematical representation of the Greek tragedies considered in our analysis. The collection of texts ("corpus") consists of 28 plays from Aeschylus, Euripides, and Sophocles. This entails the entire extant body of tragic work from each playwright barring *Iphigenia at Aulis* and *The Phoenician Women*. Also excluded are Aeschylus' *Prometheus Bound* and Euripides' *Rhesus* due to their spuriousness. The texts were obtained from the Perseus Digital Library (Crane n.d.) and cleaned using the Classical Languages Toolkit in Python (Johnson et al. 2021). Cleaning involved changing any modern *tonos* accent marks into the ancient *oxia* equivalents and normalizing the Unicode transcriptions of the Greek text (that is, ensuring a character such as ἐ is parsed as one whole rather than a sequence of ᾽ and ε). After cleaning the texts were imported into R for tokenization via the `stylo` package (Eder et al. 2016a) and for the necessary analyses. The observations were grouped into two classes: 0 for texts written by Sophocles or Aeschylus and 1 for texts written by Euripides. Given that the spurious sections of *Iphigenia at Aulis* and *The Phoenician Women* are only a few lines long, we will follow the approach used in Manousakis and Stamatatos (2023) of dividing each text of the corpus into chunks of 50 lines each. Thus, out of 28 Greek tragedies, we obtain 941 observations, with 413 in class 0 and 528 in class 1. These comprise our training set of texts with known au-

thorship. Our test set of texts to which we seek to assign authorship consists of 7 observations: 4 from the conclusion of *Iphigenia at Aulis* and 3 from *The Phoenician Women*.

A common approach to text representation for ancient Greek authorship analysis involves the use of character n-grams, where a given text is partitioned into overlapping character strings of length n. For example, character 2-grams of the phrase "the quick brown fox" will include {"th", "he", "e ", " q", "qu"} and so on. This is used to obtain a list $L = \{L_1, \ldots, L_p\}$ of p n-grams with the most predictive ability from across the entire corpus, selected by the method in Section 3.1.1. These p n-grams are then used in the design matrix $\boldsymbol{X}$. So, for a text chunk j and an n-gram L_i, x_{ij} corresponds to the relative frequency of L_i in j. This relative frequency is calculated as a ratio of the total number of occurrences of L_i in j to the total number of unique n-grams in j. In our case, $j = 1, \ldots, 941$. Note that $\sum_{i=1}^{n} x_{ij} \neq 1$ necessarily, since a text chunk may contain n-grams that are not in L. For example, the n-gram "qu" might appear in text chunk j, but is not included in L. It would only be included in the denominator of the relative frequencies of j.

3.1.1 Information Gain

The overlapping nature of n-grams means they cannot serve as predictors in a model in which the observations are assumed to be independent. Therefore, to apply a generalized mixed-effects regression, we follow the approach employed in Koppel et al. (2009) involving content words as predictors. Given

the 10,000 most common single words in the corpus, the parameters for the final design matrix are chosen from those that discriminate between the classes of the data set the most, according to the concept of information gain described by Quinlan (1986). Let C_0, C_1 be the two classes of our response variable. Then the entropy, or uncertainty, of the class assignment yielded by a classification model for a text in the corpus can be expressed as

$$I(C_0, C_1) = -\frac{n_{C_0}}{n}\log_2(\frac{n_{C_0}}{n}) - \frac{n_{C_1}}{n}\log_2(\frac{n_{C_1}}{n}),$$

where n_{C_0} is the number of observations in C_0, n_{C_1} is the number of observations in C_1, and $n = n_{C_0} + n_{C_1}$. Let W be a set containing all observed frequency values of a word w from the potential 10,000. The word w is considered highly discriminating between C_0 and C_1 if its information gain, calculated as

$$I(C_0, C_1) - \sum_{i \in W} \frac{n_{C_0}^{(i)} + n_{C_1}^{(i)}}{n} \cdot \left(-\frac{n_{C_0}^{(i)}}{n^{(i)}} \log_2 \left(\frac{n_{C_0}^{(i)}}{n^{(i)}} \right) - \frac{n_{C_1}^{(i)}}{n^{(i)}} \log_2 \left(\frac{n_{C_1}^{(i)}}{n^{(i)}} \right) \right)$$

is high, where $n^{(i)}$ indicates the number of observations where word w appears with a frequency of i.

For example, if the word καὶ ("and") appears in one text chunk from Euripides (C_1) and one from Aeschylus or Sophocles (C_0) at frequencies 0.002 and 0.01, respectively, then its information gain would be

$$I_{\text{καὶ}} = -\tfrac{416}{941}\log_2(\tfrac{416}{941}) - \tfrac{256}{941}\log_2(\tfrac{256}{941}) - \left[\tfrac{0+1}{941}\left(-\tfrac{1}{1}\log_2\tfrac{1}{1}\right)\right] - \left[\tfrac{1+0}{941}\left(-\tfrac{1}{1}\log_2\tfrac{1}{1}\right)\right]$$

where the bracketed terms correspond to $W_{0.002}$ and $W_{0.01}$, respectively.

Koppel et al. used this approach to select both 1000 words and 1000 3-grams from lists of 10,000 for each, obtaining high predictive accuracy when used to train multinomial regression and SVM on a literature corpus. We employ the same for word-based SVM, and use this process to select the p most informative n-grams for SVM implementations based on both 3-grams and 4-grams. We use the `FSelector` R package (Romanski et al. 2023).

3.1.2 Principal Components Analysis

This selection process risks the filtering in of feature words that are correlated with one another: for example, we may capture a character name that only ever appears in the same texts as another character or place name, also captured as a viable, highly discerning predictor. To mitigate the issue this causes for models that assume noncorrelation of predictors, we employ principal components analysis (PCA). PCA is a method to reduce the number of predictors in a data set from p to a pre-determined quantity $M < p$ without loss of information. We create linear combinations $Z_1, \ldots, Z_M$ of the p predictors that represent the directions in space along which the greatest variation in the observations can be found. A principal component Z_i is constructed such that

$$Z_i = \phi_{1i}\frac{(X_1 - \bar{X}_1)}{\sigma_{X_1}} + \phi_{2i}\frac{(X_2 - \bar{X}_2)}{\sigma_{X_2}} + \ldots + \phi_{pi}\frac{(X_p - \bar{X}_p)}{\sigma_{X_p}},$$

where $i = 1, 2, \ldots, M$ and the new predictors $\boldsymbol{\phi}_i$, called the principal component loadings, dictate that direction of greatest possible variance while subject to the restriction that $\sum_{j=1}^{p} \phi_{ji} = 1$. Further, the linear combinations are constructed such that Z_{i+1} is strictly uncorrelated to $Z_1, \ldots, Z_i$. In this way the information provided by our predictors is not reduced but rather projected from $(p+1)$-dimensional space into $(M+1)$ dimensions instead, decreasing the inflated variance caused by correlated predictors. We use the `prcomp` function in R (R Core Team 2013).

3.2 Support Vector Machines

Support vector machines (SVMs) are among the most common classification algorithms for text categorization and authorship attribution on account of their high accuracy (Koppel et al. 2009, Yu 2008), so we employ them here for baseline comparison. As binary classifiers, they function by constructing a hyperplane defined by

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = 0,$$

where $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_p]'$ is the vector of parameters and $\mathbf{X}$ is the $n \times p$ design matrix. A point will be assigned to a class depending on whether it yields a value greater or less than 0 by the equation. The theory behind SVM builds off of the concept of a maximal margin classifier, where the hyperplane is constructed such that the margin—the space between the hyperplane and

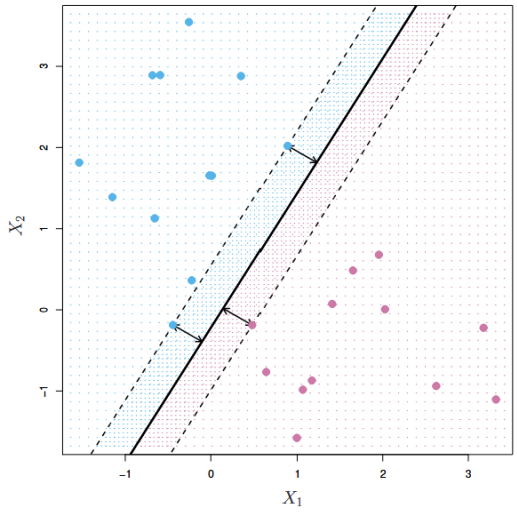

Figure 3.1: A hyperplane dividing a data set. The dotted lines indicate the outer edges of the margin, with three support vectors sitting on, and therefore dictating, where the hyperplane lies. Image from James et al. (2021).

the closest observations from either class, called support vectors—is as large as possible, as shown in Figure 3.1. However, because a change in any of the support vectors may lead to a pronounced movement of the hyperplane, as shown in Figure 3.2, SVMs instead allow some points to sit within the margin or even on the wrong side of the boundary. The location of a point i relative to the margin or hyperplane is dictated by a slack variable ϵ_i: $\epsilon_i = 0$ if point i is on the correct side of the margin and $\epsilon_i > 0$ if not; if $\epsilon_i > 1$, it is on the wrong side of the hyperplane entirely. The number of points for which $\epsilon_i > 1$ is dictated by a tuning parameter C, which, in authorship attribution problems, is typically set to 1. Thus, given the dimensions of our data, we construct the hyperplane by solving the optimization problem

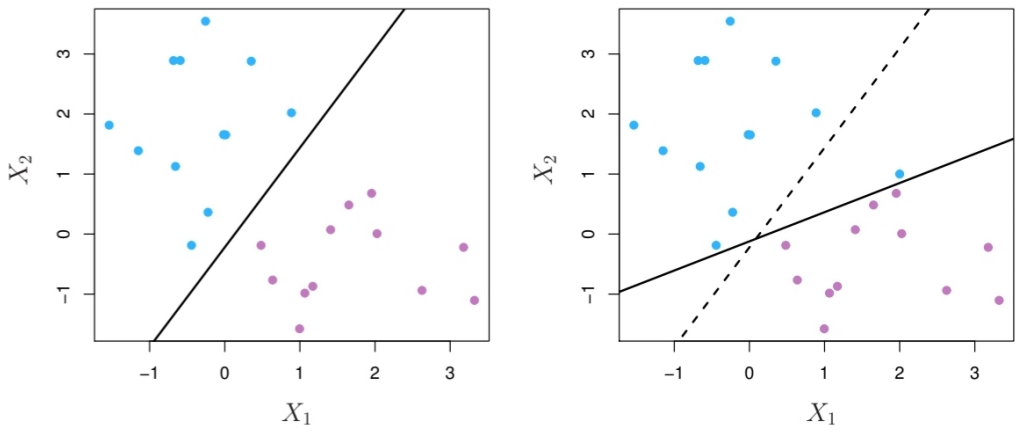

Figure 3.2: On the left, a hyperplane dividing a data set with two parameters with maximal distance from the nearest points. On the right, the addition of a new point within the blue class completely shifts the location of the boundary. Images from James et al. (2021).

$$\begin{aligned} \max_{\boldsymbol{\beta},\boldsymbol{\epsilon},M} \quad & M \\ \text{subject to} \quad & \sum_{j=1}^{p} \beta_j^2 = 1 \\ & y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i) \\ & \epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C \end{aligned}$$

where M represents the margin around the hyperplane and $y_1, \ldots, y_n \in \{-1, 1\}$ are the class assignments for each observation. The first two constraints ensure that the desired number of points lie on the correct side of the margin, where $y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})$ represents the distance of point i from the hyperplane. The solution of the optimization problem gives a linear boundary between both classes—in cases where this does not suit, a kernel

is applied to expand the feature space via higher-order polynomials. Since authorship attribution problems operate in a large feature space by their nature, the default linear kernel suffices. For this work we use the `stylo` package in R (Eder et al. 2016b).

3.3 Mixed Models

Mixed models are an extension of standard regression models applicable in cases of hierarchical or correlated data. This can happen in the case of a longitudinal study, when multiple observations are taken from the same study participants over time, or when the data are clustered due to the sampling procedure used—say, when surveying students who have been sampled by their school affiliation. In these cases, the relationships between the predictor variables and the response may vary on an individual or cluster level, respectively, but to include all of these effects in the parameter matrix $\boldsymbol{\beta}$ would be computationally troublesome or infeasible (Fahrmeir et al. 2013). Thus, a distinction is made between the "fixed" effects $\boldsymbol{\beta}$ and the individual- or cluster-level "random" effects $\boldsymbol{\gamma}$ which are estimated jointly. So, if the model equation for a linear fixed-effects regression is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$, the model equation for a linear mixed model becomes

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{U}$ represents the design matrix for the random effects $\boldsymbol{\gamma} \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{G})$ such that $\boldsymbol{G}$ is the positive definite covariance matrix. Aside from the distributional assumptions on $\boldsymbol{\epsilon}$ and $\boldsymbol{\gamma}$, we further assume that the predictor variables are not perfectly correlated with one another (such that one predictor can be used as a perfect substitute for another) and that they are linearly related to $\boldsymbol{Y}$.

In this study we will seek to model authorship with mixed-effects logistic regression, a type of generalized linear mixed model (GLMM). Where this departs from typical implementations of logistic regression for text classification is in the assumption that some correlation exists between text chunks that pertain to the same topic, and that these errors should be estimated as a random effect. As would be the case in fixed-effects logistic regression, a logit function is applied to predict the probability that a given text chunk was written by an author. Let T be the number of topics identified and n_i the number of text chunks included in the ith topic. We assume that the relationship between word frequencies and authorship will be similar from topic to topic, so cluster-specific heterogeneity will be modeled by a random intercept γ_{0i} for each topic. Then, for a text chunk j in topic i, our

mixed-effects model follows

$$\mathrm{P}(y_{ij} = 1) = \pi_{ij} = \frac{\exp(\boldsymbol{x'}_{ij}\boldsymbol{\beta} + \gamma_{0i})}{1 + \exp(\boldsymbol{x'}_{ij}\boldsymbol{\beta} + \gamma_{0i})},$$

where $i = \{1, \ldots, T\}$, $j = \{1, \ldots, n_i\}$, $\boldsymbol{x'}_{ij}$ is the vector of covariates for point ij, and $\boldsymbol{\beta}$ is the $(p+1)$ parameter vector. In this model we make much the same assumptions as in the linear mixed-effects case, but instead assume that $\mathrm{P}(y_{ij})$ is related to the linear predictor $\boldsymbol{x'}_{ij}\boldsymbol{\beta} + \gamma_{0i}$ by the function for π_{ij}.

The inference for this model follows a maximum likelihood framework. Observe that the likelihood function

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{Y}) = \mathrm{P}(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) \cdot \mathrm{P}(\boldsymbol{\gamma}).$$

Because $\boldsymbol{Y} \sim \mathrm{Binomial}(\mathbf{1}, \boldsymbol{\pi})$ and $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{G})$, we have that

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{Y}) = \boldsymbol{\pi}^{\boldsymbol{Y}}(1 - \boldsymbol{\pi})^{1-\boldsymbol{Y}} \cdot (2\pi)^{-\frac{n}{2}}|\boldsymbol{G}|^{-\frac{1}{2}} \exp(-\frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{G}^{-1}\boldsymbol{\gamma}).$$

Taking the logarithm yields

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{Y}) \propto \boldsymbol{Y}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma}) - \log(1 + \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma})) - \frac{1}{2}\boldsymbol{\gamma'}\boldsymbol{G}^{-1}\boldsymbol{\gamma}.$$

Note that $-\frac{1}{2}\boldsymbol{\gamma'}\boldsymbol{G}^{-1}\boldsymbol{\gamma}$ acts as a penalty term for deviations from $\mathbf{0}$ in the random effects. Because there is no closed form for $\hat{\boldsymbol{\beta}}^{MLE}$ or $\hat{\boldsymbol{\gamma}}^{MLE}$, we

proceed with iterative estimation by either a Newton-Raphson procedure or Fisher scoring. For this work we use the `lme4` package in R (Bates et al. 2015).

3.3.1 Latent Dirichlet Allocation

In the GLMM, we need the topic of each text. To identify those topics in the corpus, we employ latent Dirichlet allocation (LDA). In such a model, each document is viewed as a mixture of topics, and each word within a document as appearing according to a topic-dependent distribution. Finding the mixture of topics for each document becomes a question of evaluating the posterior distribution

$$\Pr(\boldsymbol{z}|\boldsymbol{w}) = \frac{\Pr(\boldsymbol{w}|\boldsymbol{z})\Pr(\boldsymbol{z})}{\Sigma_{\boldsymbol{z}}\Pr(\boldsymbol{w}|\boldsymbol{z})\Pr(\boldsymbol{z})},$$

where $\boldsymbol{w} = \{w_1, \ldots, w_V\}$ is the vector of V unique words in the corpus and $\boldsymbol{z}$ is the vector of topic assignments for each word. The likelihood of $\boldsymbol{w}$ given topic z_j is assumed to follow a Multinomial distribution with probability $\phi_w^{(j)}$ such that $\phi_w^{(j)} \sim$ Dirichlet(β) and β is a hyperparameter. Similarly, the prior $z_j|d, \theta$ follows a Multinomial distribution with probability $\theta_j^{(d)}$ such that $\theta_j^{(d)} \sim$ Dirichlet(α), α is a hyperparameter, and $d = 1, \ldots, D$ designates one of the D documents of the corpus. Given its efficacy in authorship attribution problems (Seroussi et al. 2011), we employ the implementation of LDA discussed in Griffiths and Steyvers (2004) where $\Pr(\boldsymbol{z}|\boldsymbol{w})$ is computed

via Gibbs sampling and where the hyperparameters are set at $\alpha = \frac{50}{T}$ and $\beta = 0.1$. The number of clusters specified in the GLMM model corresponds to the number of topics found through LDA—a text will be assigned to the cluster corresponding to the topic of plurality in its mixture.

In sum, we implement and compare five methods for authorship verification: SVMs with 3-grams, 4-grams, and full words (all of which are identified via the information gain process), logistic fixed-effects regression trained on M principal components, and logistic mixed-effects regression trained on M principal components and T topic clusters identified through LDA.

4 Results

4.1 Authorship Prediction

We first discuss model performance. For both logistic models a decision boundary of 50% was found to minimize the rate of misclassifications; thus, in all regression results shown, a text chunk was assigned to Euripides if the predicted probability of his authorship was above 50%.

Table 4.1 displays the true positive and false positive rates averaged from 10-fold cross validation with each method. The true positive rate (TPR) is calculated as the number of correct "positive" (in this case, correct "Euripides") predictions divided by the total number of true Euripidean text chunks in the corpus. Finally, the false positive rate (FPR) indicates the proportion of Sophoclean or Aeschylean texts that were incorrectly attributed to Euripides. We want TPRs close to 1 and FPRs close to 0. In conjunction, these metrics tell us how often a model correctly identifies Euripides' style and how often a model incorrectly identifies him as the author of a text he did not write.

In line with the literature, the n-gram based SVMs performed the best, with the lowest FPRs and perfect identification of Euripidean documents in the 3-gram case. Conversely, the SVM implementation with words showed the highest FPR of all methods. The logistic mixed-effects regression seemed to perform moderately worse than the logistic fixed-effects regression, though their performances are within one percent of each other. Either logistic option

Performance Metric	SVM: 3-grams	SVM: 4-grams	SVM: Words	Logistic: Fixed	Logistic: Mixed
True Positive Rate	1.000	0.993	0.956	0.962	0.958
False Positive Rate	0.006	0.023	0.066	0.031	0.036

Table 4.1: True positive and false positive rates for all five models after 10-fold cross validation on the training set.

is reliable, perhaps more so than SVM with words, though neither is as reliable as using n-grams. All models, to some extent, are prone to identifying Euripidean material where none exists, though some less so than others. The SVM models are less likely to miss authentic work altogether, and the logistic models are about equally likely to produce a false positive. Performance of the logistic mixed-effects model could change with a different choice of random effect, such as clustering by a variable other than topic of plurality. Overall, though, all five models perform relatively well.

We apply each of the five models onto select sections from *Iphigenia at Aulis* (IA) and *The Phoenician Women* (PN). These comprise seven text chunks with debatable authorship:

- IA1: Iphigenia's final monody, lines 1475-1509: potentially spurious
- IA2: the final choral song, lines 1510-1531: potentially spurious
- IA3: the beginning half of the messenger's report, lines 1532-1582: certainly spurious

- IA4: the final half of the messenger's report and some concluding dialogue, lines 1583-1629: certainly spurious
- PN1: lines 1-50, with certainly spurious 1-2
- PN2: lines 549-559, with certainly spurious 558
- PN3: lines 760-810, with certainly spurious 778

In the case of IA we discuss the model's ability to capture the spuriousness of lines 1532-1629 and whether or not they attribute lines 1475-1509 to Euripides. For PN, we discuss how the model predictions are affected by the presence of small spurious lines placed within otherwise reliable Euripidean material. The results are presented in Table 4.2.

All models reach the conclusion that IA chunks 1 and 2, corresponding to lines 1475-1509, are Euripidean; on the other hand, all but the SVM 3-gram model catch the spuriousness of lines 1583-1629 in IA4. The `stylo` R package, with which SVM was employed, did not produce probabilistic estimates; it is possible that the assignment to Euripides of IA3 in the 3-gram model was uncertain, with a value close to 0 by the hyperplane equation. None seem to catch issues in IA3, PN1, or PN2—in fact, the logistic models predict lines 1-50 of *The Phoenician Women* to certainly be Euripidean despite the problematic opening. The results for PN3 are more mixed: the SVMs unanimously find the passage to be inauthentic, but the logistic models suggest it is more likely authentic than not. Perhaps line 778, identified by Mastronarde (1994) as likely being an unknown reader's clarifying annotation,

Model	IA1	IA2	IA3	IA4	PN1	PN2	PN3
SVM: 3-grams	1	1	1	1	1	1	0
SVM: 4-grams	1	1	1	0	1	1	0
SVM: Words	1	1	1	0	1	1	0
Logistic: Fixed	0.999	0.999	0.985	0.262	1.000	0.999	0.648
Logistic: Mixed	0.999	0.999	0.987	0.305	1.000	0.999	0.677

Table 4.2: Predictions for each of the seven test chunks using each model. For SVM, the response is binary: 0 (non-Euripidean) and 1 (Euripidean). For logistic regression, the response is probability of authenticity.

stands out enough from the authentic material around it to confuse the models; compare this to line 558, the spurious representative in PN2, where the final word in the line, ἐφήμερος ("short-lived"), is used in a sense "not common in poetry" but seen in another play by Euripides (Mastronarde 1994). Similarly, Mastronarde remarks that lines 1-2 of *The Phoenician Women* are not non-Euripidean in their language or context, but are simply unattested in many of the ancient manuscripts and commentaries of the play, and, in conjunction with the subsequent lines, create "an odd heaping of participles"—a grammatical quirk these models are not trained to catch.

This may be what is at play in segments 1-3 of IA. Weiss (2014) noted in her discussions of *Iphigenia at Aulis* that the debated monody and choral song could both have been "intended by Euripides, even if he himself did not write them—in which case they were probably composed by . . . [someone]

trying to reproduce the tragedian's style." With both those lines and IA3 its possible that the interpolators are able to emulate Euripides' style enough to evade model detection. This is not without precedent— Manousakis and Stamatatos (2023), in evaluating their SVMs trained on plays from each of the four Greek playwrights (Euripides, Sophocles, Aeschylus, and Aristophanes), observed that their method incorrectly ascribed to Euripides a segment in Aristophanes' *Thesmophoriazousae* meant to parody the former's style. They took this as a sign in favor of their method's robustness, but the examples from IA and PN present another perspective on such an outcome: these models (which catch lexical style only) can mislead by mislabeling a skilled imitator as the original.

In any case, the topic clustering component of the mixed-effects model did not seem to provide an improvement on the original model designs.

4.2 Selection of the Predictors

We discuss the selection of predictors through cross validation. Figure 4.1 shows the results of 10-fold cross validation on SVM models using the top 100 to 900 features selected by the information gain process. The performance of each set of predictors was assessed based on the accuracy observed, calculated as the number of correct predictions made by the model divided by the total number of predictions. In the 3-gram case the choice is clearest: at 600 3-grams the accuracy for SVM reached nearly 100%. The 4-gram accuracy results plateau after 600, leading to the same option there. The greatest

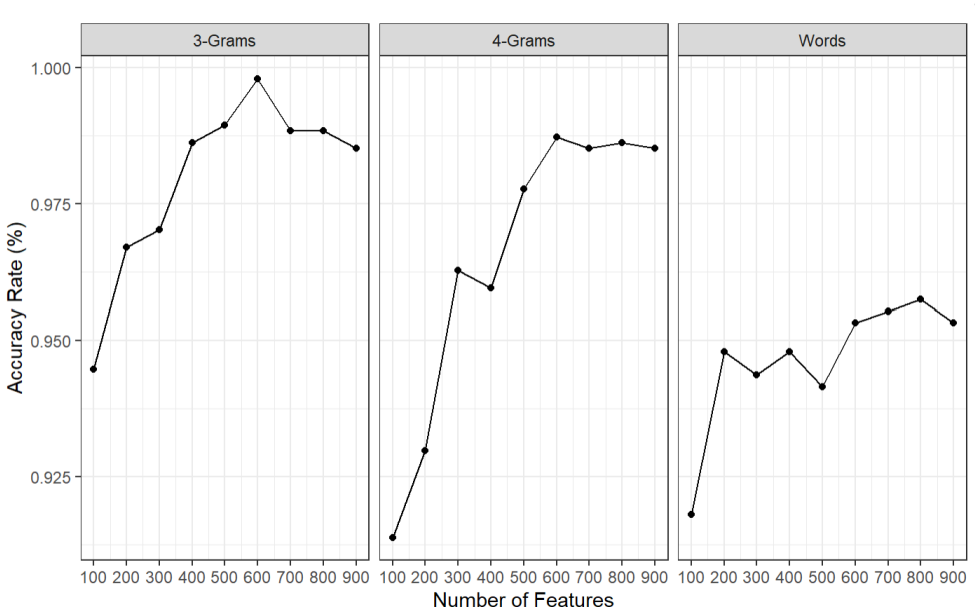

Figure 4.1: Accuracy rates after 10-fold cross validation of SVM using 3-grams, 4-grams, and words, from 100 features to 900.

accuracy for full words occurred from 600 to 900, with 200 and 400 about a percent behind. We use 200 since it offers high performance with lower dimensionality.

Figure 4.2 shows the cross validation process to select the M principal components from the 200 feature words. For each value of M from 1 to 25, 10-fold cross validation was performed on a fixed-effects logistic regression model. The AIC and BIC metrics of a candidate model are meant to estimate prediction error through a formula that rewards models under which the observations have the greatest likelihood of occurring and penalizing models that are overfit. We choose the value of M that yields the lowest AIC and BIC. In the figure we see a sharp decline in both values from $M = 1$ to

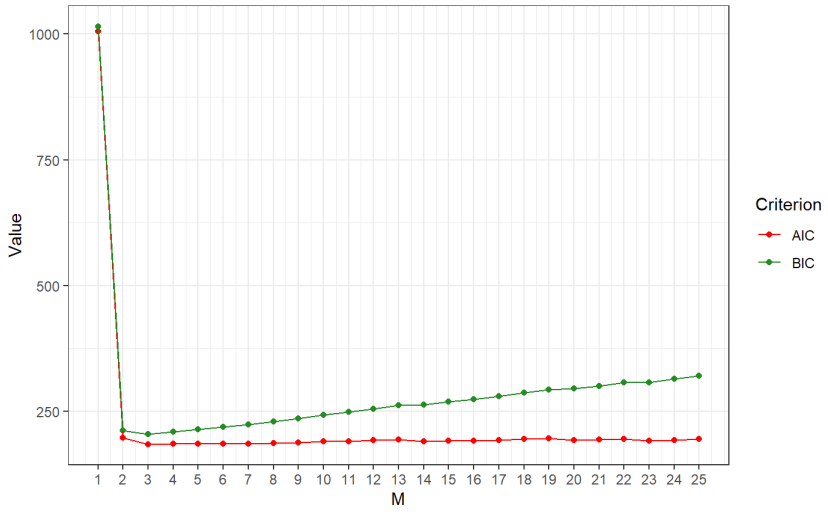

Figure 4.2: AIC and BIC values obtained from PCA applied to a fixed logistic regression model, from $M = 1$ to $M = 25$.

$M = 2$, after which the values steadily climb. Consequently, we use two principal components, since adding more would not significantly improve the predictive ability of the model.

4.3 Selection of the Topics

We employ the `FindTopicsNumber` function from the `ldatuning` R package (Murzintcev 2020) to select the number of topics T, with results shown in Figure 4.3. The function evaluates four different metrics at each supplied option for T used in the GLMM, computing topic similarities or distances in order to identify the value of T where the topics are at their most distinct from one another (Deveaud et al. 2014). The choice to assign text chunks to topic clusters based on the topic of plurality means that, as T gets larger,

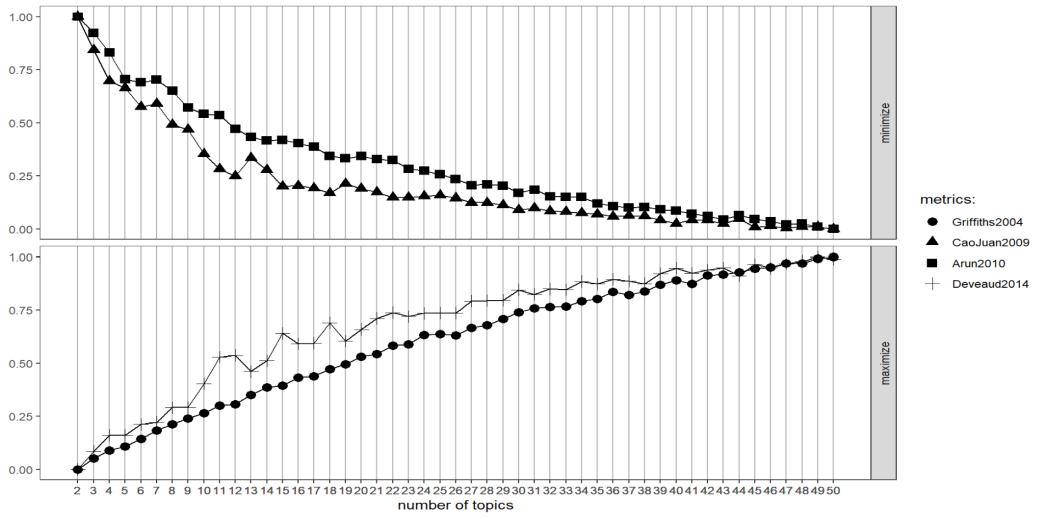

Figure 4.3: Values yielded from the four metrics of `FindTopicsNumber` for $T = \{2, 3, \ldots, 50\}$. The metrics are organized by whether they should be maximized or minimized. A low value for CaoJuan2009 or Arun2010, as well as a high value for Griffiths2004 or Deveaud2014, signifies a distribution of topics such that each topic is as different from the others as possible given their associated words.

the percentages of each text that correspond to each topic will become much smaller, and thus the designated clustering topic of choice for a document may not represent much of its content, especially given the short lengths of the chunks in the present data. Therefore we pick a value of T that allows each topic to be highly distinct without making them too trivial. Given the output shown, we use $T = 30$ as a viable middle-ground number. We employ LDA with the `topicmodels` package in R (Grün and Hornik 2011).

5 Conclusion

This study presented a mixed-effects logistic regression model to predict authorship of Greek tragedy such that topic-based clustering of the data is taken into account. We compared this method to models traditionally used for authorship analysis, evaluating their ability to identify the known authorship of the extant tragic corpus and comparing their predictions on dubious texts attributed to the tragedian Euripides.

These methods appear to be sensitive to any approximation of Euripides' style, leading to misattributions. Thus, it is hard to say definitively that the concluding monody and choral song of *Iphigenia at Aulis* are indeed Euripidean, or assess reliably how well these models can catch the embedded interpolations of *The Phoenician Women*. Regardless, each model performed remarkably well in the cross validation task, with true positive rates above 95% and the largest false positive rate at 6.6%. So, even if the mixed-effects regression did not substantially improve upon the accuracy of the other methods presented, it can still make good predictions; perhaps a corpus with more topic polarization would generate better results. It may also be possible to improve accuracy if a different clustering variable correlated with authorship is used. It may have been the case that the topic clusters, which were estimated based on the same word frequency patterns used for the fixed effects, did not improve accuracy because they did not provide brand new information.

The 3-gram SVM model performed the best, though, at first glance, it did not catch the spuriousness of lines 1583-1629 of *Iphigenia at Aulis* when the others did. In this study, the logistic regressions were much more informative in providing probabilistic estimates for authorship instead of the binary results from the SVMs. Manousakis and Stamatatos (2023) utilized an implementation of SVM that generated a "degree of affinity" rather than a binary response, which was not available with the `stylo` package employed here. This provided them with a percentage indicating the likelihood that an author in their corpus wrote a given text, yielding more information for interpreting model results. Future work could incorporate such an implementation instead.

Another avenue for future work may be to assess these models not in an authorship verification setting (that is, the Euripides/not-Euripides binary used here) but instead for authorship attribution, where each playwright's corpus is kept separate. This was the approach taken by Manousakis and Stamatatos (2023), where they trained multiple SVMs on each possible pair of authors. Such a design, as well as the use of multinomial regression instead of binomial regression, could have changed the results obtained. Performance may improve even more if the writing styles of Aeschylus and Sophocles are kept distinct in the training process.

It is also worth noting a limitation in the way the results from latent Dirichlet allocation were used: only the most prevalent topic in each text chunk was taken into consideration. Previous work has shown that the topic

mixtures alone can serve as highly accurate predictors for authorship analysis (Seroussi et al. 2011), suggesting that there is value in incorporating all topics observed rather than just one. Future work may consider an alternative set-up that does not create this loss of information.

References

Armstrong, R. H. (2015). A wound, not a world: Textual survival and transmission. In M. Hose & D. Schenker (Eds.), *A companion to greek literature* (pp. 27–40). John Wiley & Sons.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Burrows, J. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, *17*(3), 267–287.

Crane, G. R. (Ed.). (n.d.). *Perseus Digital Library*. Tufts University. Retrieved October 15, 2024, from http://www.perseus.tufts.edu

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

Eder, M., Rybicki, J., & Kestemont, M. (2016a). Stylometry with R: A package for computational text analysis. *The R Journal*, *8*(1), 107–121. https://doi.org/10.32614/RJ-2016-007

Eder, M., Rybicki, J., & Kestemont, M. (2016b). Stylometry with r: A package for computational text analysis. *R Journal*, *8*(1), 107–121. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, methods and applications*. Springer Berlin. https://doi.org/10.1007/978-3-642-34333-9

Field, A. (2017). Authorship analysis of Xenophon's Cyropaedia. *Computing Research Repository*. http://arxiv.org/abs/1711.01684

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30. https://doi.org/10.18637/jss.v040.i13

Hall, E. (2010). *Greek tragedy: Suffering under the sun*. Oxford University Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer New York. https://doi.org/10.1007/978-1-0716-1418-1

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. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 20–29. https://doi.org/10.18653/v1/2021.acl-demo.3

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, *60*(1), 9–26. https://doi.org/10.1002/asi.20961

Manousakis, N. (2020). *Prometheus bound. Vol. 98. Trends in classics—supplementary volumes: A separate authorial trace in the aeschylean corpus* (F. Montanari & A. Rengakos, Eds.; *Vol. 98*). De Gruyter.

Manousakis, N., & Stamatatos, E. (2018). Devising Rhesus: A strange 'collaboration' between Aeschylus and Euripides. *Digital Scholarship in the Humanities*, *33*(2), 347–361. https://doi.org/10.1093/llc/fqx021

Manousakis, N., & Stamatatos, E. (2023). Authorship analysis and the ending of Seven Against Thebes: Aeschylus' Antigone or updating adaptation? *Classical World*, *116*(3), 247–274. https://doi.org/10.1353/clw.2023.0007

Manousakis, N., & Stamatatos, E. (2024). Authorship analysis and the authenticity of Euripides' Electra 518–44: Preserving character constistency. *Classical Philology*, *119*(3), 338–353. https://doi.org/10.1086/730675

Mastronarde, D. J. (1994). *Euripides: Phoenissae*. Cambridge University Press.

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, *9*(214), 237–249.

Modaber Dabagh, R. (2007). Authorship attribution and statistical text analysis. *Metodološki zvezki*, *4*(2), 149–163.

Mosteller, F., & Wallace, D. L. (1964). *Inference and disputed authorship: The federalist*. Addison-Wesley.

Murzintcev, N. (2020). *Ldatuning: Tuning of the latent dirichlet allocation models parameters* [R package version 1.0.2]. https://CRAN.R-project.org/package=ldatuning

Pavlopoulos, J., & Konstantinidou, M. (2023). Computational authorship analysis of the homeric poems. *International Journal of Digital Humanities*, *5*, 45–64. https://doi.org/10.1007/s42803-022-00046-7

Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106. https://doi.org/10.1007/BF00116251

R Core Team. (2013). *R: A language and environment for statistical computing* [ISBN 3-900051-07-0]. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/

Romanski, P., Kotthoff, L., & Schratz, P. (2023). *Fselector: Selecting attributes* [R package version 0.34]. https://CRAN.R-project.org/package=FSelector

Seroussi, Y., Zukerman, I., & Bohnert, F. (2011). Authorship attribution with latent dirichlet allocation. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 181–189. https://doi.org/10.5555/2018936.2018957

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–556.

Stamatatos, E. (2018). Masking topic-related information to enhance authorship attribution. *Journal of the American Society for Information Science and Technology*, *69*(3), 461–473.

Weiss, N. A. (2014). The antiphonal ending of Euripides' Iphigenia in Aulis (1475–1532). *Classical Philology*, *109*, 119–129. https://doi.org/10.1086/675252

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, *23*(3), 327–343. https://doi.org/10.1093/llc/fqn015