

Authorship Attribution of the Federalist Papers Using Stylometry and Machine Learning

Robert Thomas Nitto
Information Sciences, Syracuse University
Spring 2026

Abstract—Authorship attribution is the task of identifying the most likely author of a document based on linguistic and stylistic features. This study investigates authorship of the Federalist Papers using stylometry and machine learning text classification techniques. A structure corpus of papers authored by James Madison and Alexander Hamilton was constructed and used to train and evaluate Multinomial Naïve Bayes, Logistic Regression and Support Vector Machine classifiers. Text representation was performed using Term Document Inverse Term Frequency (TF-IDF) enabling high dimensional sparse feature modeling for classification. Model performance was assessed using cross-validation and holdout evaluation. Trained models were applied to the historically disputed Federalist Papers. Experimental results show strong classification performance across all models with Logistic Regression and Support Vector Machines achieving the highest accuracy and most consistent predictions. The findings demonstrate that combining stylometric features with TF-IDF-based linear classifiers is an effective approach for authorship attribution.

Keywords—authorship attribution, stylometry, Federalist Papers, TF-IDF, text classification, machine learning, linear classifiers

I. INTRODUCTION

Authorship attribution is the task of determining the most likely author of a document based on linguistic and stylistic characteristics. Computational methods can distinguish between authors by identifying patterns in language use and provide evidence for disputed or unknown authorship.

The Federalist Papers are a well-established benchmark dataset in authorship attribution research in part to their historical significance and disputed authorship between James Madison and Alexander Hamilton. The majority of essays have been confidently attributed to Madison, Hamilton or John Jay although historically the authorship of several papers was disputed between Madison and Hamilton. As a result, the Federalist Papers have been widely used to evaluate stylometric, statistical and machine learning approaches to authorship attribution[1].

Traditional authorship attribution research has relied on stylometry and examines features like function-word usage, vocabulary size, sentence structure and punctuation patterns. Machine Learning techniques have more recently been applied to text classification using vector space representations such as Term Document Inverse Document Frequency (TF-IDF). These approaches present authorship attribution as a supervised learning problem and provide a basis for comparing classification methods [2], [3].

This study investigates authorship attribution of the Federalist Papers using both stylometry and machine learning text classification techniques. A corpus of known Madison and Hamilton papers was assembled and used to train Multinomial Naïve Bayes, Logistic Regression and Support Vector Machine classifiers. Model performance was evaluated using cross-validation and holdout testing. The trained classifiers were used to attribute authorship of the disputed papers and assessed for consistency. The results demonstrate strong classification performance across all models. Logistic Regression and Support Vector Machine models achieve the highest accuracy and produce attribution results consistent with modern consensus.

II. LITERATURE REVIEW

A. Authorship Attribution and the Federalist Papers

Authorship attribution is a well-established problem in computational linguistics focusing on identifying the most likely author of a text based on stylistic and linguistic patterns. The Federalist Papers represent a canonical dataset in this domain due to their historical significance and disputed authorship between James Madison and Alexander Hamilton[1].

The papers were originally published under the pseudonym “Publius” and have been extensively analyzed to determine individual authorship contributions. Although the disputed papers have since been largely resolved through historical and computation analysis, the Federalist Papers remain a benchmark for evaluating stylometric and statistical authorship attribution methods[1].

Early computational studies demonstrated that authorship can be inferred from consistent stylistic patterns even when topics are highly similar across documents. These findings establish the Federalist Papers as a foundational case study for authorship attribution research.

B. Classical Statistical Approaches to Authorship Attribution

Early authorship attribution techniques treated the task as a statistical inference problem based on differences in word usage patterns. These methods relied heavily on frequency-based representations of text and assume authors have consistent writing styles[2].

Function words like articles, prepositions and conjunctions are often used for identifying authors. These words carry minimal topical meaning and their usage patterns tend to reflect stable stylistic choices rather than subject matter making them effective features for classification tasks[2].

Bayesian methods such as Naïve Bayes classifiers have been used in authorship attribution and estimate the likelihood that a text was written by a particular author by modeling the probability of an author given observed word distributions. These simple, effective methods can serve as baseline models for text classification and stylometry[1].

These statistical methods establish the foundation for frequency based and probabilistic approaches to authorship attribution and remain relevant in modern machine learning applications.

C. Stylometric Feature Engineering

Stylometry is the quantitative study of writing style and has long been used a foundation for authorship attribution research. Instead of focusing on the content of a document, stylometry seeks to identify consistent linguistic habits that distinguish one author from another. Early studies of the Federalist Papers demonstrated that stylistic patterns can provide strong evidence of authorship[1].

Common stylometric features include function-word frequencies, vocabulary size, type-token ratio, sentence length and punctuation usage patterns. These features are designed to capture stable aspects of writing styles across documents and are less dependent on writing styles [1], [4].

These features represent author-specific writing habits and continue to play an important role in modern authorship research. Traditional stylistic measures are often used in addition to machine learning techniques to provide interpretable insights into differences between authors and to serve as predictive features for classification models[4].

D. Vector Space Models and TF-IDF Representations

Text classification methods commonly represent documents using bag-of-words vectors. Term frequencies are transformed into high-dimensional feature space suitable for statistical learning. Term Frequency Inverse Document Frequency weighting is often applied to reduce the influence of frequently occurring but less informative terms and emphasizing words that are more distinct across documents. This approach has been shown to be effective in a wide range of text classification tasks[2], [3].

Bag-of-words and TF-IDF representations produce sparse, high-dimensional feature matrices that are well-suited for linear classification models. They are able to capture sufficient lexical despite being simple and are common input representations for classifiers such as Logistic Regression and Support Vector Machines [2], [3].

E. Linear and Probabilistic Models for Text Classification

Probabilistic and linear models are widely used for text classification and perform particularly well with sparse, high-dimensional representations like TF-IDF vectors [2], [3].

Multinomial Naïve Bayes is a probabilistic classifier that estimates class membership based on observed term frequencies and conditional probability distributions. It is a frequently used as a baseline model for text classification and author attribution due to its simplicity and computational efficiency[2].

Logistic Regression and Support Vector Machines have been historically two of the most successful linear approaches for document classification. Logistic Regression produces probabilistic predictions while learning a decision boundary. Support Vector Machines maximize the margin between classes to improve generalization performance. These classifiers are particularly well suited for high-dimensional text data and effective when paired with TF-IDF[2], [3]. They also serve as the foundation for experimental evaluation in this study.

III. PROPOSED METHODOLOGY

A. Build Corpus

The Federalist Papers corpus was collected from Project Gutenberg and used as the primary data source for analysis[5]. The downloaded text was segmented into individual papers using headers and document markers. For each paper, the paper number, raw text and attributed author were extracted and stored in a structured DataFrame. This process produced a structured corpus of individual essays that could be further processed and analyzed.

B. Label Disputed Papers

Authorship of a subset of the Federalist Papers has been disputed historically between James Madison and Alexander Hamilton making these documents suitable for authorship attribution analysis. Papers 49-58 and 62-63 were identified as disputed based on literature review. A binary indicator was added to structured DataFrame to distinguish disputed papers from those with known authorship. This labeling allowed the exclusion of disputed papers from the training dataset and their use as a hold-out evaluation set for final attribution experiments.

C. Preprocess Text

Raw text was preprocessed to reduce noise introduced by formatting and punctuation inconsistencies across documents allowing for its use in machine learning analysis. Text cleaning operations included the removal of headers and metadata, as well as normalization of whitespace and punctuation. These steps ensured that meaningful textual content was retained for analysis.

All text is further normalized through lowercasing. This standardization reduced variability caused by capitalization and formatting differences while preserves stylistic features. The output of this process was a clean, structured textual representation of each Federalist Paper which served as input to the Feature Extraction and Modeling steps.

D. Extract Linguistic Features

Natural Language Processing was performed using the spaCy library to create structured linguistic representations of each document. spaCy was selected for its efficient pipeline for tokenization, sentence segmentation and linguistic annotation making it well-suitable for stylometry.

Each document was processed to extract token-level and sentence-level representations. Tokenization identified individual word units and sentence segmentation divided each paper into coherent sentence boundaries enabling lexical and structural feature analysis.

Lemmatization was then applied to reduce words to their base form and minimize morphological variation. The resulting output consisted of structured spaCy objects including tokens, sentences and lemmas which were used to engineer stylometric features.

E. Engineer Stylometric Features

Stylometric features are constructed in addition to the TF-IDF representations to capture author-specific writing patterns independent of lexical content. These features are widely used in authorship attribution tasks due to their ability to reflect consistent stylistic habits rather than topic-specific vocabulary[3].

Lexical features were computed to capture word usage patterns including word count, vocabulary size, type-token ratio and average word length. Syntactic features average sentence length, sentence length standard deviation and average paragraph length were used to characterize writing structure and complexity.

Punctuation usage rates and function-word frequencies are also computed to measure stylistic preferences. Function-words are specifically valuable in authorship attribution because their usage tends to stay stable across topics and reflects individual writing style[1].

These stylometric features provide complimentary information to the TF-IDF representations by capturing structural and behavioral elements of writing style not directly encoded in word-based frequency models.

F. Define Feature Views

Two complementary data representations were constructed in order to support different experimental settings. This separation allowed for stylistic analysis and predictive modeling to be conducted using appropriately structured inputs.

A stylometric feature view was created to store linguistic and structural features including lexical, syntactical, punctuation-based and function-word based statistics. This representation was primarily used for exploratory data analysis and blah blah evaluation of stylistic differences between authors.

A separate text-based modeling view was also constructed with the clean text corpus. This representation served as the input for machine learning pipelines in TF-IDF vectorization and classification models.

The separation of these two views ensured a distinction between feature engineering for interpretability and feature engineering for predictive modeling.

G. Partition Data

A train-test partitioning strategy was used to ensure robust model evaluation and prevent information leakage ensuring performance metrics reflected generalization to unseen data. The train corpus consisted of exclusively undisputed papers authored by James Madison and Alexander Hamilton. These documents formed the labeled dataset used for model training and validation.

An 80/20 stratified split was applied to the training corpus to preserve the proportional representation of each author across

training and validation sets. This ensured that class imbalance did not bias the model evaluation during cross-validation.

The disputed Federalist Papers were excluded from the training process and reserved as a hold-out set for final authorship inference. This hold-out strategy ensured no information from disputed papers influenced model learning or hyperparameter selection.

This partitioning approach prevents evaluation bias and ensures final authorship attribution results are derived from models trained only on texts with known authorship.

H. Train Attribution Models

The objective of the modeling stage was to predict authorship of a given Federalist Paper based on its textual representation. This was formulated as a supervised machine learning classification task using known-authorship documents.

Three machine learning approaches were evaluated under a unified experimental framework: Multinomial Naïve Bayes, Logistic Regression and Support Vector Machines. Each model was trained using the TF-IDF representations of the text to ensure consistent input features across all classifiers.

All models were trained and evaluated under identical data partitioning preprocessing conditions to ensure fair comparison. This controlled setup allowed performance differences to be attributed to model behavior as opposed to differences in input representation or data sampling.

The performance of each model was assessed using cross-validation and held-out evaluation metrics including accuracy, precision, recall, F1-score, ROC-AUC and log-loss. Detailed experimental results and comparisons are presented in the following section.

IV. EXPERIMENTS AND RESULTS

A. Stylometry

Exploratory data analysis examined descriptive statistics and word frequency distributions across the corpus to characterize stylistic differences between authors. Table I shows the average values per author and reveals stylistic patterns relevant to authorship attribution.

TABLE I. AUTHOR DESCRIPTIVE STATISTICS

Statistic	Author	
	<i>Madison</i>	<i>Hamilton</i>
Word Count	3100	2464
Vocabulary Size	807	717
Sentence Count	83	67
Punctuation Rate	0.11	0.10
Average Word Length	4.90	4.79
Average Sentence Length	34.29	33.82
Average Paragraph Length	119.10	112.08
Sentence Length Std. Dev.	21.22	20.21

The results indicate Hamilton wrote shorter papers with an average word count of 2464 compared to Madison at 3100. Vocabulary size reveals that Madison also used 90 more unique words suggesting a broader range of lexical diversity in his papers. Hamilton however had a higher type-token ratio indicating he used more unique words relative to document length.

There were also structural characteristics shared by the two authors. The difference in average sentence length was less than one word and average word length and punctuation rate reveal minor differences. The similarities identified suggest that stylistic measurements by themselves may not be enough to attribute authorship.

There are some differences in the way the papers are organized. Madison wrote slightly longer paragraphs and had higher sentence length standard deviation. These subtle variations may be useful signals for supervised machine learning models.

The descriptive statistics suggest that Madison and Hamilton have many writing style characteristics in common although subtle differences in document organization. These findings motivate the use of machine learning techniques to identify those complex patterns that cannot be observed with descriptive statistics alone.

B. Function-Word Frequency Analysis

Function-Word Frequency Analysis further investigates the stylistic differences between authors. Function Words are often analyzed in authorship attribution because they can reveal patterns in an author's unconscious writing habits and are less influenced by document subject matter and content-specific vocabulary. Frequencies were computed for each author and compared.

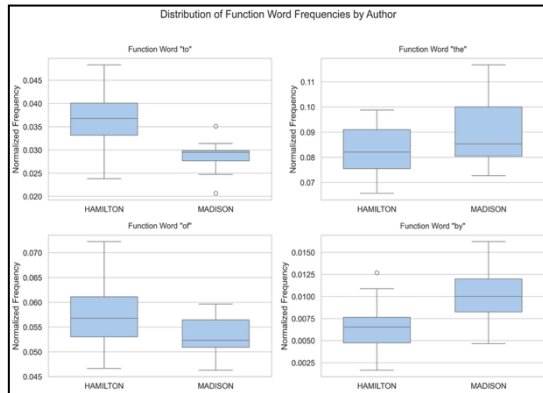


Fig. 1. Distribution of Function Word Frequencies by Author.

The analysis revealed a few distinctions between the two authors. Madison used “the” and “by” more often than Hamilton while he used “to”, “of” and “in”. The largest usage difference between the two authors was for the word “to”, appearing in Hamilton’s papers 3.66% compared to 2.86% in Madison’s papers.

The absolute mean differences were relatively small although consistent across the papers. This result is consistent

with prior stylometric research demonstrating that function word usage often provides discriminative signals for authorship attribution[1]. The observed usage pattern suggests a meaningful stylistic differences between the two authors despite having the same historical context and subject matter. These findings provide evidence that authorship signal exists within the corpus and motivation to apply supervised machine learning techniques to classify authorship.

C. Naïve Bayes Classification

A multinomial Naïve Bayes classifier was evaluated as a baseline for authorship attribution using TF-IDF feature representations and stratified five-fold cross-validation. The model demonstrated strong classification performance summarized in Table II.

TABLE II. CROSS-VALIDATION RESULTS

Accuracy	0.87 +/- 0.06
Precision	0.85 +/- 0.13
Recall	0.88 +/- 0.06
F1-Score	0.85 +/- 0.10
ROC-AUC	0.96 +/- 0.05
LogLoss	3.90 +/- 2.21

The model performed well separating the two authors. The confusion matrix shows that the majority of documents were correctly classified for both authors with few misclassifications. These results indicate that a basic probabilistic model can capture meaningful differences in writing style.

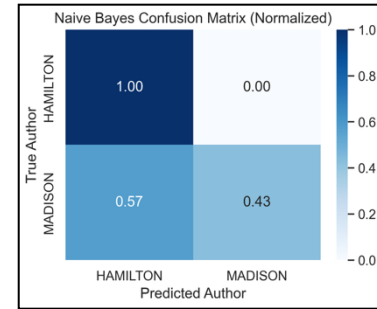


Fig. 2. Confusion Matrix (Normalized) – Naïve Bayes.

The model produced very high probability scores when attributing authorship indicating it is overconfident in its predictions. This is a documented limitation of Naïve Bayes when working with high-dimensional sparse text data. Small differences in word usage can result in large probability shifts[2].

D. TF-IDF and Logistic Regression

The Logistic Regression model displayed strong classification performance. 11 out of 14 documents authored by Madison were classified correctly while 3 were mislabeled Hamilton. All Hamilton papers were classified correctly.

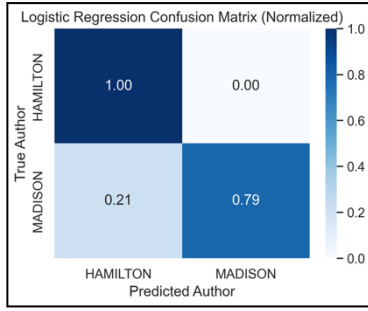


Fig. 3. Confusion Matrix (Normalized) – Logistic Regression.

This results in a perfect recall for Hamilton and slightly slower for Madison indicating a slight bias towards choosing Hamilton although overall misclassification is low. The high test accuracy and F1 score confirm the Logistic Regression classifier is effective at distinguishing between writing styles using TF-IDF representation.

TABLE III. CROSS-VALIDATION RESULTS

Accuracy	0.95 +/- 0.09
Precision	0.92 +/- 0.13
Recall	0.95 +/- 0.09
F1-Score	0.93 +/- 0.13
ROC-AUC	1.00 +/- 0.00
LogLoss	0.62 +/- 0.01

E. TF-IDF and Support Vector Machine

The Support Vector Machine model achieved the strongest classification performance among the evaluated techniques. 13 out of 14 documents authored by Madison were classified correctly with only 1 being misclassified. All papers authored by Hamilton were classified correctly.

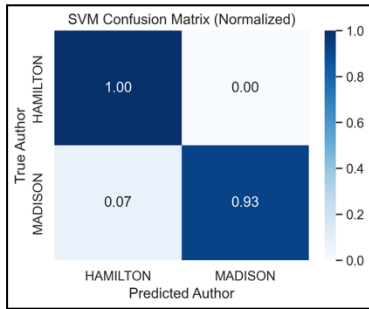


Fig. 4. Confusion Matrix (Normalized) – Support Vector Machine

This resulted in a perfect recall for Hamilton and a near-perfect recall for Madison. Overall misclassification was extremely low and the high accuracy, precision, recall and F1-score confirm the SVM classifier is highly effective at distinguishing between Hamilton's and Madison's writing styles using TF-IDF feature representations.

TABLE IV. CROSS-VALIDATION RESULTS

Accuracy	0.95 +/- 0.09
Precision	0.92 +/- 0.13
Recall	0.95 +/- 0.09
F1-Score	0.93 +/- 0.13
ROC-AUC	1.00 +/- 0.00
LogLoss	0.62 +/- 0.01

F. Disputed Paper Attribution

Predictions from the Naïve Bayes, Logistic Regression and Support Vector Machine classifiers were aggregated and compared across the disputed documents, papers 49-58 and 62-63. Table 5 summarizes class predictions and confidence scores for each model.

TABLE V. AUTHORSHIP CLASS PREDICTIONS

Paper No.	Model		
	Naïve Bayes (NB)	Logistic Regression (LR)	Support Vector Machine (SVM)
49	Madison	Madison	Madison
50	Madison	Madison	Madison
51	Madison	Madison	Madison
52	Madison	Madison	Madison
53	Hamilton	Madison	Madison
54	Madison	Madison	Madison
55	Madison	Hamilton	Madison
56	Hamilton	Madison	Madison
57	Hamilton	Madison	Madison
58	Madison	Madison	Madison
62	Madison	Madison	Madison
63	Madison	Madison	Madison

All 3 classifiers attributed the majority of the authorship of the disputed papers to James Madison. Logistic Regression and Support Vector Machine models showed strong agreement in nearly all documents, disagreeing only on paper 55 while the Naïve Bayes classifier deviates further on papers 54, 56 and 57. This corresponds to unanimous agreement on authorship of 9 out of 12 disputed papers across all models.

TABLE VI. MODEL AGREEMENT SUMMARY

	Model Pair		
	NB vs. LR	NB vs. SVM	LR vs. SVM
Agreement Rate	0.67	0.75	0.91

The confidence of the classification varies across models and in most cases favor Madison with the documents near the decision boundary. This result suggests the documents are not

strongly separable in the higher dimensional feature space but still bias towards Madison.

These findings are consistent with recent research and scholarly findings attributing authorship of the disputed Federalist Papers to James Madison. The strong agreement between Logistic Regression, Support Vector Machine and in many cases the Naïve Bayes model provides evidence to support this claim within the TF-IDF-based stylometric framework used here.

V. CONCLUSION

This study investigated authorship attribution of the Federalist Papers using a combination of stylometric feature analysis and machine learning text classification techniques. A structured corpus of papers authored by James Madison and Alexander Hamilton was constructed and used to train Multinomial Naïve Bayes, Logistic Regression and Support Vector Machine classifiers. Stylometric features we extracted to capture lexical, syntactic and structural patterns in writing style.

Strong classification performance was observed across all models with Logistic Regression and Support Vector Machines

achieving the highest accuracy and most consistent predictions. The models also demonstrate strong agreement in attributing the Federalist Papers primarily to James Madison which aligns with modern consensus of authorship.

The study is limited by the size of the dataset and restriction to a binary classification between the two authors. Future work could expand on this framework to include richer stylometric features or explore neural language models for comparison to classical machine learning techniques. These findings demonstrate an effectiveness of combining stylometry with TF-IDF-based machine learning models for authorship attribution.

REFERENCES

- [1] F. Mosteller and D. L. Wallace, *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*, 2nd ed. New York, NY, USA: Springer-Verlag, 1984.
- [2] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [3] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. draft, Stanford University, 2023. Available: <https://web.stanford.edu/~jurafsky/slp3/>.
- [4] N. M. Demir, "A study in authorship attribution: The Federalist Papers," *Southeast Europe Journal of Soft Computing*, vol. 4, no. 1, pp. 29–34, 2015. [Online]. Available: <https://doi.org/10.21533/scjournal.v4i1.86>