**Who Wrote the Federalist Papers?**

**Part 2**

Ben Harwood

**Introduction**

During the recent impeachment trial, members of both the House prosecutors and the President’s defense team repeatedly referenced the Federalist Papers. These 85 articles and essays were written as an exhaustive explanation and defense of the proposed Constitution that had been submitted to the states for ratification in 1787. There had been a collection of articles and letters written in criticism of the Constitution, and it was Alexander Hamilton that sought to win over the hearts and minds of the “Anti-Federalists” as they became known.

Hamilton was not alone in this endeavor. He initially enlisted John Jay who wrote four of the first five essays (Jay did write one of the later essays, but this was his last due to illness). Hamilton and Jay also recruited James Madison, who contributed numerous papers and became Hamilton’s primary collaborator. The articles were published in New York newspapers beginning on October 27, 1787. They were published under pseudonyms because the men had attended the Constitutional Convention, though many people were able to determine the authors.

Hamilton provided a list that associated the authors with the paper numbers when demand for a more official publication of the collected papers arose. The authorship of seventy-three of the articles is generally well-established, but there is some debate about the remaining twelve among scholars. This is partly due to Madison providing a different list of authorship that disputed Hamilton’s list[[1]](#footnote-1), indeed it known that Hamilton attributed authorship of paper 54 to John Jay, when in fact Jay wrote paper 64. Much, much more can be said about the Federalist Papers however the purpose of this article is to use data mining techniques to see if the authorship of the disputed papers can be established.

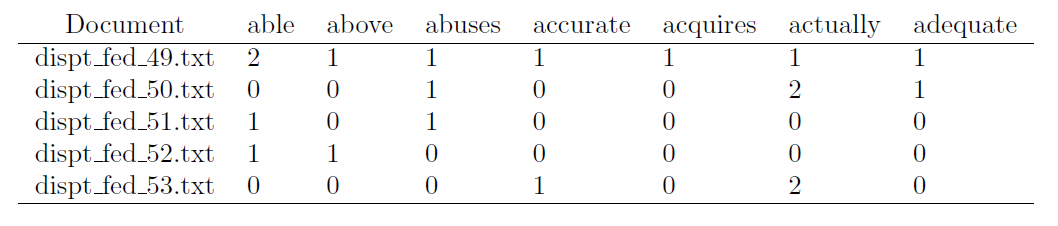
This study is a continuation of the author’s previous examination of the Federalist Papers, this time using a different data mining technique to confirm (or refute) the previous findings.

**Section 1: Analysis and Models**

As mentioned in the introduction, there are 85 papers total in the collection. The idea of the study is to look at the word usage in each of the papers to see what similarities in writing styles and vocabulary exist between them. As in the previous study, the documents will be turned into a collection of values measuring the frequency each word appears in each document, but where this study goes its own way is that instead of dividing the papers into clusters based on word frequency a sort of tree will be formed (aptly called a decision tree) by splitting the data through a series of basically yes or no questions.

**Section 1.1: Data Cleaning and Preparation**

Thankfully, the Federal Papers are widely available. To begin the study, the papers are collected together into a digital corpus[[2]](#footnote-2) and then a so-called “document term matrix” (DTM going forward) is formed with the content of the papers. The DTM is a matrix where each row is a different document, and the columns are the individual words from any of the papers. The elements of the matrix are the frequencies in which the words appear in the documents. The below example shows the word counts for the first seven (alphabetically) words that appear in any of the papers[[3]](#footnote-3).



Certain very commonly used words (such as “I”, “it”, “is”, etc.) are left out of this formation, as they frequently appear in everyday writing or conversation, and their inclusion would not improve the chances of determining the author of the disputed papers.

The values in the matrix are then normalized to reflect the frequency of the different words relative to the paper they appear in. This is done by dividing the number of times the word appears in a certain paper by the total number of words in said paper. This effectively puts the different words on an even playing field because if the word “constitution” appears 40 times in a paper A and 12 times in paper B, those 50 times in A could be much less impactful than the 12 times in B, and normalizing reflects this. Additionally, because it is known that John Jay wrote five of the papers and that three of the papers were cowritten by Hamilton and Madison, those eight documents should not be part of the analysis. And to allow for identification, another variable was created for the author of each paper, either “disp”, “ham”, or “mad”.

To get a feeling for each author’s word usage, here are word clouds for each author.

A close up of a piece of paper

Description automatically generated

Figure : Disputed Author word cloud

A close up of a piece of paper

Description automatically generated

Figure : Hamilton word cloud

A close up of a logo

Description automatically generated

Figure : Madison word cloud

**Section 1.2: Analysis**

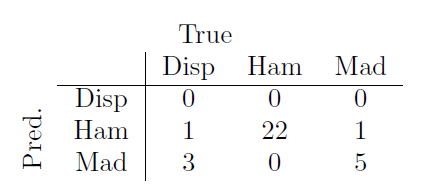
Decision trees are a supervised learning technique. By this it is meant that a model is developed based on a subset of a dataset, the so-called “training data” and then this model is applied to the remaining data, the “test data”. These two subsets should be chosen without bias, so a random sample subset of 60% of the papers was selected as the training data. It consisted of 7 disputed papers, 32 Hamilton, and 9 Madison.

The first tree that was produced was… less than revealing.

A picture containing object, clock

Description automatically generated

Notice that no branches terminated with a disputed author, giving no indication of who may have authored them. The following confusion matrix comparing the accuracy of the authors in the test set to the predicted values based on this tree also indicates that three of the disputed papers were predicted to be authored by Madison and one by Hamilton.

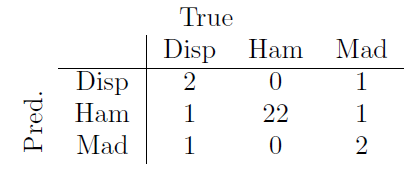


To this end, a second tree was developed, this time with a set minimum number of observations required before a splitting can occur (2, to be specific). This resulted in the followed, much more enlightening tree:

A close up of a map

Description automatically generated

Notice here that the disputed papers do appear as terminal nodes, and more specifically that they are direct splits from a split that resulted in Madison. Here is the confusion matrix for this model:



But here something interesting can be observed: the first model was actually more accurate in its predicting power than the second. Notice that 2 of the disputed papers were correctly predicted as disputed while the other 2 were split between Hamilton and Madison. Notice also that one Madison paper was predicted as a Hamilton paper, and one Madison paper was predicted as a disputed paper. Thus, while the second tree showed more information and seemed to paint a better picture, the model created by the first tree more accurately predicted the authorship of the papers.

**Section 2: Results**

In the previous section, decision tree models were developed to help confirm or refute the findings of the previous study to determine who wrote the 11 disputed Federalist Papers. The first model predicted authors for each of the four disputed papers in the test set (three for Madison and one for Hamilton), while the second model predicted that two of the disputed papers were still disputed, and one was authored by each of the two men. The second model was also less accurate than the first. These results somewhat corroborate the results of the previous study.

Finally, using all of the confirmed authorship papers as the training set, a decision tree model was developed that predicted 10 of the 11 disputed papers were written by Madison, including papers 53 and 56 (that were not able to be predicted last time) and paper 54 was written by Hamilton though the clustering techniques previously used determined Madison authored paper 54.

**Conclusion**

At the risk of sounding like a broken record., it was noted in the introduction that much more could be said about the Federalist Papers. For instance, paper 84 is notable in that it is an opposition to what would come to be the Bill of Rights, one of the foundational notions of the US Democracy. The gap left from the previous study (papers 53 and 56, which ironically both pertain to the House of Representatives) appears to be settled. Potentially, at least, as the techniques in this study contradict the result of the last study regarding the authorship of paper 54.

It is somewhat widely accepted that papers 18-20 were collaborative efforts between Hamilton and Madison. There is the possibility, of course, that more of the papers than these three were indeed collaborations, and the contradictory results of this study and the previous one lends some merit this to possibility. A study for another time, perhaps.

1. Hamilton provided his list to his lawyer mere days before his infamous duel with Aaron Burr, and Madison suggested the discrepancies in the list were due to this [↑](#footnote-ref-1)
2. Corpus: a collection of written texts, especially the entire works of a particular author or a body of writing on a particular concept [↑](#footnote-ref-2)
3. After removing “stop words” such as ‘a’, ‘is’, ‘it’, etc., as well as excessively lengthy words. Also, all words are converted to lowercase, and numbers are removed [↑](#footnote-ref-3)