

Fixing the Federalist: Correcting Results and Evaluating Editions for Automated Attribution

Shlomo Levitan (levishl@iit.edu) and Shlomo Argamon (argamon@iit.edu)
Linguistic Cognition Laboratory
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616, USA

Introduction

In the history of authorship attribution, the analysis of The Federalist Papers plays an important role. However, most previous non-traditional (stylistic) authorship studies have been flawed, mainly due to the use of improper editions, as documented by Rudman (2005). Our goal in this work was to perform a correct study by using a revised corpus of the Federalist papers based in large part on Rudman's critique. We used machine-learning techniques for analyzing the use of lexical features for authorship attribution of the papers. Another goal of our study was to explore how different corruptions of the corpus may affect the accuracy of the classification results, and the differences between them.

The Federalist Papers were written during the years 1787 and 1788 by Alexander Hamilton, John Jay and James Madison. These 85 propaganda tracts were intended to help to get the U.S. Constitution ratified, and were all published anonymously under the pseudonym "Publius". According to Avalon project (Yale Law School) Hamilton wrote 51 of the papers, Madison wrote 15, Jay wrote five, while three papers were written jointly by Hamilton and Madison, and 11 papers have disputed authorship – either Hamilton or Madison, although most evidence points to Madison as the author.

The Federalist Papers have presented a classic research problem for authorship attribution, and many studies on them have been done since Mosteller and Wallace's seminal attribution study (1964). Professor Joseph Rudman (2005) argues that in previous studies, textual flaws in the corpus were not properly addressed. To summarize Rudman's critique, problems in previous versions of the corpus included: Wrong letters and misspellings, due to letters typed by mistake or typos; inconsistency in inclusion of greetings and signatures in the papers; inclusion of foreign language words and quotations from other sources in the analysis; and inconsistent treatment of footnotes. Therefore, there is a doubt regarding the results of previous studies based on a flawed corpus.

The Corpus

We constructed a new corpus based on the one made available by the Avalon project, which is a collection of various papers that are related to United States history available on-line under the supervision of Yale Law School. The Avalon version is currently the most accurate on-line version of the Federalist Papers. To create a more accurate edition of the Papers, we compared the Avalon corpus to the edition of the Federalist book collection (*THE FEDERALIST: A COLLECTION OF ESSAYS*, 1788, Special Collection, Northwestern University) and corrected the flaws mentioned above. We used XML tags to group words by marking them specifically in the corpus in order to include or exclude them while processing the corpus. This option was used for comparison between the corrupt and corrected versions of the corpus or between the corpus with or without the groups. For example, footnotes, the remarks that the editor included in the corpus, were ignored by using XML tags while processing the corrected version. We addressed several problems in the corpus wrong letters and spelling in the papers, which were marked and fixed, according to the original source. Quotes and footnotes, were marked and fixed according to the original source in order to process the text with and without the marked data. Endings and openings that marked the text openings ("the People of the State of New York:" etc.) and the endings of the papers ("PUBLIUS"), were used in order to process the text with and without the

marked data. These corrections address most of Rudman's criticism. In this work we disregarded punctuation because processing of the corpus was done automatically by extracting words, and this processing method doesn't take any punctuation issues into consideration.

Attribution Study

After we fixed the corpus, we counted various words to compute feature values. This operation included the extraction of frequent words and frequent collocations from the fixed corpus. We defined frequent words as the k most frequent words in the corpus where k is a parameter of the system. Frequent collocations (Argamon) and (Hoover) were defined as pairs of words occurring within a given threshold distance (window size) between them (for example, "for", "are" and "sure" appearing within 10 words of each other in a sentence, with a window size of 10). Given such a threshold, the most frequent such collocations were determined over the whole corpus. Given each particular feature set (frequent words, or collocations), the method was used to represent each document as a numerical vector, each of whose elements is the frequency of a particular feature of the text. For example, the word "the" was represented in a numerical vector as the number of times it occurs in the text divided by the number of words in the text. We then applied the SMO learning algorithm (Platt) with default parameters, which gives a model linearly weighting the various text features. SMO is a support vector machine algorithm; these have been applied successfully to a wide variety of text categorization problems (Joachims).

To analyze our results and to examine the accuracy of the classification we used 10-fold cross-validation, which is a common and reliable technique, to examine the generalization error of the model. 10-fold cross-validation divides the whole data set into 10 subsets of equal size; trains 10 times, each time leaving out one of the subsets from training, and then averages the results.

To find out the effect of corruption of the corpus, we performed 10-fold cross validation on the known Hamilton and Madison documents in both corrupt version and the corrected version of the corpus and compared the accuracy of results. We also examined the consistency attribution by building a model using all of the known training documents and classifying the 11 disputed papers, comparing the attributions derived from the corrected corpus as well as various corrupt versions of the corpus and various features sets. This experiment allowed us to evaluate the effect of a corrupt corpus and to evaluate the benefit of creating a corrected corpus as suggested by Rudman (2005).

Results

The 10-fold cross validation experiment on the corrected corpus vs. the corrupt corpus produced results that allowed us to examine which corpus is better suited for further research. Our findings are described in figures 1, 2 and 3 indicate that the experiments performed on the corrected corpus using different feature sets produced either slightly more accurate or similar results to the ones produced on the corrupt corpus. Our findings support the argument that the corrected corpus is a better source for further study of the federalist papers attribution problem.

We also addressed the authorship attribution problem of the 11 disputed papers, by building a model using the known papers as training documents and classifying the 11 disputed papers. Our results are summarized in figure 4 and show that the experiment conducted on both corpora, the corrected and the corrupt, produced the same results. The results clearly attribute the authorship of the disputed papers to Madison. Our findings are thus consistent with the universal accepted allocation that disputed papers were authored by Madison (Carey and McClellan).

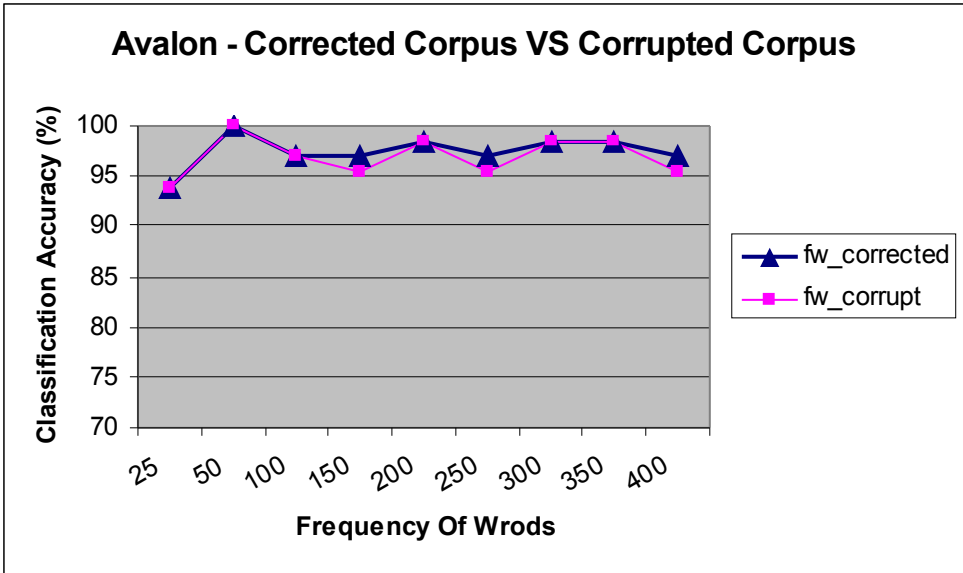


Figure 1. Classification Accuracy results for 25 – 400 most frequent words on Avalon corrected corpus VS. Avalon corrupt Corpus.

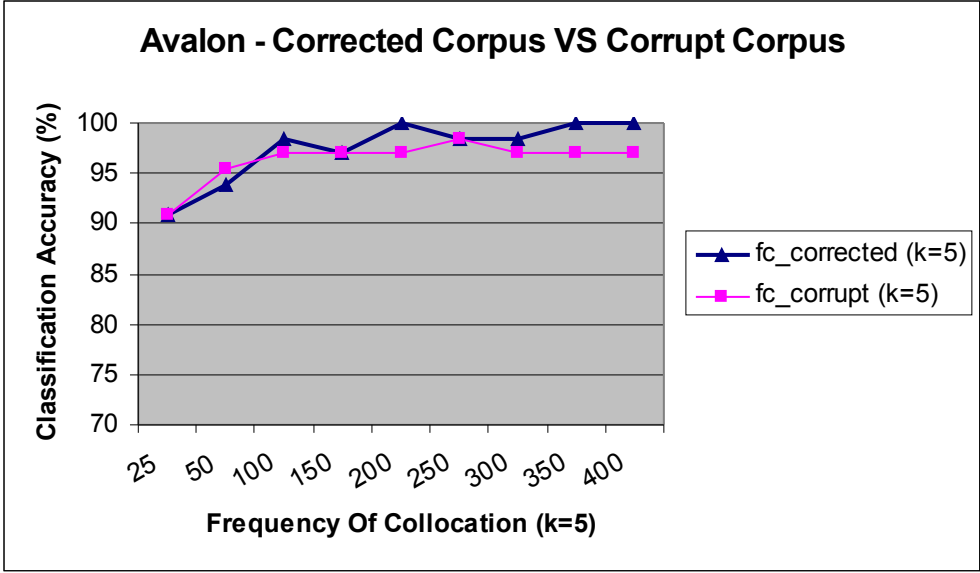


Figure 2. Classification Accuracy for 25 – 400 most frequent collection (window size of 5) on Avalon corrected corpus VS. Avalon corrupt corpus.

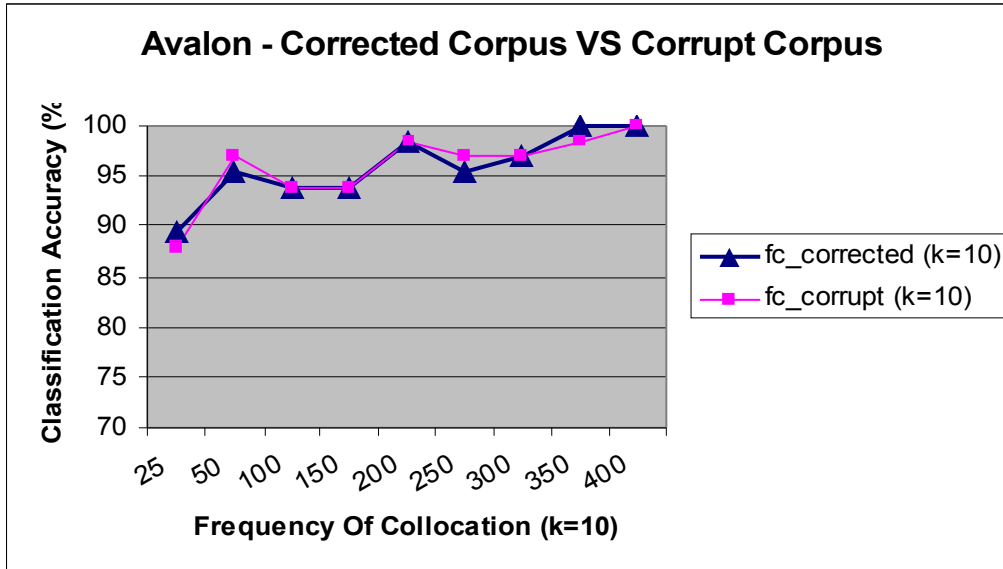


Figure 3. Classification Accuracy for 25 – 400 most frequent collection (window size 10) on Avalon corrected corpus VS. Avalon corrupt corpus.

Frequency Of Words	Papers Classified as Madison's Corrected Corpus	Papers Classified as Madison's Corrupt Corpus
25	10	10
50	10	10
100	10	10
150	11	11
200	11	11
250	11	11
300	11	11
350	11	11
400	11	11

Figure 4. Disputed papers classification for most frequent words on Avalon corrected corpus VS. Avalon corrupt corpus.

Discussion

We have constructed a more accurate corpus of the Federalist papers, which was corrected and is now available for further research. Furthermore, the evaluation and analysis on both corpora were done by using modern machine learning methods that are completely automated. Our results show that using the corrected corpus for authorship attribution studies produces slightly better results than the corrupt corpus. Thus, our corrected corpus may be a better source for further study of the federalist papers attribution problem.

Moreover, this study provides additional support to the almost universally accepted allocation that Madison is the author of the disputed Federalist papers. We will be making our new corpus available for public use and we hope that it will become a useful tool for the research community in the future.

Conclusions

The corrected corpus of the Federalist papers that we have generated in this study was found to

be a better source for further study of the Federalist papers attribution problem than the corrupt version. The experiments we conducted on the corrected and the corrupt corpus by using different feature sets provided results that support this argument. Furthermore, our study supports the universal opinion that Madison is the author of the disputed Federalist papers.

Acknowledgments

The authors would like to thank Joseph Rudman for his advice on all aspects of this project.

References

Rudman, J. (2005). *The Non-Traditional Case for The Authorship of the Twelve Disputed "Federalist" Papers: A Monument Built on Sand*. Association for Computers and the Humanities and the Association for Literary and Linguistic Computing.

Rudman, J. (2005). *Unediting, De-editing, and Editing in Non-traditional Authorship Attribution Studies: With an Emphasis on the Canon Of Daniel Defoe*. The Papers of the Bibliographical Society of America, 99.1 pp 5-36

Mosteller, F. and **Wallace, D. L.** (1964). *Inference and Disputed Authorship: The Federalist*. Reading, Mass.: Addison Wesley.

Avalon project Yale Law School (<http://www.yale.edu/lawweb/avalon/federal/fed.htm>)

THE FEDERALIST: A COLLECTION OF ESSAYS (As agreed upon by the federal convention September 17, 1787, in two volumes). Publisher: New-York: Printed and sold by J. and A. M'Lean 1788. The source was found in Library of Special Collections/Northwestern University.

Argamon, S., Levitan S., (2005). *Measuring the Usefulness of Function Words for Authorship Attribution*. Association for Computers and the Humanities and the Association for Literary and Linguistic Computing.

Hoover, D.L. (2004). *Frequent collocations and authorial style*. Literary and Linguistic Computing 18.3 261-282

Goutte, C. (1997). *Note on free lunches and cross-validation*, Neural Computation, 9(6):1246-9.

Joachims, T. (1998). *Text categorization with Support Vector Machines: Learning with many relevant features*. In Machine Learning: ECML-98, Tenth European Conference on Machine Learning, pp. 137-142.

Platt, J. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*, Microsoft Research Technical Report MSR-TR-98-14.

The federalist, the Gideon edition, Edited by George W. Carey and James McClellan.