Author Attribution of Thomas Paine Work

Smiljana Petrovic, Gary Berton, Robert Schiaffino, Lubomir Ivanov

***Abstract*—Thomas Paine is one of the most significant historical figures who, through his political, philosophical, and socio-economic writings influenced – and continues to influence – the course of history. While Paine's major works are widely known, there are period writings of unknown or disputed authorship which may be attributed to Paine.**

The main goal of this project is to develop a methodology for automated authorship attribution, and apply it to documents of disputed origin. The results generated by the software are cross-referenced with facts about Paine's life and work by experts in Paine studies and 18th-19th century literature and history.

The authorship attribution is performed using machine learning software, trained through the use of works of undisputed authorship to recognize unique features of Paine's style compared to other authors of the period. Once trained, the software is applied to documents of questioned authorship, yielding probabilistic results, further verified by human experts. The results have been both surprising and inspiring: For some disputed documents the software strongly points to Paine as the author. In other cases Paine's previously presumed authorship has been refuted. These results will help better understand Paine's impact on literature, philosophy, history, and politics.

Keywords: Authorship Attribution, Thomas Paine, Machine Learning, Interdisciplinary

I. Introduction

THOMAS Paine is an inherently controversial historical figure whose story has been shrouded in disinformation. Sentenced to obscurity after his death by a power structure that feared him, Paine is the most important historical actor to have been marginalized by academia. Barely mentioned by leading historians for over 200 years when writing about the American and French Revolutions, scholarship on Paine was left largely unexplored. Academic interest in Paine began in earnest in the 1960's, but drew initially upon faulty biographies written to discredit Paine or upon over-enthusiastic biographies by Paine supporters. Most early Paine studies were, therefore, inaccurate due to large factual gaps and biased personal opinions. Even good studies of Paine ([1-4]) have suffered from a lack of an exhaustive factual reservoir to draw from. The biographies of Thomas Paine have been hampered by a lack of knowledge of Paine's early life and writings, and that vacuum has been filled with speculations. The difficulty in studying Paine's life and works is compounded by the fact that Paine wrote anonymously until 1791 when Rights of Man appeared. Thus, there exist a number of writings that may have been created by Paine but have never been attributed to him. There are also works which may have been erroneously attributed to Paine or whose authorship is disputed. While Paine's enormous contribution can stand on his major works alone (Common Sense, American Crisis, Rights of Man, Age of Reason, Agrarian Justice), to fully appreciate Paine's significance and impact on literature, philosophy, history, and politics, a clarification of his authorships is essential.

In 2011, the Institute for Thomas Paine Studies - a collaboration between Iona College, New Rochelle, NY and the Thomas Paine National Historical Association - began a multi-directional text analysis project whose main goal is to develop a solid scientific methodology for authorship attribution, and use it to verify the authorship of a number of documents that may have been written by Thomas Paine. The developed methodology is based on rigorous scientific principles, and takes advantage of modern computer technologies and techniques.

While the results generated by the authorship attribution software may be indicative of a strong possibility that a particular paper may or may not be attributed to Thomas Paine, they can never be absolutely conclusive. Thus, once a trend is uncovered by the authorship attribution software, a further verification and cross-reference of the facts is carried out by experts in Thomas Paine studies, American history, and 18th-19th century literature. Among the issues considered are the conformity of the work to the ideological content of the author's other writings and a match with the historical circumstances and personal idiosyncrasies of the author.

This paper focuses primarily on the automatic authorship attribution aspects of the work while highlighting the interaction between the software-based and the human-expert based aspects of the project.

II. Author Attribution

Authorship attribution is the task of identifying the author of an anonymous text or a text whose authorship is in doubt [5]. While many text mining applications analyze the content of a document as an important indicator for classification, authorship attribution usually focuses on the

DMIN'14: LATE BREAKING PAPER

Smiljana Petrovic is with Iona College, 715 North Av. New Rochelle, NY, 10801 (email: spetrovic@iona.edu, phone: 914-740-4620)
Gary Berton is the Coordinator of the Institute for Thomas Paine Studies at Iona College, New Rochelle, NY, USA, and the historian for the Thomas Paine National Historical Association, 715 North Av. New Rochelle, NY, 10801 (email: gberton@iona.edu, phone: 914-633-2648)
Robert Schiaffino is with Iona College, 715 North Av. New Rochelle, NY, 10801 (email: rschiaffino@iona.edu, phone: 914-633-2338)
*Lubomir Ivanov (**Contact Author**) is with Iona College, 715 North Av. New Rochelle, NY, 10801 (email: livanov@iona.edu, phone: 914-633-2342)*

style of the document rather than its contents; typically, all candidate authors write about similar topics and use similar, topic-specific words and phrases. However, stylistic features are often used unconsciously and consistently and, if correctly identified, may correctly reveal identity of the author.

We approach authorship attribution as a classification task: here, to classify a document means to assign it to the class of documents written by the same author. One way to perform classification is through supervised machine learning: Special algorithms use documents of known authorship (training examples) to train the system to recognize each author's writing style. Once training has been completed, the created model can be used to attempt to identify the creator of a document of disputed authorship.

A. Lexical Features

Among the most common "off-the-shelf" lexical features are function words, N-grams of characters and words, and sentence lengths [6].

Function words are the most common words (articles, prepositions, pronouns, etc.) in the English language. Since function words are topic-independent, they are usually excluded from the feature set of a topic-based text classification. However, since function words are often used in a subconscious manner, they well reflect the author's style and are among the best features for authorship attribution. In this work, we used function words as defined by Mosteller-Wallace in their Federalist papers study [7].

Word N-grams and character N-grams are also standard features used in text analysis. Word 2-grams consider sequences of 2 words from a given text. For example, word-2-grams of the text "Author Attribution of Paine and his Contemporaries" are "Author Attribution", "Attribution of", "of Paine", etc. Similarly, character-2-grams consider sequence of 2 characters from a given sequence of characters. For example, character-2-grams associated with the text "Author Attribution" are "au", "ut", "th", etc. N-grams features are simple, language independent and often very effective in text mining applications.

Our approach is to extract the fifty most frequent words from each document. Their union is a pool from which the fifty most frequent words are used to create the vector of features. The normalized vectors of frequencies of those words represent our training examples. For example, the vector of the most frequent functional words in one of our experiments was (to, but, for, no, by, every, has, been, who, of, were, are, more, his, would, any, on, had, be, such, so, or, and, shall, not, that, than, will, this, can, have, one, from, was, if, all, is, with, may, it, a, at, as, the, in, should, which, an, their, our). The normalized vector of frequencies of those words in Paine's "Forester Letters" was (0.07546, 0.01189, …, 0.01276). That vector is labeled as "Paine" and considered one training instance. Vectors of all documents of known authorship represent training data for one experiment.

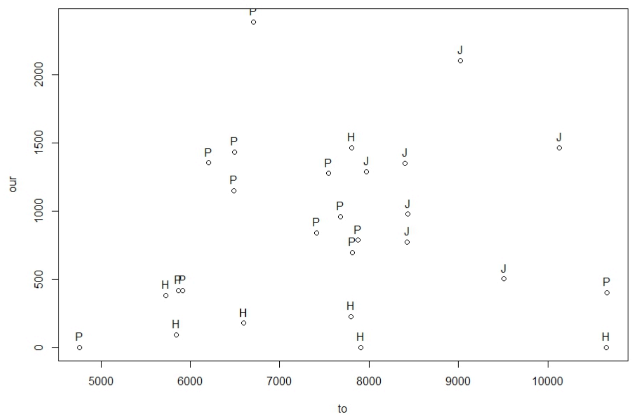

Figure1: Scatter plot of relative frequencies of two functional words, "to" and "our", in the 28 documents written by Paine (P), Jefferson (J) and Hopkinson (H)

Consider, for example a set of twenty eight documents written by Hopkinson, Jefferson and Paine in the two-dimensional space of relative frequencies of the words "to" and "our" (Figure 1). We can see that Jefferson relatively often uses "to", while Hamilton tends to use the word "our" less frequently that the other two authors. Ideally, all documents of the same author would have similar relative frequencies and would be clustered together: they would have small inter-cluster distance and large intra-cluster distance. Our example, however, illustrates that documents are often not clearly separated, and that documents by the same author may have very different feature values.

To attribute a document of an unknown or disputed authorship, a vector of its relative frequencies is created and submitted to the software. The machine learning algorithm attempts to extrapolate from its stored knowledge, and attribute the document to the author with the closest existing vector. In our example from Figure1 an unattributed document represented by vector (7000, 1200) would be probably attributed to Paine and one with vector (9000, 1000) to Jefferson by most methods. However, the classification of a document with vector (8200, 500) would be uncertain and probably different from method to method. Vector proximity is based on specific criteria characteristic of each leaning method. Thus, different learning methods may produce different attributions.

B. Learning methods

Linear Support Vector Machines (LSVM) method seeks a hyperplane in the n-dimensional input space which best separates points corresponding to different candidate authors. The best separator is the hyperplane that maximizes the distance to the closest training data points of different authors. To attribute a disputed document, we evaluate on which side of the hyperplane the point corresponding to that document lays.

Table 1: Features used in our analysis and their descriptions

Style marker	Description
MW Function Words	Considers function words as defined by Mosteller-Wallace in their Federalist papers study.
Word *n*-grams	Considers sequence of *n* items from a given sequence of words (we used the values 2, 3 and 4 for n)
Character *n*-grams	Considers sequence of *n* characters from a given sequence of characters (we used the values 2, 3 and 4 for n)
Part of Speech	Marking up a word in a text as corresponding to a particular part of speech; identification of words as nouns, verbs, adjectives, adverbs etc. Uses the Maxent Tagger developed by the Stanford NLP Group [9]
Vowel-initial Words	Considers words beginning with vowels
Sentence Length	The number of words in a sentence
Special Words	Special words selected from the documents that are frequent or have an atypical spelling (e.g. hath, juster, willful)
Prepositions	Most common prepositions
Suffixes	The last 3 letter of every word
First Word in Sentence	First Word in Sentence

Centroid Nearest-Neighbor approaches represent each author by its centroid vector - a vector whose coordinates are averages of coordinates of all training instances. An unknown document is associated with the author with the nearest centroid. Distance can be measured using different metrics. In our work, we used Euclidian distance (L2 metric) and cosine distance (normalized scalar product distance).

In order to check the accuracy of the model, the given documents are usually divided into training and testing sets. The training set is used to build the model, and then the model is tested on the remaining documents. In our work we adopted a "leave-one-out" validation: *n-1* of the available *n* documents are used for training, and validation is carried out using the remaining one document. The procedure is repeated n times, so every document is at some point used for validation. The percentage of correctly classified documents constitutes the "leave-one-out" accuracy of the method.

To further improve performance, we used a weighted sum of supports of different methods for different authors. Each method independently makes a choice (supports one author). We associate with each method a weight proportional to its leave-one-out accuracy. The weighted sum method selects the author with the largest weighted sum of all supports the author received from different methods. In our experiments, the weighted sum usually outperforms any individual method.

Table 2: Learning methods used in this study

Learning Method	Description
Support Vector Machine with Linear Kernel	Generates a linear separator to divide the feature space into regions, each corresponding to a specific author
Centroid with Histogram Distance	Nearest-neighbor approach using Euclidian Distance (L2 metric)
Centroid with Cosine Distance	Nearest-neighbor approach using normalized dot product distance

III. Experimental Design and Results

There are two major components that determine accuracy of learning - the set of lexical features considered and the choice of a machine learning (classification) algorithm. We consider sixteen lexical features (Table 1). Some of them are common, "off the shelf" lexical features, widely used in authorship attribution literature, such as function words, N-grams of characters and words, and sentence lengths. We also developed some domain-specific features based on our general knowledge of the documents: set of special words (e.g. "hath", "juster", "willful") and prepositions. For training we use a corpus of sixty-nine documents of ten authors [see Appendix A]. Results are obtained using the JGAAP (the Java Graphical Authorship Attribution Program), open source software [8] and programs written by authors of this paper. We extracted sixteen standard features from the documents and paired each with each of three learning approaches (Table 2). We considered fifty most common values for each feature.

The next table provides precision and recall data for each author. The *precision* of an author is the fraction of documents attributed to him that are indeed his work. The *recall* of an author is the fraction of his documents that were attributed to him. If a classification method attributes all training documents to one author, the recall will be 100%, as the software correctly classified all documents of that author, but precision will be low, as many documents were incorrectly attributed to that author. In the first experiment, we consider sixty-nine documents of ten candidate authors (Table 3). The leave-one-out validation correctly classified 90% documents. Note that if authors were assigned by random guessing, expected accuracy would be 10%. Seven authors (Adams, Benezet, Jefferson, Paine, Price, Priestley and Rush) had all their documents correctly identified (100% recall). Five authors (Adams, Benezet, Hopkinson, Price and Priestley) had 100% precision, indicating that they were associated only with their own documents. Four of them (Adams, Benezet, Price and Priestley) were associated with all of their documents, and with no other documents.

Table 3: Recall and precision of each author when learning was on sixty-nine documents and ten candidate authors. Accuracy of leave-one-out cross validation is 90%.

	Recall	***Precision***
Adams	100	100
Benezet	100	100
Franklin	89	80
Hopkinson	50	100
Jefferson	100	88
Paine	100	80
Price	100	100
Priestley	100	100
Rush	100	86
Witherspoon	71	83

In the next two sections we present experimental results using the selected learning methods and the weighted sum method when different candidate authors were considered for each of two unattributed/disputed-attribution documents. We start with 10 authors in the first experiment, and then select authors based on supports they received in previous experiments. We removed from consideration authors that received less than 10% support and repeated experiment with the narrowed set of candidate authors. The comprehensive list of all documents appears in Appendix A. For the four selected methods we report leave-one-out accuracy and their attribution choice. We choose generally well performing lexical features combined with linear support vector machine based learning. We report the leave-one-out accuracy and choice made by the weighted sum approach. We also provide break out of percentages of supports that each candidate author received from the weighted sum approach. For support we consider only methods that were correct on at least half of the documents in the leave-one-out validation. Ideally, all methods would choose the same author, giving him the 100% support.

A. The Dream Interrupted

Experimental Results

The performance of four experiments with different candidate authors and distribution of support to each author by weighted sum is shown in Table 4. All experiments selected Hokinson as the author of "The Dream Interrupted". The most successful four methods were character 3-grams, functional words, character 4-grams and words 2-grams in combination with LSVM, with accuracies of 87%, 77%, 74% and 72% respectively. Hopkinson was also voted as a probable author by the weighted sum method with support of 43%. In the second experiment, we eliminated from consideration authors that received less than 10% support, and repeated the experiment with only Adams, Hopkinson, Paine and Price. We, then, ran another experiment with Hopkinson, Paine and Price, and finally considered Hopkinson and Paine as the only candidates (Table 5). In all experiments, Hopkinson was selected as the author. Individual methods voted for Hopkinson vs. Paine with 65% support.

Table 4 Accuracy and choice made by four lexical features and weighted sum in experiments with different candidate authors on The Dream Interrupted

Method	***Accuracy***	***Choice***
All 10 authors (69 documents)		
Function words	77%	Hopkinson
Word 2-grams	71%	Adams
Character 3-grams	87%	Hopkinson
Character 4-grams	74%	Hopkinson
Weighted sum	**90%**	**Hopkinson**
Adams, Hopkinson, Paine, Price (32 documents)		
Function words	72%	Hopkinson
Word 2-grams	84%	Adams
Character 3-grams	84%	Hopkinson
Character 4-grams	81%	Hopkinson
Weighted sum	**91%**	**Hopkinson**
Hopkinson, Paine , Price (23 documents)		
Function words	83%	Hopkinson
Word 2-grams	87%	Hopkinson
Character 3-grams	83%	Hopkinson
Character 4-grams	78%	Hopkinson
Weighted sum	**87%**	**Hopkinson**
Hopkinson, Paine (16 documents)		
Function words	75%	Hopkinson
Word 2-grams	75%	Hopkinson
Character 3-grams	81%	Hopkinson
Character 4-grams	75%	Hopkinson
Weighted sum	**81%**	**Hopkinson**

Notice that Hopkinson has very low recall in all of our experiments. We believe that it is due to his experimental writing style that is hard to capture. As the number of training documents decreases, the misclassified Hopkinson's documents account for higher percentage of misclassifications, hence the overall accuracy decreases. However, the percentage of support to Hopkinson increases. More importantly, in all experiments, Hopkinson has 100% precession: every document attributed to him was truly his work. This fact additionally supports the attribution of "The Dream Interrupted" to Hopkinson.

Table 5: Support, recall and precision of each author in experiments with different candidate authors on The Dream Interrupted

	Support	***Recall***	***Precision***
All 10 authors (69 documents)			
Adams	16	100	100
Benezet	2	100	100
Franklin	3	89	80
Hopkinson	**43**	**50**	100
Jefferson	3	100	88
Paine	22	100	80
Price	10	100	100
Priestley	0	100	100
Rush	0	100	86
Witherspoon	0	71	83
Adams, Hopkinson, Paine, Price (32 documents)			
Adams	13	100	100
Hopkinson	**42**	62	100
Paine	25	100	73
Price	20	100	100
Hopkinson, Paine, Price (23 documents)			
Hopkinson	**54**	62	100
Paine	28	100	73
Price	18	100	100
Hopkinson, Paine (16 documents)			
Hopkinson	**65**	62	100
Paine	35	100	73

Human Expert Cross-Verification

The article “The Dream Interrupted” appeared in the Pennsylvania Magazine in May, 1775. Philip Foner described the article as “an interesting example of Paine's ability to use different literary techniques to bring home a vital political message” [2]. While it is true that Paine used different styles to get his message across, this is not one of them.

Our Author Authentication method suggests that this article was written by Francis Hopkinson. Hopkinson was a frequent contributor to the Pennsylvania Magazine, writing under the pen names of B, or the Old Bachelor (also used by others), but mostly unsigned. Hopkinson resided in Bordentown, NJ, where Paine eventually purchased property and spent a great deal of time. The friendship between the two men is well documented, and Paine’s ties to Bordentown probably stem from this association at the Pennsylvania Magazine. Hopkinson was a strong advocate of independence, and the politics of “The Dream Interrupted” fit his views as well as Paine’s.

The context of “The Dream Interrupted” should have also raised some questions regarding Paine’s authorship. The fact that it took place in Bucks County (signed “Bucks County”), which is across the river from Bordentown in Pennsylvania, though not definitive, would not fit in with Paine’s movements in the first five months in America. The reference to a “fatiguing journey from Virginia” in the first sentence of the article should also have raised questions, since there is no indication that Paine could have made time to venture such a trip. Bucks County would be the last leg of a trip from Virginia to Bordentown. It would not be part of an itinerary from Virginia to Philadelphia where Paine resided at that time. Paine’s physical involvement in Bordentown did not begin until 1778.

Hopkinson also used the device of dreams in other Pennsylvania Magazine articles, such as “The Revery” and “The Extraordinary Dream”. Hopkinson’s more classical style fits this employment of dreams, as does his sweeping grandiose metaphors. In this period of his life, Hopkinson had set aside his musical compositions to focus on experimental writing styles, hence his frequent contributions to the magazine.

B. The Magazine in America

Experimental Results

Our experiments (Tables 6 and 7) strongly point to Paine as the author of “The Magazine in America”. Among the ten candidates, only Paine and Price received a support higher than 10%, with Paine’s support of 77% and Price’s of 18%. In an experiment with only those two candidates, Paine received overwhelming support of 93%. Among highly weighted methods, the part of speech and character-2-grams supported Price, all others favor Paine. In all other experiments in which Paine was included, the weighted sum algorithm highly supported him.

Human Expert Cross-Verification

Scholars have questioned the authorship of this article which appeared in the Pennsylvania Magazine in January 1775 because it was written soon after Paine arrived in America with an acute, debilitating illness. This was the premiere issue. However, our analysis supports Paine’s authorship. Paine was present in Philadelphia, and looking for employment, and his subsequent involvement in the magazine all favor his authorship. The article demonstrates Paine’s hands-on role at the magazine, his early selection of pieces to publish, and the editorial position he assumed. This reinforces those scholars who give weight to articles appearing in the magazine as reflective of Paine’s political and philosophical leanings.

Table 6: Accuracy and choice made by four lexical features and weighted sum in experiments with different candidate authors on The Magazine in America

Method	***Accuracy***	***Choice***
All 10 authors (69 documents)		
Function words	77%	Paine
Word 2-grams	71%	Paine
Character 3-grams	87%	Paine
Character 4-grams	74%	Paine
Weighted sum	**90%**	**Paine**
Paine, Price (15 documents)		
Function words	93%	Paine
Word 2-grams	93%	Paine
Character 3-grams	93%	Paine
Character 4-grams	93%	Paine
Weighted sum	**100%**	**Paine**

Table 7: Support, recall and precision of each author in experiments with different candidate authors on The Magazine in America

	Support	*Recall*	*Precision*
All 10 authors (69 documents)			
Adams	0	100	100
Benezet	0	100	100
Franklin	0	89	80
Hopkinson	0	50	100
Jefferson	0	100	88
Paine	**77**	100	80
Price	18	100	100
Priestley	0	100	100
Rush	5	100	86
Witherspoon	0	71	83
Paine, Price (15 documents)			
Paine	**93**	100	100
Price	7	100	100

IV. Future Work

The team is currently working to increase the size of our training corpus by adding additional documents written by the authors that we have been studying, and by identifying other authors from this period and including their works. Under our current classification methods, an unknown document will be always attributed to one of the candidate authors whose documents were used during the training. The possibility that a document is authored by someone not among the selected set of authors is not currently supported in the software model. Hence, ensuring the inclusion of the names and representative documents of additional relevant candidate authors is crucial.

Developing new features relevant for authorship attribution is another focus of our work. While we have already considered several well-known lexical features and learning methods, we intend to enhance our current methodology by including new methods for text analysis (e.g. artificial neural networks) and new lexical features such as the use of alliterations.

As the size of the corpus expands and the number of features grows, the computational complexity is expected to increase drastically. Therefore, we are investigating the use of advanced computing techniques so that some of the computations can be carried out in parallel rather than sequentially.

Additional challenges we face in our future work include the reliability of training documents and the influence of incorrectly attributed training documents, which could have serious impact on accuracy of the model and its predictions.

Finally, our future work will consider relationships among authors. We would like to be able to consider not only direct collaborations on some documents, but also evidence of influences of one author on another, in ideas as well as in style. An additional challenge is detecting the influence that editors or even typesetters may have had on final versions of the published documents. We would also like to attempt detecting common characteristics of authors that intentionally experimented with different styles (e.g. Hopkinson).

These challenges necessitate a close collaboration with historians, political scientists, and 18th/19th century literary scholars. The results of our analysis should be considered directions for further study by historians. Only through a close collaboration can the true nature of the life and work of Thomas Paine and his global impact on posterity be truly revealed.

Appendix A

Table 8: A corpus of work of ten authors we used for training. Some of large publications were broken into multiple documents.

Author	***Work***
Adams	Defense
	John Adams Thoughts
Benezet	Guinea (Sans Quotes)
	Some Observations On The Situation (Sans Quotes)
	Caution To Great Britain And Her Colonies (Sans Quotes)
Franklin	Correspondence and essays from 1768 to 1776
Hopkinson	Improved Plan Of Education (Sans Quotes)
	Consolation For The Old Bachelor (Sans Quotes)
	The Ambiguity Of The English Language
	A Revery (Sans Quotes)
	A Prophecy
	A Pretty Story (Sans Quotes)
	Extraordinary Dream
	Translation Of The Letter
Jefferson	Correspondence 76-77

	Correspondence 78-79
	Drafts of documents & Correspondence 78
	On The Instructions Given To The First Delegation Of Virginia To Congress
	Correspondence Post 1780
	Correspondence Pre 1776
	Summary View
Paine	Common Sense
	Crisis Papers
	Forester Letters
	Miscellaneous Articles In 75 And 76
	Of Monarchy And Hereditary Succession
Price	Observations Civil Liberty (Sans Quotes)
	Review Of The Principal Questions In Morals
	Britains Happiness
	Discourse Of Love Country
	Evidence Of Future
	Fast Sermon
	Observations on the Importance of the American Revolution (Sans Quotes)
Priestley	An Essay On The First Principles Of Government (Sans Quotes)
Rush	A Plan For Establishing Public Schools In Pennsylvania
	An Account Of The Life And Death Of Edward Drinker
	An Address To The Ministers Of The Gospel Of Every Denomination In The United States
	Paradise Of The Negro
	Thoughts Upon Female Education
	Thoughts Upon The Amusements And Punishments Which Are Proper For Schools
Witherspoon	Aristides
	On Conducting The American Controversy
	On The Affairs Of The United States
	On The Convention With General Burgoyne
	On The Proposed Market In General Washington
	Reflections
	Thoughts On American Liberty

References

[1] Claeys, Gregory, *Thomas Paine: Social and Political Thought. Boston: Unwin Hyman*, 1989.

[2] Foner, Philip, *The Complete Writings of Thomas Paine,* 2 vols. New York: Citadel Press, 1969.

[3] Williamson, Audrey, *Thomas Paine*, New York: St. Martin's Press, 1973.

[4] Fruchtman, Jr., Jack. Fruchtman, Jr., Jack, *Thomas Paine: Apostle of Freedom, Four Walls Eight Windows*, New York, 1994

[5] Love, H., *Attributing Authorship: An Introduction*, Cambridge, Cambridge University Press, 2002.

[6] Stamatatos, Efstathios, *A survey of modern authorship attribution methods,* Journal of the American Society for Information Science and Technology 60, no. 3 (2009): 538–556.

[7] Mosteller, Frederick, and David L. Wallace, *Inference and disputed authorship: The Federalist*, Reading, MA: Addison-Wesley, 1964.

[8] Juola, Patrick, *Authorship Attribution*, Foundations and Trends in Information Retrieval 1, no. 3 (2006): 233-334.

[9] Toutanova, Kristina, Dan Klein, and Christopher Manning. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*. HLT-NAACL, 2003: 252-259.