Linguistic Approach for Authentic Authorship

Rishi R. Singh[1], Deepika Koundal[2] and Rajeev Tiwari[3]

[1] *School of Computer Science University of Petroleum & Energy Studies, UPES, Dehradun, India*
[2,3] *Department of Virtualization, School of Computer Science University of Petroleum & Energy Studies, UPES, Dehradun, India*

Abstract

This work is an analysis of the linguistic approach of author's and by this, the work identifies the correct author for the debated federalist papers with a special focus on Federalist 64 paper. This work was assumed to be written by Hamilton but is debated by others with the assumption that John Jay has written them. Three different methods are utilized to analyze the papers to test the correct authorship of the papers. The paper uses a set of thirty commonly used words to apply Chi-squared which will help identify similarities of differences between the authors. After applying and going through the three techniques it can be concluded that the Burrows' Delta method gives the most accurate prediction of the Author for a given piece of text..

Keywords

Linguistic methods, Fake Authorship Attribution, NLTK, John Burrows' Delta, Kilgariff's Chi-Squared, Curves of Composition.

1. Introduction

The presented work does an analysis of author stylometry [1, 2] by using three different algorithms namely Mendenhall's Method of Composition [3], Chi-Squared by Kilgariff's [4], and Delta calculation by John Burrows' [5]. Since the 1960s, statisticians, and literary experts have tried to analyze texts to resolve questions and disputes about authorship, and this approach is called stylometry.

With this presented work, "Comparative analysis on Linguistic Approach, this paper aims to do a comparison of Linguistic Approach with the help of quantitative study of literary styles, of different authors, with reading methods using computational techniques. The focus will be on the Federalist Papers part of Project Gutenberg. It is since authors write in consistent ways which can be recognized and are unique in a lot of ways. For example, Different author or person use their vocabulary, which is unique in a lot of ways. Since different people use unique vocabulary, it is said that two people will not use punctuations in the same way. Eustachio Stamatas makes the same point in the survey of historical and current stylometric methods, that authors use function words unconsciously without correlation to the topic. A lot of research has in which differences have been studied between the ways in which all genders write [7] or are written about. The Federalist Papers which has a few debated papers have not been analyzed for authorship comparison using Natural Language processing [8]. The authorship of Federalist 64 has been debated; the work will try to find the correct author for it by using three different methods. Identifying an author based on the number of words being used per sentence is sometimes possible. This work will try to identify it, the test case will be the Federalist Papers taken from Project Gutenberg.

The contributions of this work are as:

1. The work of three authors has been analyzed to verify the authenticity of Authorship.

International Conference on Emerging Technologies: AI, IoT, and CPS for Science & Technology Applications, September 06–07, 2021, NITTTR Chandigarh, India

EMAIL: mr.rishisingh@gmail.com (A. 1); dkoundal@ddn.upes.ac.in (A. 2); rajeev.tiwari@ddn.upes.ac.in (A. 3)
ORCID: Not Available (A. 1); Not Available (A. 2); Not Available (A. 3)

2. We have used Curves of Composition – Mendenhall, Chi-squared measure of distance – Adam Kilgariff's and Burrows' Delta to analyze Federalist 64, for comparison.
3. The metric curves of composition give a rough idea of the correct author, chi-squared metric builds on it to improve the accuracy of the results and finally Z score metric gives the most accurate result.
4. We finally conclude based on the analysis that the correct author for Federalist 64 was John Jay.

The rest of the paper is structured as below: Section 2 covers the dataset being used along with the methods to run the analysis. Section 3 talks about the results and the conclusion based on that. Finally, section 4 concludes the conclusion.

2. Material and Methods

2.1. Dataset

This work uses the Federalist Papers [9] which is an archive containing 85 documents, it also has the original Project Gutenberg ebook [10] from which these 85 documents have been taken. When unzipped, a folder named data is created, which will be used as the working directory. These 85 documents were published under a single pen name "Publius" [11] — but were written by three American prominent politicians: Alexander Hamilton, James Madison, and John Jay. Since they were initially published under a common pen name there are a few which are debated for authorship.

The following tools have been used for research and analysis: Python 3.x, nltk - Natural Language Toolkit, and matplotlib.

2.2. Methods

The three different methods have been used on Federalist Papers to show three different stylometric approaches and to identify the correct author of Federalist 64 Paper [12]. Curves of Composition – Mendenhall, Chi-Squared – Kilgariff and Delta Method - John Burrow.
The presented work has been split into six different categories, as below:-

- The fifty-one papers are said to be written by Hamilton.
- The fourteen papers are said to be written by Madison.
- The set of 4 papers from the five which are thought to be written by Jay.
- The three papers assumed to be written together by Madison and Hamilton
- Twelve papers are debated between Hamilton and Madison.
- A separate category for Federalist 64

The analysis will be done based on the below methods:

- Curves of Composition - Mendenhall
- The chi-squared measure of distance - Adam Kilgariff's
- Burrows' Delta to analyze Federalist 64.The fifty-one papers are said to be written by Hamilton.

2.2.1. First Linguistic Test: Curves of Composition

As per Mendenhall, an author's linguistic signature can be figured out if one counts the length of different words used by the author and how often the length of words is different. Length of words can count in a number of 1,000-word or 5,000-word blocks for a given book and plot a graph of word lengths, the plotted curve will almost be the same for a graph plotted from any part of the book [13]. Mendenhall suggested that if one counts a large sample of words taken from various works of a writer, the characteristic curve for the usage of word length will be so accurate that it would remain the same for all his writings.

2.2.2. Second Linguistic Test: Kilgariff's Chi-Squared Distance

Adam Kilgarriff [14] wrote a paper in 2001 which recommended using chi-squared statistics for determining the correct author. The method implements statistics to calculate the distance between vocabularies used in a couple of text groups. This is how it will apply the statistic for authorship attribution:

- The focus will be to use papers of a couple of Authors to be analyzed and will merge them to create a single large collection.
- This step will involve sum up the tokens for each word which it will find in the larger collection.
- The most used and common 'n' number of words will be selected.
- Now divide the number of tokens that were found in a larger collection, based on the comparative size of contributions from both authors to the larger collection.
- Work on the chi-squared distance by using the below equation. C is the count of tokens for a feature and E is the count expected for it.

$$(O-E)^2/E \tag{1}$$

Considering the two collections similarly depending on their chi-squared value, the minor the value will be the more confirmation will be there for collections to be similar. The chi-squared values for the difference of Madison and clashed collection will be calculated, and another one for the difference between the Hamilton and Clashed collection; this will help to identify who Madison or Hamilton is the most likely author for the Debated set. Jay is not part of this as from the first test it can deduce that Jay cannot be the author.

2.2.3. Third Linguistic Test: John Burrows' Delta

John Burrows' Delta statistic [15], is comparatively more complex as compared to the first two but is a very prominent method used for stylometric analysis. John Burrows' Delta measures the distance between a text whose authorship one wants to find when compared to another collection, this is similar to Kilgariff's chi-squared. With the Delta value, the divergence between the unknown/known texts can be measured and when they are all put together. Equal weight is assigned to each feature that is used in the measurement. Burrows' algorithm is as follows:

- Gather a collection of texts, written by a random author, for example, x.
- Features will be the n most used words.
- Calculate the percentage of the total count of words for the share for each of the x authors whose sub-collection has the representation of these n features.
- Calculate mean and standard deviation for all the x values and use those values for this feature over the whole collection.
- A z score will be calculated for the n features and x sub-collection, which will describe how much further away from the collection is the usage of the feature in the sub-collection. Then reduce mean of means for the feature from the frequency of each feature from the sub-collection and divide the outcome by the standard deviation of the feature [16].
- The z-scores for all features will have to be calculated, for a text for which authorship is needed.
- The last step is to calculate the delta score which will compare the anonymous paper with each author's sub-collection. For this, the average of complete values of the difference between the z-scores for each feature between the unacknowledged paper and the author's sub-collection. This will assign equal weight to each feature.
- The identity of the author will be identified based on the smallest delta score between the test case and part of the collection.

2.2.4. Test Case: Federalist 64

The Federalist 64 will be considered for the final test. This was claimed to be written by Alexander Hamilton, however, a draft was also recovered in Jay's papers, so there is the possibility of it being authored by John Jay. As Burrows' Delta Method is based on a random group of authors, to confirm Federalist 64's linguistic impression will be matched with the following: Hamilton's work, Madison's writings, Jay's prior works, shared, and texts which have a clash for Hamilton and Madison. Using the Delta method will be beneficial in this case and should be able to identify the actual author.

2.3. Methodology

The sub-collection will be combined into one collection for Delta. This will help in calculating the standard to assist our work with it. Several words will be selected to be used as features. For calculating Kilgariff's chi-squared 500 words were used but for this test, a set of 30 words will be used, as features.

- To calculate features for each sub-collection, the frequencies of each feature will have to be investigated for each author's sub-collection and take as a part of the total tokens in the sub-collection.
- A mean of means along with standard deviation will be calculated covering all the features, this will be used to calculate average and standard deviations of features.
- For Calculating z-scores the observed feature frequencies will be transformed in the five authors' sub collection [17], this will describe the distance from the collection norm these observations were taken.
- Delta calculation will be done by the Burrows [18,20,21] formula to get an exclusive score which will compare Federalist 64 with the three authors and clashed and shared set. The minute the Delta score for the author the more it will match to the author.

$$\Delta \mathrm{B}=\sum_{i=1}^{n}(z(D1)_{i}-z(D1)_{i} \tag{2}$$

ΔB is the Burrows Delta score, n is the number of a number of features, z is the frequency of word distance and D is the distance.

3. Results and Discussion

The results will be analyzed based on the three different methods that were used to identify authors. The author identification for the Debated paper 64 will be based on the most accurate John Burrows Delta Method. Let's discuss and analyze the results for all three methods one by one.

3.1. Results Based on Mendenhall's Method

The graphs for Authors Hamilton as shown in Fig 1, Madison as can be seen in Fig 2, and Jay as depicted in Fig 3, depict the length of words and their usage. For example, Jay uses 2 and 3 letter words the most, Hamilton word length steadily declines from 2 letters to words to 15 letter words and Madison is somewhere in between Jay and Hamilton. The graphs depicted for using Mendenhall's Method for the clashed works, Fig 4 looks like a mix and match for works of Madison and Hamilton. There is not much help with the authorship attribution that is being analyzed as is displayed in Fig 5. The below graphs are plotted with the number of occurrences on the Y-axis and word length in the X-axis.

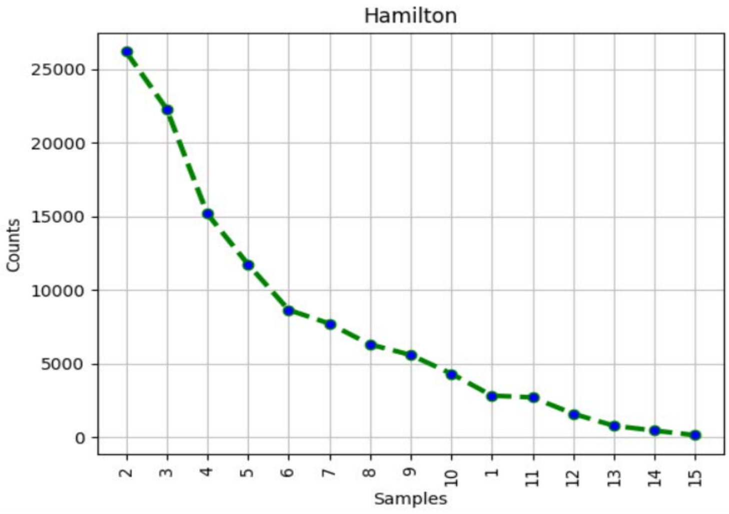

Figure 1: Hamilton’s graph

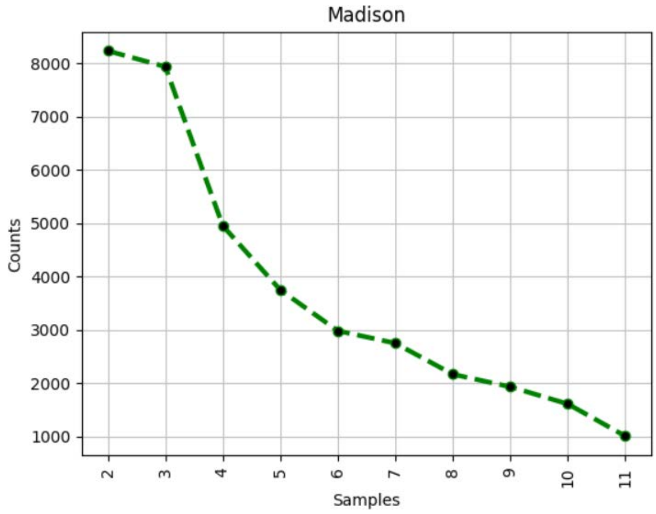

Figure 2: Madison’s graph

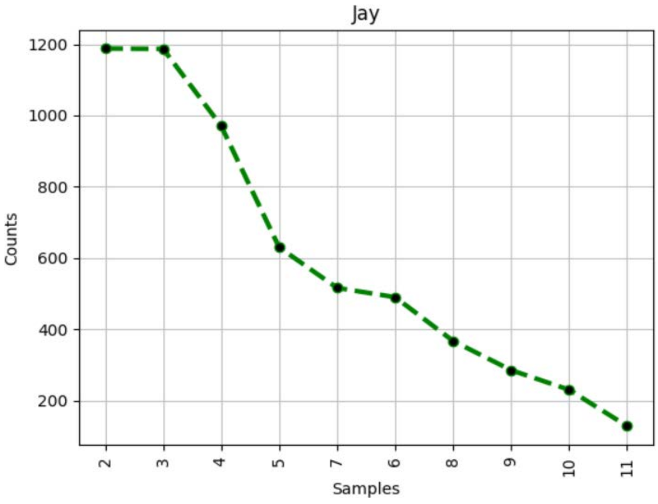

Figure 3: John Jay’s graph

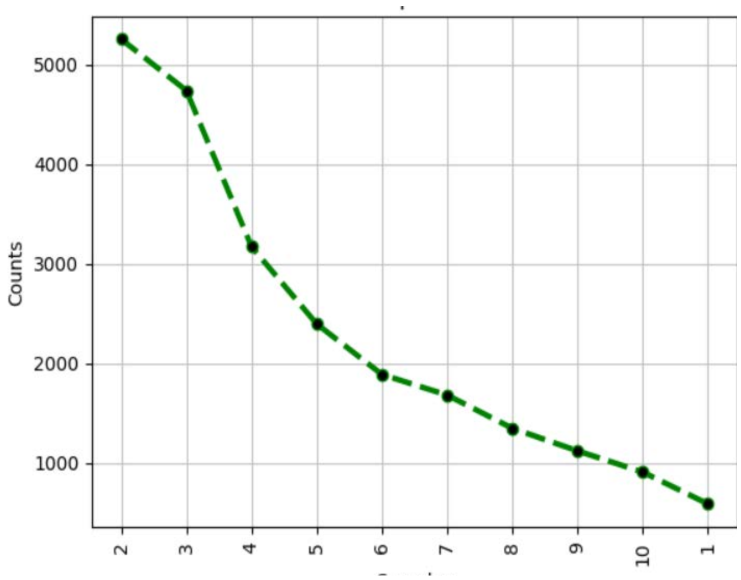

Figure 4: Debated graph

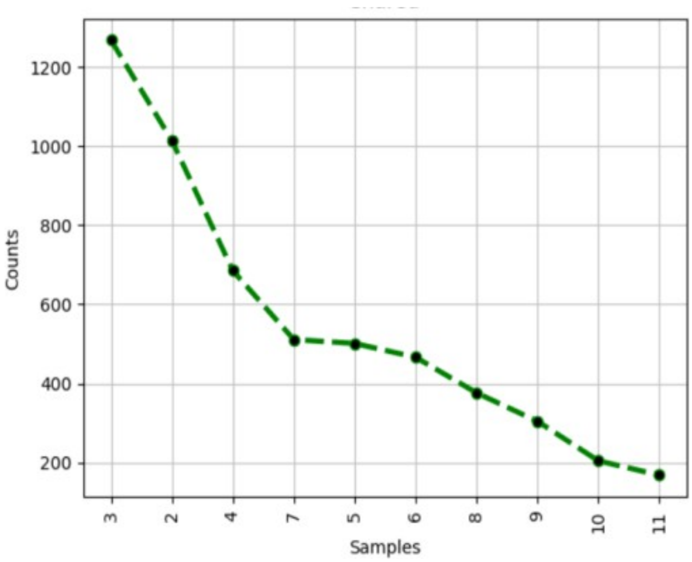

Figure 5: Shared graph

3.2. Results based on Kilgariff's Chi-Squared Method

The results based on Kilgariff's chi-squared method as presented in table 1, the difference in chi-squared values between debated papers and Hamilton's collection is comparably larger as compared to the difference between the Madison and Debated collection [19]. This confirms Madison was the author for the twelve papers in the debated collection rather than Hamilton and the results don't give any connection with John Jay for the clashed papers.

Table 1
Chi-squared Statistic results for authors Hamilton and Madison

Author Hamilton	3463.5516225
Author Madison	1907.5992915

3.3. Results based on Kilgariff's Chi-Squared Method

The results of features based on z-scores for the Federalist paper 64 are as presented in table 2, the Z scores have been calculated for thirty common syllables used by authors.

Table 2

Z scores for the common words used

Feature	Z Score	Feature	Z Score
the	-0.590727	for	-0.862375
of	-1.818085	would	-0.840871
to	1.097095	have	2.327511
And	1.054642	will	1.497289
In	0.759606	or	-0.238417
A	-0.795675	from	-0.502357
Be	1.027406	their	0.862943
that	1.959876	with	-0.040503
it	0.211485	are	7.796343
is	-0.879192	on	-0.038225
which	-2.055245	an	-0.714673
by	1.219933	they	5.366035
as	4.552016	states	-0.721953
this	-0.651326	government	-2.045708
not	0.843206	may	0.983682

Now based on the below Delta scores shown in table 3, it can be confidently said that John Jay was the author for Federalist 64 and not by Hamilton as has been falsely claimed. Thus, putting an end to all the false claims and controversies.

Table 3

Delta scores for the three authors, including Debated and Shared

Deltascore for author Hamilton	1.752279
Deltascore for author Madison	1.598792
Deltascore for author Jay	1.516661
Deltascore for Debated	1.536483
Deltascore for Shared	1.906091

4. Conclusion

Based on the three different methods used, it can be concluded that the Burrows' Delta method gives the most accurate prediction of the Author for a given piece of text. Mendenhall's method confirms that Jay certainly was not the author of the debated papers, his curve is the most distinct than the other two, lengths 6 and 7 are even inverted in his graph. Kilgariff's Chi-Squared method confirms Madison is the author for the twelve papers in the debated collection rather than Hamilton. The other two, Mendenhall's and Kilgariff's chi-squared method though good are not very accurate to predict who is responsible for a specific text. John Burrow's method looks like the most accurate way to identify the authorship.

5. Acknowledgment

I wish to express my deep gratitude to my guide Dr. Deepika Koundal, for the advice, encouragement, and constant support she has given me for my research work. It would have been

challenging to work without her support and valuable suggestions. Lastly, I would like to express my sincere gratitude to my parents for every support they have given to me.

6. References

[1] Linguistic Approaches to Interval Complex Neutrosophic Sets in Decision Making, in IEEE Access,vol. 7, pp. 38902-38917, 2019

[2] Krause S.S. (2021) Federalist Papers und Antifederalists. In: Campagna N., Hidalgo O., Krause S.S.(eds)Tocqueville-Handbuch.J.B. Metzler, Stuttgart. 2021;

[3] Vysotska, V., Lytvyn, V., Hrendus, M., Kubinska, S., Brodyak, O.: Method of textual informationauthorship analysis based on stylometry. In: 2018 IEEE 13th International Scientific and TechnicalConference on Computer Sciences and Information Technologies (CSIT), vol. 2, pp. 9–16. IEEE(2018)

[4] Boccia,F,Sarnacchiaro,P.Chi-squaredautomaticinteractiondetectoranalysisonachoiceexperiment:An evaluation of responsible initiatives on consumers' purchasing behavior. Corp Soc Resp Env Ma.2020; 27:1143– 1151

[5] JohnBurrows,"'Delta':aMeasureofStylisticDifferenceandaGuidetoLikelyAuthorship",LiteraryandLinguistic Computing,vol. 17, no. 3(2002), pp.267-287.

[6] EfstathiosStamatatos,"ASurveyofModernAuthorshipAttributionMethod,"JournaloftheAmerican Society for Information Science and Technology, vol. 60, no. 3 (December 2008), p. 538–56, citationonp.540, https://doi.org/10.1002/asi.21001.

[7] JanRybicki,"VivеLaDiffe´rence:Tracingthe(Authorial)Gender Signal by MultivariateAnalysis of Word Frequencies," Digital Schol- arship in the Humanities, vol. 31, no.4(December2016), pp. 746–61, https://doi.org/10.1093/llc/fqv023. Sean G. Weidman and James O'Sullivan, "TheLimitsofDistinctiveWords:Re-EvaluatingLitera-ture'sGenderMarkerDebate,"DigitalScholarshipinthe Humanities,2017,https://doi.org/10.1093/llc/fqx017.

[8] Ted Underwood, David Bamman, and Sabrina Lee, "The Transformation of Gender in English-LanguageFiction",CulturalAnalytics,Feb. 13,2018,DOI: 10.7910/DVN/TEGMGI.

[9] Heath, J. Benton. "From the Spirit of the Federalist Papers to the Endof Legitimacy: Reflections onGundyv. United States." Nw. UL Rev.114 (2019):1723.

[10] Rheingold, David M. "The Longest Words Using The Fewest Letters." Word Ways, vol. 53, no. 4,2020,p. NA. Gale AcademicOneFile, . Accessed 22Apr. 2021.

[11] Kincaid,John."Publius:TheJournalofFederalismfromitsFoundingtoHalf-CenturyMark."(2020):541-543.

[12] Rink, Jonah. "A Republic in its Own Time: The Re-Imagining of Republican Theory in the FederalistPapers."(2020).

[13] T. C. Mendenhall, "The Characteristic Curves of Composition", Science, vol. 9, no. 214 (Mar. 11,1887), pp.237-249.

[14] AdamKilgarriff,"ComparingCollection",InternationalJournalofCol-lectionLinguistics,vol.6,no.1 (2001),pp.97-133.

[15] JohnBurrows,"'Delta':aMeasureofStylisticDifferenceandaGuidetoLikelyAuthorship",LiteraryandLinguistic Computing,vol. 17, no. 3(2002), pp.267-287.

[16] Huang, W.-H.Control ChartsforJointMonitoringoftheLog- normal Mean and StandardDeviation.Symmetry2021,13, 549. https://doi.org/10.3390/sym13040549

[17] Chalmer,B.J.(2020).UnderstandingStatistics.UnitedStates:CRCPress.

[18] Plecha´c,P.(2020). On anUnknown AncestorofBurrows'DeltaMea- sure.arXiv preprintarXiv:2012.04796.

[19] Omar, A., Hamouda, W.I. (2020). The Effectiveness of Stemming inthe Stylometric

AuthorshipAttribution in Arabic. International Journal of Advanced Computer Science and Applications, 11(1),116-121.

[20] Sharma, Y., Bhargava, R., & Tadikonda, B. V. (2021). Named Entity Recognition for Code Mixed Social Media Sentences. *International Journal of Software Science and Computational Intelligence (IJSSCI)*, *13*(2), 23-36.

[21] Bouarara, H. A. (2021). Recurrent Neural Network (RNN) to Analyse Mental Behaviour in Social Media. *International Journal of Software Science and Computational Intelligence (IJSSCI)*, *13*(3), 1-11.