A stylometric analysis of two sentimental letters by Albert Einstein and Simón Bolívar

Raúl Isea[1,*]

[1]Fundación IDEA. Hoyo de la Puerta, Baruta, Venezuela

Abstract

On social networks often circulate emotional letters from various authors that appear after their death. The supposed letter Albert Einstein wrote to his daughter Lieserl and the letter Simon Bolivar wrote to his cousin Fanny are two examples. A Python-developed tool is used to do a stylometric study to determine whether these authors were. The language styles of Einstein and Bolivar were described using eight and six letters, respectively. The results show that they never wrote these letters.

Research Article

Open Access &

Peer-Reviewed Article

DOI: 10.14302/issn.2998-4122.jlr-23-4739

Corresponding author:
Raúl Isea, Fundación IDEA. Hoyo de la Puerta, Baruta, Venezuela.

Keywords:

Albert Einstein, Simón Bolívar, Stylometric, Linguistic footprint, Letter.

Received: September 04, 2023

Accepted: October 30, 2023

Published: November 29, 2023

Citation:

Raul Isea (2023) A stylometric analysis of two sentimental letters by Albert Einstein and Simón Bolívar. Journal of Language Research - 1(1):15-20. https://doi.org/10.14302/issn.2998-4122.jlr-23-4739

Introduction

Letters written by great people in history are circulating on social networks. An example is the letter that the German physicist Albert Einstein (1879-1955) wrote to his daughter Lieserl Einstein. In this letter, he shows that love is the only answer to the survival of our species. There is also a letter that Simon Bolívar (1783-1830) claims to have written to his cousin Fanny du Villars, written on her deathbed, telling her how important she was in his life.

In fact, Albert Einstein's letter to his daughter analyzes the concept of love as a universal force that science has not yet fully developed. This letter says that love is seen as light, gravity and life. It is the only energy in the universe that humans have not learned to control at will. Suggests that love is the most powerful force because it has no limits.

Simón Bolívar wrote a letter to his cousin Fanny du Villars, obtained from the archives of the National Academy of Medicine of Colombia (https://www.interacademies.org), to express his emotional affection for her. In the letter, he describes the surrounding nature and emphasizes the beauty of the Caribbean Sea, the magnificent mountains and the colorful sky, and tells her how important she was in his life. The strange thing is that the real name of the French woman was Louise Jeanne Nicole Arnalde "Fanny" Denis de Keredern de Trobriand (1775-1859).

On the other hand, all the papers of Albert Einstein are collected in various databases such as The Collect Papers of Albert Einstein (https://einsteinpapers.press.princeton.edu/), Einstein Papers Project (https://www.einstein.caltech.edu/), and so on. None of them have any proof of the existence of this letter to his daughter.

The same applies to the documents of Simón Bolívar available at Biblioteca Virtual Miguel de Cervantes (https://www.cervantesvirtual.com). There is also no information in these files about the existence of this emotional letter. For all this, these letters are analyzed according to the stylometric methodology as described below.

Stylometric

Stylometric has been summarized as a statistical methodology based on the frequency analysis of each author's own words (14, 6, 11). The term was coined by the Polish writer Wincenty (8) when he determined the chronology of Plato's dialogues (8), and was originally based on a chi-square analysis of word length-word and their frequency (9).

More recently, it has been used to determine the authorship of some unknown or questionable works, such as a comedy originally attributed to Miguel Bermúdez by the National Library of Spain, but found by this method to be the author of Lope de Vega. (3).

Another example was Ulysses, published in 1922 by the Irishman James Joyce, but stylometric studies conclude that it was written by five other people who were not mentioned in the work. (12).

Based on the above, the paper analyzes eight letters written by Albert Einstein and finds out if the letter to his daughter was really written by him. This procedure is repeated in a letter that Simón Bolívar allegedly wrote to his cousin, as described below.

Methodology

Daelemans (4) describes the methodology for performing a stylometric analysis and is summarized below. A corpus is created from the letters of the same author. The selected letters and the numbers were chosen at random (details in Table 1).

For each author, the word frequency is determined according to the Delta function, which is a linguistic measure capable of distinguishing the authorship of texts according to the definition proposed by (2).

Some articles questioned the use of the Delta function, but Burrows showed that it is an excellent technique for identifying the authorship (13). This function simply determines the frequency variation of the most frequent words in the text using z-scores (13). Remember that z-scores are a measure of relative frequency difference minus the word mean divided by the standard deviation (16).

After that, a matrix was calculated based on the distance obtained from the frequency of the words (15). This calculation uses one, two or more words, and this number is indicated as n-grams (15).

Based on this frequency, the distance Remember that there were two possible element types used to generate n-grams: words and characters. Character n-grams show how frequently certain letters, capital letters, punctuation marks, or numbers are used at the alphabetic level of a language, while the word n-grams and vocabulary in a document are connected. In addition to word frequency, these characteristics also include sentence length, word length distribution, richness of vocabulary, and lexical mistakes. These can be used as the initial tokenization step for any language [10].

For other part, it is usually calculated using the Manhattan, Euclidean, and so on (15). The result is usually visualized with a dendogram, a tree where similar distances are grouped based on a certain number of words into a single conglomerate or cluster, simply abbreviated as MRW (ie., Meaning Most

Words).

The paper considers that sentences written by the same person should be grouped into the same branch or node (5). All calculations were performed in the Python programming language.

Table 1. Details of the letters written in English by Albert Einstein, and the letters in Spanish by Simón Bolívar.

Abbreviation	Who wrote the letter	Date	To whom the letter is addressed
B362	Bolívar	1830, May 26	General Sucre
B363	Bolívar	1830, May 26	Juan de Dios Amador
B364	Bolívar	1830, May 31	Juan de Dios Amador
B365	Bolívar	1830, Jun 17	Pedro Medrano
B367	Bolívar	1830, Jul 31	Manuela Garaycoa de Calderón
B371	Bolívar	1830, Oct 17	Joaquín de Mier
Fanny	Bolívar	1807, Sep 14	Fanny du Villers
Freud	Einstein	1932, Jul 30	Sigmund Freud
God	Einstein	1954, Jan 3	Mr. Gutkind
Borns	Einstein	1924, Apr 29	Max Born
Roosevelt	Einstein	1945, Mar 25	F. D. Roosevelt
Curie	Einstein	1911, Nov 23	Marie Curie
Switzer	Einstein	1953, Apr 23	J. S. Switzer
Palestine	Einstein	1948, Apr 10	Shepard Rifkin
Szilard	Einstein	1939, Aug 2	F. D. Roosevelt
Daughter	Einstein	¿1903,Sep 19?*	Lieserl Einstein

(*) There is no agreement at this dated.

Results

Figure 1 shows the result of normalization of the frequency of words obtained from the English letters of Albert Einstein when the n-gram is equal to two (ie two words). It can be seen that the alleged letter to daughter does not follow the same pattern as her other letters, with the words with the biggest difference such as "*of the*", *"it is*", *"in a*", "*is the*", "*and the*", "*to me*", among others.

Figure 2 shows the result of the normalization of the frequencies of words obtained from the letters of Simón Bolívar using n-gram equal to 1 and MRW = 50. This graph shows differences between the letters, such as "*mi*", "*los*", "*para*", "*he*", "*si*", "*dinero*", among others.

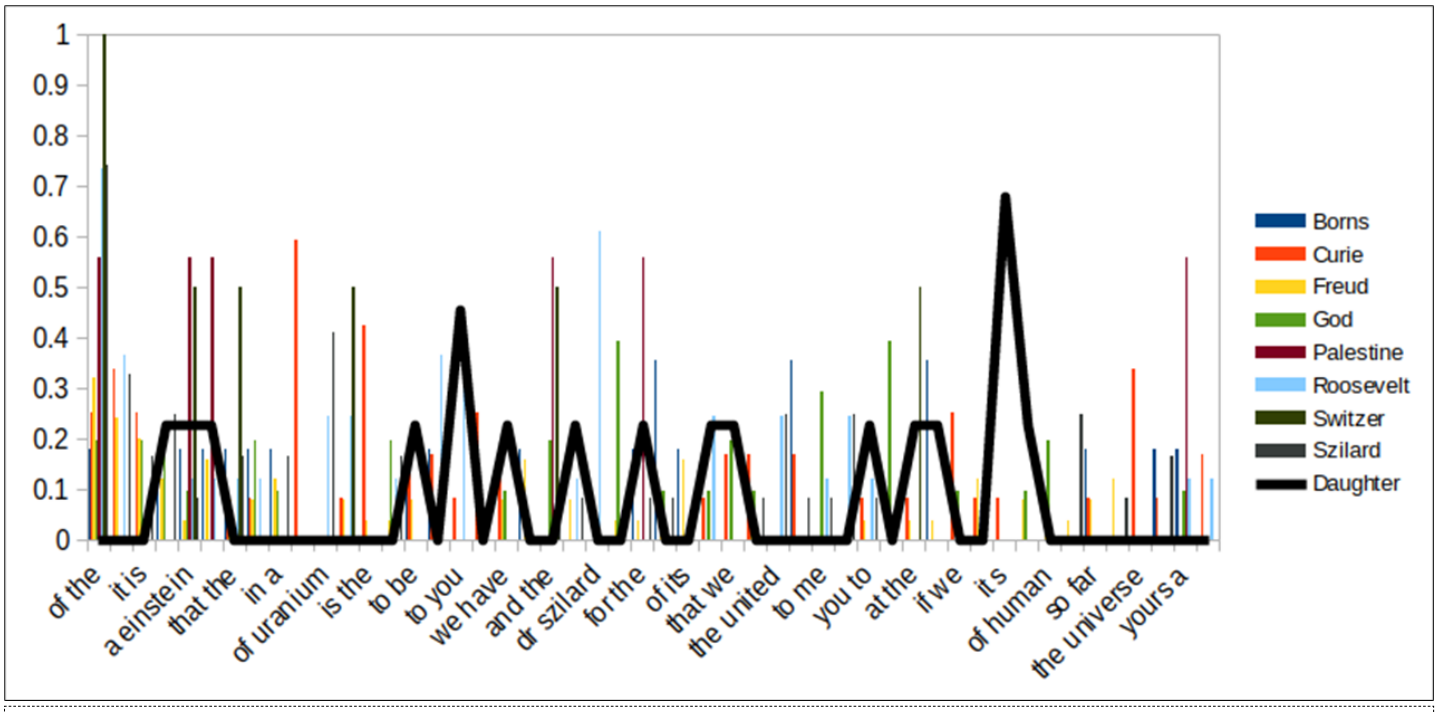

Figure 1. Normalization of the frequency of appearance of words obtained from Albert Einstein's letters, using the Manhattan distance, and MRW equal to 35.

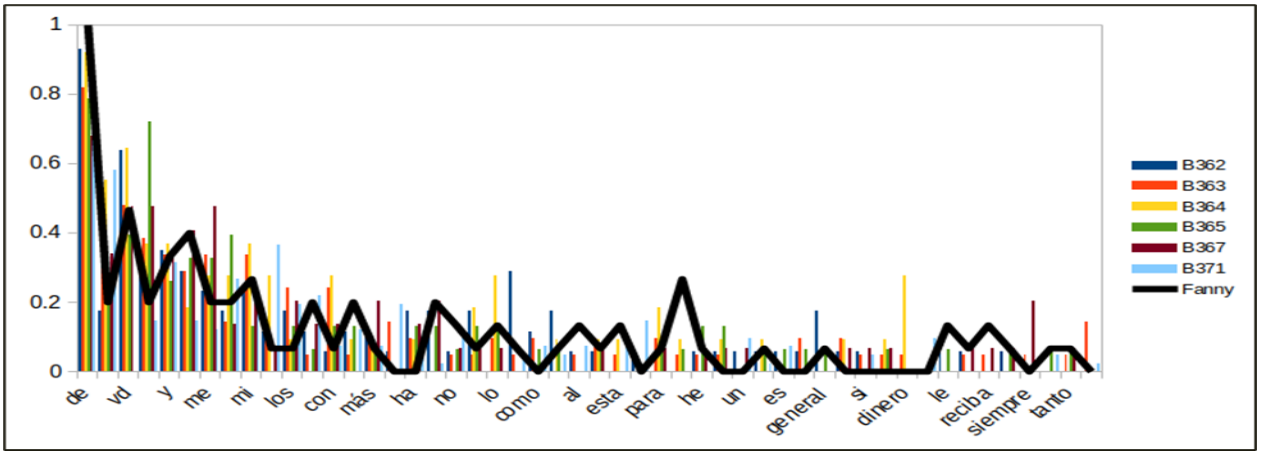

Figure 2. Normalization of the frequency of appearance of words obtained in the letters of Simón Bolívar using Manhattan's distance.

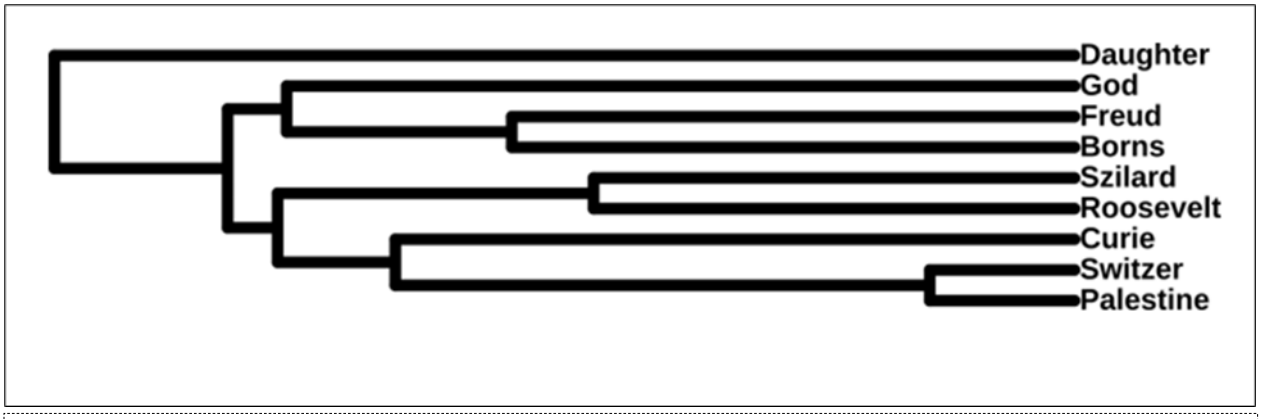

Figure 3. Rectangular dendrogram obtained from a stylometric analysis of Albert Einstein's letters, visualized with iTOL (Letunic, 2021).

Figure 3 shows a rectangular dendogram of Albert Einstein's letters. It was observed that the supposed letter to the daughter does not reproduce the linguistic styles of the other eight letters, that is, the supposed letter (labeled Daughter) does not belong to any of the three groups (clusters) that make up the rest of letters.

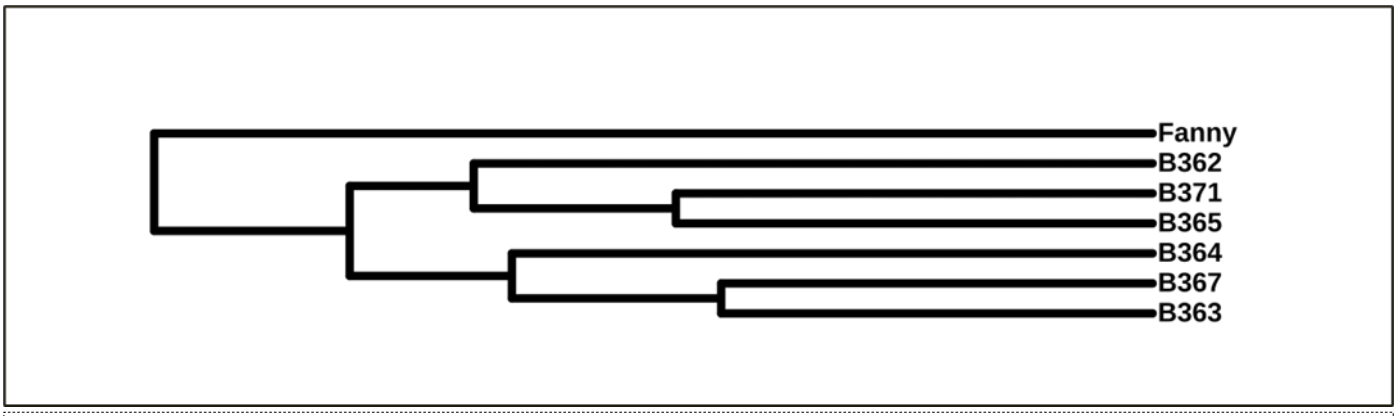

Figure 4. Rectangular dendrogram obtained from the stylometric analysis of the Bolívar letters, visualized with iTOL (Letunic, 2021).

Figure 4 shows a rectangular dendogram of Simón Bolívar's letters, showing that his cousin Fanny's letter does not belong to any of the two large conglomerates that repeat the style of Bolivar. Therefore, Fanny's letter was not written by Simón Bolívar.

Conclusion

The paper's goal was to determine whether Albert Einstein and Simon Bolivar actually penned amorous letters to their daughter and cousins, respectively, by looking for linguistic evidence in their letters. It is established that they did not write these tearful letters based on the findings of a stylometric examination.

References

1. Bensalem, I., Rosso, P. y Chikhi, S. (2019). On the use of character n-grams as the only intrinsic evidence of plagiarism, *Language Resources and Evaluation*. 53(3), 363–396. https://doi.org/10.1007/s10579-019-09444-w
2. Burrows, J.F. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3), 267-287. https://doi.org/10.1093/llc/17.3.267
3. Bvmc (2014, January 23) "Descubierta una comedia inédita de Lope de Vega" Available at https://blog.cervantesvirtual.com/descubierta-una-comedia-inedita-de-lope-de-vega/
4. Daelemans, W. (2013). Explanation in Computational Stylometry. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2013. Lecture Notes in Computer Science, vol 7817. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37256-8_37
5. Eder, M. (2017), Visualization in stylometry: Cluster analysis using networks, *Digital Scholarship in the Humanities*, 32(1), April 2017, 50–64. https://doi.org/10.1093/llc/fqv061
6. Fuller, S. y O'Sullivan, J. (2017). Structure over Style: Collaborative Authorship and the Revival of Literary Capitalism. *Digital Humanities Quarterly*, 11 (1). http://dx.doi.org/10.17613/M6BH0D
7. Letunic, I. y Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic

tree display and annotation. *Nucleic Acids Research*, 49(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

8. Lutoslawski, W. (1898). Principes de stylométrie appliqués à la chronologie des œuvres de Platon. *Revue des Études Grecques*. 11(41), 61–81. https://doi.org/10.3406/reg.1898.5847.
9. Mendenhall, T. (1887). The characteristic curves of composition. *Science*, 214, 237-249.
10. Ríos-Toledo G, Posadas-Durán JPF, Sidorov G, Castro-Sánchez NA. (2022)- Detection of changes in literary writing style using N-grams as style markers and supervised machine learning. PLoS One. Jul 20;17(7):e0267590. doi: 10.1371/journal.pone.0267590.
11. Schaalje, G.B., and others (2011). Extended nearest shrunken centroid classification: A new method for open-set authorship attribution of texts of varying sizes, *Literary and Linguistic Computing*, Volume 26, Issue 1, April 2011, Pages 71–88, https://doi.org/10.1093/llc/fqq029
12. Schoenbaum, S. (2018). Internal evidence and Elizabethan dramatic authorship; an essay in literary history and method, p. 196. Northwestern University Press (15 Octubre 2018) ISBN 0810138662
13. Škorić, M., Stanković, R., Ikonić, N. M., Byszuk, J., Eder, M. (2022). Parallel Stylometric Document Embeddings with Deep Learning Based Language Models in Literary Authorship Attribution. *Mathematics*, 10(5):838. https://doi.org/10.3390/math10050838
14. Stamatatos, E. (2009). A survey of modern authorship attribution methods. *JASIST*, 60 (3), 538–556. https://doi.org/10.1002/asi.21001
15. Stanikunas, D., Mandravickaite, J. y Krilavicius (2017). Comparison of distance and similarity measures for stylometric analysis of Lithuanian texts. CEUR Workshop proceedings [electronic resource]: ICYRIME 2017 : proceedings of the symposium for young researchers in informatics, mathematics and engineering, Kaunas, Lithuania, April 28, 2017. Aachen : CEUR-WS, 2017, Vol. 1852
16. Stefan, E., et al (2017). Understanding and explaining Delta measures for authorship attribution, *Digital Scholarship in the Humanities*, 32(suppl_2), ii4–ii16. https://doi.org/10.1093/llc/fqx023.