

Carnegie Mellon University
HeinzCollege

INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

94806-Z3

An Examination of the Impact of Stylometry, Artificial
Intelligence/Machine Learning (AI/ML) on Privacy in Social Media

Authors: Robert Conrad, Arthur Neumann, Elizabeth Sims
{rconrad, aneumann, esims}@andrew.cmu.edu

Professor: Alessandro Acquisti

Table of Contents

Abstract	3
Introduction	4
Types of Stylometry	4
Forensic Linguistics	4
Authorship Attribution	4
Plagiarism detection	4
IQ Estimation	5
Techniques	5
Markov Models	5
N-grams and Skip-grams	5
Hapax Legomenon	5
Machine Learning	6
Readability Index	6
Distant and Close Reading	6
History and Background	7
Highlights	7
Wincenty Lutoslawski	7
Plato's Dialogues	8
Federalist Papers	8
Shakespeare	8
Unabomber	9
Related Research	9
IQ Testing	9
Music	9
Paintings	9
Defeating Stylometry	9
Experiment	10
Methodology	11
Technology Background and Limitations	11
ChatGPT	11
JGAAP	12
Researcher Developed Python Application	12
Process	12
User selection	12
Data selection	12
Data Normalization	13
Data Ingestion	14

Results	14
ChatGPT	14
JGAAP	16
Custom Python Application	17
Findings	19
Limitations of Research	20
Ability to De-Anonymize	21
Implications to Privacy	21
Conclusion	22
References	24
Appendix	27
Appendix A: Custom Python Code	27

Abstract

This paper delves into the practice of stylometry, which involves analyzing writing samples to attribute authorship of previously anonymous texts. With the advancement of machine learning and artificial intelligence, stylometric predictions are becoming increasingly viable. However, the complexities of adversarial authorship, ghostwriting, and adjustments to writing style based on context make it imperative to apply stylometry with proper care.

The paper begins by reviewing publications on stylometry, including successful and unsuccessful attempts to apply the concept and real-world use cases. The limitations and challenges of applying stylometry are then discussed, including the technical acumen required to implement existing tools or create new ones selecting which features of the writing to focus on, and collective viable data sets to fuel analysis (Ding, Fung, Iqbal, & Cheung, 2017).

The paper also highlights significant privacy concerns associated with using stylometry to identify individuals on anonymous platforms. Stylometry could compromise individual privacy, track individuals across different platforms and websites, and even lead to false accusations and mistaken identities. As such, it is important to carefully consider the legal and ethical implications of using stylometry to identify individuals, given its significant potential as a useful tool.

The experimentation conducted in this paper attempted to apply stylometry using a variety of platforms for training and employing ML models. The purpose of the experimentation was to identify whether an average user, without extensive knowledge of machine learning, powerful computing hardware, or abundant funding, could reasonably use it to de-anonymize someone online.

Overall, while stylometry is still likely several years away from being a commonly available tool in the commercial industry, it is a valuable area of study that will inevitably make its way to and through the most relevant fields. It is critical to safeguard the privacy and personal data of individuals through regulations to ensure an appropriate balance between the benefits and drawbacks of using stylometry to identify individuals on anonymous platforms.

Introduction

Stylometry is the practice of analyzing various writing samples in an attempt to attribute authorship of previously anonymous texts. Stylometry is not a concept that is unique to the digital age, but with the advancements of technologies like machine learning and artificial intelligence, the ability to parse mass amounts of data to fuel predictions is dramatically increasing the applicability of the concept. That said, even with today's technology, there are still limitations to stylometry, and it should not be used to claim authorship definitively but rather provide a probability. In the digital age, many people find comfort in the privacy they achieve through anonymity on online platforms. With advanced capabilities becoming more accessible to everyday users, privacy through anonymity may become a thing of the past. Through this project, we conducted a thorough review of publications on the topic of stylometry, including both successful and unsuccessful attempts to apply the concept and real-world use cases. We then conducted our own experiments to attempt to conduct stylometry, using a variety of platforms for training and employing ML models. The purpose of this experimentation will be to identify where the barrier to entry lies in applying stylometry for average users. To do these, we conducted tests using techniques with varying levels of difficulty in application and varying levels of expected and actual success. Finally, based on both the literature review and our experiments, we will comment on what effects stylometry and machine learning may have on intrusions to privacy for those using anonymous digital platforms, now and in the future.

Types of Stylometry

Forensic Linguistics

This is a field that involves the analysis of language data to help solve crimes. Stylometry is one of the techniques used in forensic linguistics to identify the authorship of disputed documents. Forensic linguistics also includes other areas of analysis, such as discourse analysis and phonetics (Coulthard, Johnson, & Wright, 2016).

Authorship Attribution

This is a sub-field of stylometry that focuses specifically on identifying the author of an anonymous or disputed text. It looks at various techniques such as word length, syllables, parts of speech distribution, function words, and entropy, among others (Holmes, 1994).

Plagiarism detection

Stylometry can also be used to detect cases of plagiarism by comparing the linguistic style of a text to that of other texts to identify similarities in language and writing style. One technique is to use Latent Semantic Indexing (LSI) to identify relationships between words. For example,

by finding the words "dog," "cat," and "fish" together, one can assume the text is about animals (Alsallal, Iqbal, Amin, & James, 2013).

IQ Estimation

Stylometry has been used to estimate a person's IQ by analyzing linguistic features in written texts. While this is not a direct application of stylometry, it demonstrates how stylometric techniques can be used in other fields of study. Estimating IQ from written texts is reasonably new, and stylometric analysis has been proposed to estimate an individual's IQ by analyzing the number of SAT words in a body of text (Adebayo & Yampolskiy, 2022).

Techniques

Markov Models

Markov models compare a text's linguistic style sequence based on the preceding words in the series. The likelihood of the next word only depends on the immediately preceding words. For example, a first-order Markov model would predict the probability of the next word based on the preceding word. A second-order Markov model would consider the two preceding words. Higher-order Markov models are more complex and accurate but require more data and computational resources. Speech recognition, machine translation, and data types all use Markov models. In stylometry, it is another tool to determine the likelihood of an author's word choices (Goldman & Allison, 2008).

N-grams and Skip-grams

N-grams, are sequences of n items from a given text sample, where the items can be characters, syllables, or words. Text mining, natural language processing, and stylometry all use N-grams. For example, "education" is a 9-gram of characters, and "need to know" is a 3-gram of words. In specific applications, as with Twitter, determining authorship may prove even more challenging due to the limited number of characters allowed in digital texts.

Skip-grams are an N-gram model where the focus is not on consecutive sequences of words but on the co-occurrence of words within a particular space. Natural language processing uses skip-grams for word embedding and text classification tasks. For example, a skip-gram with a window size of 4 would consider the four words to the left and right of the target word (Sharon Belvisi, Muhammad, & Alonso-Fernandez, 2020).

Hapax Legomenon

A straightforward method for attributing authorship of distinct texts to the same author is through hapax legomenon analysis. Hapax legomena are words or phrases that appear only once in a given text, and its name derives from the Greek for "something said only once."

Several factors may influence the number of hapax legomena in a text, including its length, topic, audience, and passage of time. The Token Type Ratio (TTR) is employed to evaluate the linguistic richness of a text, and it is defined as the total number of unique words (types) divided by the total number of words (tokens) in a specific segment of language. The TTR can provide insight into the reading complexity of a text, which is linked to the number of unique words present (SerHack, 2022).

Machine Learning

Stylometry analyses often use machine learning algorithms to extract features from the text. Machine learning is a broader field that involves the development of algorithms that can learn patterns in data without being explicitly programmed. While stylometry is a specific application of machine learning to textual data, machine learning can also be applied to many other types of data (Savoy, 2020).

Readability Index

The readability score of a written text is a quantitative measure of its ease or difficulty in reading. The underlying concept is that readers possess varying levels of literacy. Professional writing and editing firms that enlist ghostwriters and editors use readability indexes to standardize the readability of each paragraph. By calculating the readability index of each sentence or paragraph, one can assess the intended level of readability and identify differences in writing styles. One of the most common readability indices is the Flesch-Kincaid Grade Level. It uses the average number of syllables per word and the average number of words per sentence. Calculating using the formula: $0.39 \times (\text{Total Words} / \text{Total Sentences}) + 11.8 \times (\text{Total Syllables} / \text{Total Words}) - 15.59$ gives the grade level. This particular readability formula stresses sentence length over word length (Kincaid, Fishburne Jr, Rogers, & Chissom, 1975).

Stylometry utilizes readability indexes to contrast texts by various authors and distinguish their writing styles. For instance, a high readability index in a text might indicate an author's preference for using intricate vocabulary and sentence structures, while a low readability index might imply a tendency towards more straightforward words and sentences.

Distant and Close Reading

The term "close reading" refers to a focused examination of the text, which includes an analysis of word choice, syntax, and specific imagery, to reveal its intended meaning. The roots of this approach can be traced back to New Criticism, which emphasizes an examination of the complexities of the individual text and does not rely on historical or biographical research.

Franco Moretti introduced the concept of distant reading, which proposes that analyzing a broader range of literature through computational or archival methods can provide a broader perspective, enabling us to identify significant trends and bring systems previously hidden within the

literary study to light (Moretti, 2013).

History and Background

Stylometry is a computational linguistic approach to identifying the author of a written text through the analysis of unique writing styles. This field of study is helpful when the author of a text is unknown or disputed. Every writing style possesses subtle indicators or authorial fingerprints determined through writing analysis. Researchers identify various features in stylometric analysis, including lexical, syntactic, semantic, structural, and subject-specific features (SerHack, 2022).

The use of stylometry in authorship attribution has achieved high accuracy, even when the number of candidate authors is high (Wang, Juola, & Riddell, 2022)(Brennan, Afroz, & Greenstadt, 2012). Some studies have found that stylometry can correctly identify the author of a text with a success rate of over 90% (Ahmed, Javed, Jalil, & Iqbal, 2020). Stylometry is especially useful for identifying differences in writing styles between authors and the similarities in writing techniques between different texts written by the same author (Mahor & Kumar, 2022).

However, stylometry also has some limitations, including the requirement for a sufficient amount of pre-existing writing data for the candidate author and the need to account for the stylistic variations in writing that can occur over time and across different contexts (Wang et al., 2022). Moreover, stylometric analysis cannot be 100% accurate in all situations since individuals can vary their writing styles, and there can be confounding factors such as ghostwriting, co-authorship, or plagiarism.

Stylometry successfully determines the author of a written text with high accuracy in many cases, making it an essential tool for forensic and computational linguistics research. However, there are limitations to the approach, and the writing analysis must be performed carefully, including selecting appropriate features and using a sufficient amount of pre-existing writing data for the candidate author.

Highlights

Wincenty Lutoslawski

Wincenty Lutoslawski was a Polish scholar and philosopher. He invented the term "stylometry" in his book *Principes de stylométrie* (Lutoslawski, 1898). Lutosławski used stylometry to solve the unresolved problem of periodizing Plato's Dialogues. He developed his method based on comparing stylistic text characteristics and believed their style could determine the order of Plato's writings. Lutosławski's method relied on several premises, including the existence of individual style in texts, the possibility of deciding authorship based on style, and the analogy between stylometry and graphology. He also emphasized the hierarchy of the importance of stylistic features and the necessity of comparing samples of equal length.

Lutosławski used his stylometric method to establish a complete chronology of Plato's works by comparing disputed and undisputed texts based on their style. Despite the criticism and the development of more advanced techniques, Lutosławski's proposition is still recognized in some Hellenistic circles. However, modern published studies do not often cite Lutosławski's methodology (*Wincenty Lutoslawski*, 2007).

Plato's Dialogues

The principles behind Lutosławski's application of stylometry helped determine the chronological order of Plato's works. His method included the importance of identifying reliable information about the dating of texts, the existence of individual style in an author's writing, and the possibility of identifying an author's identity based on the stylistic properties of their writing. Some of the limitations of Lutosławski's method include the difficulty of identifying a limited set of relevant stylistic features and the problem of determining the hierarchy of importance of these features (Keyser, 1992).

Federalist Papers

Frederick Mosteller and David Wallace applied statistical methods to the problem of authorship attribution regarding the Federalist Papers. They determined the likelihood that each of several possible authors (Alexander Hamilton, James Madison, and John Jay) wrote each paper. The authors used a Bayesian approach to calculate the probabilities of each author being the actual author of each paper. They also examine the reliability of their results using various statistical tests, including a chi-squared test and a likelihood ratio test. They concluded that their approach was effective in determining authorship attribution (Mosteller & Wallace, 1964).

Shakespeare

The authorship of William Shakespeare's plays has been the topic of much debate among scholars and enthusiasts for centuries (Shapiro, 2011). While there is no definitive answer to this question, stylometric analysis has been used to determine whether Shakespeare was the author of the plays attributed to him.

Stylometry is the study of writing style, which can be used to analyze the linguistic features of an author's work and compare them to other works. One common technique is to analyze the frequency and patterns of words, sentence structures, and other aspects of language usage. In the case of Shakespeare, scholars have used a range of stylometric methods, including word frequency analysis, vocabulary analysis, and analysis of syntax and other linguistic features.

While stylometry has been used to make some claims about the authorship of Shakespeare's plays, there is still much debate among scholars about its validity and reliability. Some critics argue that stylometric methods are unreliable enough to make definitive conclusions about authorship and that the linguistic patterns used to identify Shakespeare's work could also appear in the cre-

ation of other writers. Others argue that stylometric analysis can provide valuable insights into Shakespeare's writing style and evolution (Aljumily, 2015).

Unabomber

The Unabomber's capture was the result of linguistic analysis conducted by Kaczynski's brother and sister-in-law, not by a computer, as some computational linguistics and semantics stories suggest. Computer analysis did not aid in obtaining a search warrant or prosecuting Kaczynski. However, the FBI did hire a forensic linguist, James Fitzgerald, to compare Kaczynski's writings to the Unabomber's manifesto. Fitzgerald's testimony was presented in court during the trial, including Kaczynski's remark about modern philosophers not being "cool-headed logicians," a phrase his brother had never heard anyone else use (Storage, 2012).

Related Research

IQ Testing

Psychologists have sought to quantify intelligence for the last hundred years. Using stylometry to estimate IQ with machine learning has been effective, but a lack of large data sets has limited research in this area. Studies on IQ estimation from written text using stylometry can succeed with a suitable data set (Adebayo & Yampolskiy, 2022).

Music

Musical stylometry is a branch that uses computational methods to analyze and characterize music, including its authorship, genre, and period. Machine learning algorithms analyze musical features. There is difficulty in identifying the unique features of a particular composer or time period and the subjective nature of the musical analysis at this time (Kroonenberg, 2021).

Paintings

The developing field of visual stylometry of art aims to identify the creator of a painting through machine learning applied to high-resolution digital images. In one study, tests on over 100 high-resolution digital images of impressionist paintings by Van Gogh and contemporaries show good separation between the paintings of Van Gogh and others. Using stylometric techniques can help determine a painting's actual authorship (Qi & Hughes, 2011).

Defeating Stylometry

Stylometry is a technique used to attribute authorship to anonymous or disputed texts. However, it is not always reliable and can be defeated. Some of the techniques and strategies include the following:

Altering writing style: One way to defeat stylometry is to alter one’s writing style deliberately. This can be done by changing vocabulary, sentence structure, punctuation, or other stylistic features. By doing this, an author can create a different “wordprint” less likely to match their known writing style. However, it is necessary to note that this can be difficult to do consistently over a long piece of writing.

Using ghostwriters or co-authors: Another way to defeat stylometry is to use ghostwriters or co-authors to obscure the authorship of a text. They have someone else write the text entirely or by having multiple authors contribute. This method can make it more difficult to discern the unique writing style of a single author. This strategy is commonly used in ghostwriting or in the case of a corporate or political speechwriter.

Adding noise: Adding random or irrelevant content to a text can make it more difficult to discern the true authorship. One can do this by adding extra words or sentences that do not relate to the content of the text or by intentionally introducing errors or misspellings. However, this technique can also make the text more difficult to read and understand.

Limiting the amount of available data: Stylometry algorithms require a large amount of data to attribute authorship to a text document accurately. Limiting the amount of data available can make it more difficult for these algorithms to identify the author.

Using a style transfer algorithm: A style transfer algorithm can be used to change the style of a text without changing its content. This can be done by training the algorithm on two texts: one written by the actual author and one written by a different author. The algorithm can then be used to change the style of the actual author’s text to match that of the other author, making it more difficult to attribute authorship.

Using a language translation tool: A language translation tool can be used to translate a text into a different language and then back into the original language. This can introduce errors and stylistic changes that make attributing authorship more difficult.

It is worth noting that while these techniques can make it more challenging to attribute authorship, none of them are foolproof. However, the techniques discussed above can be helpful to those wishing to remain anonymous (Brennan et al., 2012).

Experiment

In order to effectively comment on the accessibility of applying stylometry using openly available analysis tools, we attempted to implement multiple different techniques with varying levels of complexity and technical understanding in order to deanonymize an individual or predict whether the same author wrote two or more samples. Through the use of multiple techniques we attempted to identify where the barrier to entry is in successfully applying this technique. Our experiment did not seek to exhaust all potential options for deanonymizing an individual posting to social media; instead, we sought to explore three separate technologies requiring varying levels of skill and identify a potential barrier to entry required to utilize readily available technology solutions to

deanonymize an individual based on social posts made on open social media platforms.

Methodology

Our methodology focused on the implementation and analysis of three technologies requiring varying levels of configuration and a collection of 50 social media posts obtained from multiple known users on two separate social media platforms, Reddit.com and Twitter.com. In addition to testing an online AI/ML platform ChatGPT, and an open-source stylometry software suite, we attempted to implement our own machine learning application in Python on readily available consumer-grade user hardware with a foundational knowledge (less than six months) of python application development education and no prior experience with machine learning application development.

Technology Background and Limitations

As part of our experiment, we leveraged three technologies, the newly released ChatGPT Artificial Intelligence / Machine Learning (AI/ML) model for direct submission and queries, the Java Graphical Authorship Attribution Program (JGAAP), and a custom python application developed by our team leveraging imported natural language processing and machine learning modules. The submission of all testing was performed within a Linux Mint or Ubuntu virtual machine configured with 4 CPU cores, 8GB of memory, and no graphics acceleration hosted on consumer-grade hardware. To maximize repeatability, all testing could be executed on any system with internet connectivity and the ability to install a modern internet browser, the latest long-term release of the Java Runtime Environment (JRE), and a Python interpreter.

ChatGPT

ChatGPT is an online chat service provided by the AI/ML research and deployment company OpenAI (OpenAI, 2020, 2022). The model being used to support ChatGPT was refined from a model in the GPT-3.5 series that was completed in early 2022. The ChatGPT model was trained using Reinforcement Learning from Human Feedback (RLHF) using human AI trainers to refine and tune the model (OpenAI, 2022). OpenAI (OpenAI, 2022) defines the following key limitations of ChatGPT:

- Occasionally provides “plausible-sounding but incorrect or nonsensical answers” (OpenAI, 2022)
- Sensitivity to modifications of input phrasing or trying the same prompt multiple times
- “Often excessively verbose and overuses certain phrases, such as restating that it’s a language model trained by OpenAI” (OpenAI, 2022)
- Often guesses when the user request is ambiguous
- Occasionally responds to harmful or inappropriate requests

In addition to these limitations, due to the model’s training period, the model and the fact that it is not connected to the internet is not able to use internet look-ups to find information and has limited information about events occurring after 2021 (Staud, 2023).

JGAAP

Java Graphical Authorship Attribution Program (JGAAP) is an open-source authorship attribution tool developed by Evaluating Variation in Language (EVL) Lab at Duquesne University (Evaluating Variation in Language Lab, 2021). Our team implemented the model using a download of the open-source application, which was configured and executed using publicly available online resources with no prior knowledge of stylometry terminology or the application’s functionality.

Researcher Developed Python Application

Due to the limited time frame in which our research could be conducted, and the intention of determining if an individual with limited programming skills and no formal training in AI/ML could deanonymize an individual by comparing posts from two social media sources, our model was developed using python and by a student with introductory python knowledge, using only publicly available and open source online resources. As part of the execution of our code, we implemented the Python Natural Language Toolkit (NLTK) to implement natural language processing functionality and the Scikit-learn (Sklearn) package to implement machine learning functionalities.

Process

To standardize our process across each technology, we collected, obtained, and input the same raw text data set from multiple users. We had each technology analyze the data with the goal of deanonymizing an individual.

User selection

Our process began with a selection of social media posts from popular social media accounts on two platforms. We selected five users with an account on both platforms with at least 50 posts. To simplify our experiment, we chose mostly public figures with the exception of one user under the handle “iBleedOrange” who has a following on social media but is not a well-known public figure like the others.

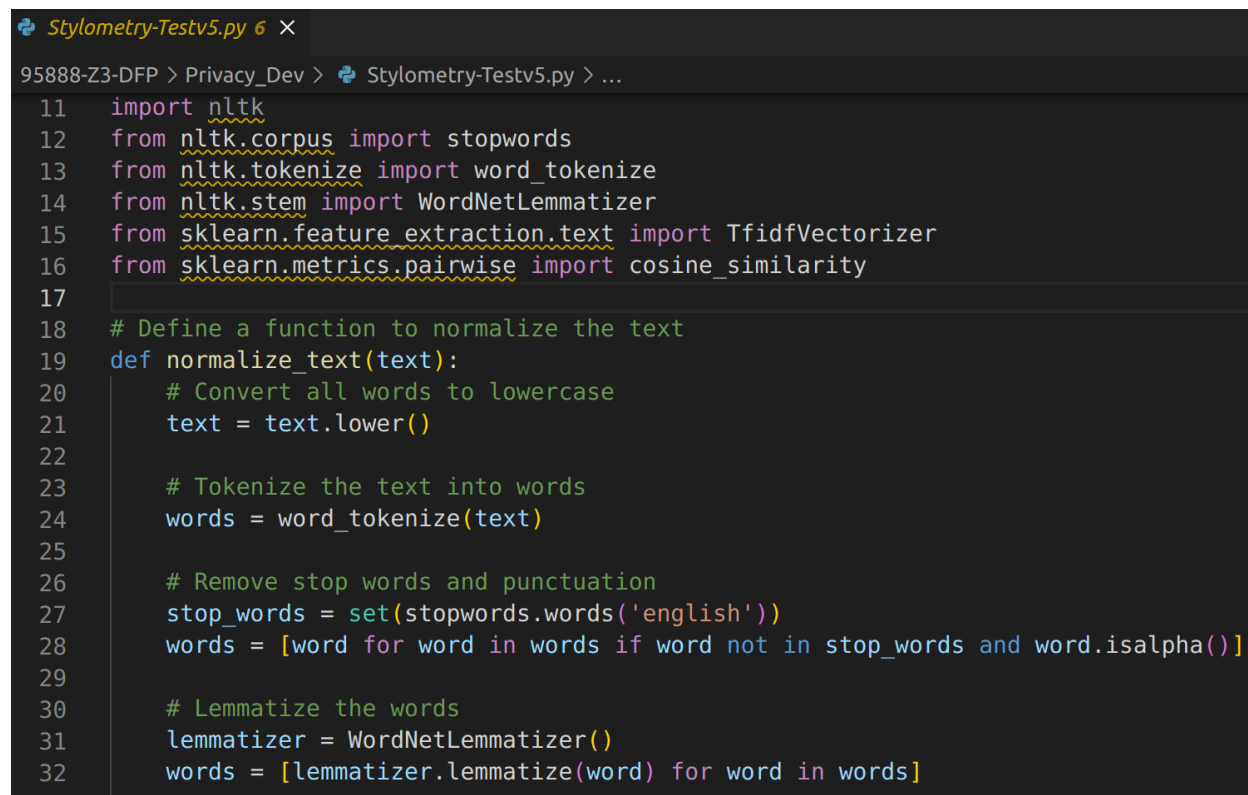
Data selection

From each user, we selected and captured 50 publicly available posts from each social media account in a raw text file. While we trended toward longer posts in an attempt to obtain the most data possible within our constraints, there were instances where some of the chosen post content was quite small (< 50 characters). Each post was placed in a text file with one blank line between

to delineate each separate post. We also validated that the user’s real name was not identified in the file name or within the file during input to any application as a precautionary measure.

Data Normalization

All posting data was imported into each different application as raw data, and normalization was applied using different methodologies. For ChatGPT, no normalization was applied as part of the data ingestion; however, following submission, control of normalization was ceded to the model. For JGAAP, the “Normalize ASCII,” and “Normalize Whitespace” options were applied as canonizers to remove non-text data and whitespace from the source material. For the custom python application, the data was normalized by converting all text to lowercase, splitting the text into a list of words (tokenizing). The list was then trimmed of all English stop words and non-textual information found using the NLTK stop word English dictionary and text analysis. Finally, each word was reduced to its base form or root (lemmatizing) using methods imported from the NLTK toolkit and then rejoined into a complete string variable with a single space inserted in between each word using the python join method (Prabhakaran, 2022). Data normalization tool imports and methods used in the custom python application are illustrated below in Figure 1.

A screenshot of a code editor window titled 'Stylometry-Testv5.py'. The code is a Python script for text normalization. It imports 'nltk' and 'sklearn' modules. It defines a function 'normalize_text(text)' which performs the following steps: 1. Convert all words to lowercase using 'text.lower()'. 2. Tokenize the text into words using 'word_tokenize(text)'. 3. Remove stop words and punctuation by creating a set of stop words from 'nltk.corpus.stopwords.words('english')' and filtering the words list. 4. Lemmatize the words using 'WordNetLemmatizer()' from 'nltk.stem'. The script is shown in a dark-themed editor with line numbers 11 through 32 visible.

```
11 import nltk
12 from nltk.corpus import stopwords
13 from nltk.tokenize import word_tokenize
14 from nltk.stem import WordNetLemmatizer
15 from sklearn.feature_extraction.text import TfidfVectorizer
16 from sklearn.metrics.pairwise import cosine_similarity
17
18 # Define a function to normalize the text
19 def normalize_text(text):
20     # Convert all words to lowercase
21     text = text.lower()
22
23     # Tokenize the text into words
24     words = word_tokenize(text)
25
26     # Remove stop words and punctuation
27     stop_words = set(stopwords.words('english'))
28     words = [word for word in words if word not in stop_words and word.isalpha()]
29
30     # Lemmatize the words
31     lemmatizer = WordNetLemmatizer()
32     words = [lemmatizer.lemmatize(word) for word in words]
```

Figure 1: Python Input Data Normalization

Data Ingestion

For input into ChatGPT, we implemented the following standardized process for each analysis:

1. The following text was input into the chat window: “Can you tell me who wrote this text and give a percentage level of confidence:”
2. The raw text from each file was copied into the chat after the initial prompt surrounded by quotations as shown below in Figures 2 and 4
3. In cases where the raw text was too large for input into the ChatGPT and caused an error return, the text was reduced in size to meet submission requirements
4. In instances where ChatGPT did not provide an initial level of confidence, we prompted for clarification by inputting ”Can you give a percentage level of confidence?” as shown below in Figure 5

Results

Each subsection below provides a detailed report of our findings for each analysis tool.

ChatGPT

Using our standardized input for requesting the identification of the author of social media posts, we were able to identify the social media posts from three public figures on both Reddit and Twitter with a confidence level ranging from 95% to 100%, a public figure only on Twitter at a greater than 99% confidence level, and were unable to identify the poster of a non-public figure. Figures 2 - 5 below demonstrate the results obtained using ChatGPT, and Table 1 below shows our findings of authorship attribution using ChatGPT.

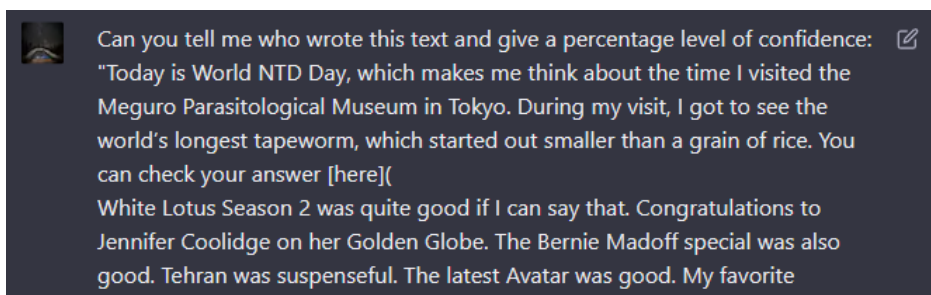


Figure 2: ChatGPT Text Input - Bill Gates Reddit Posts

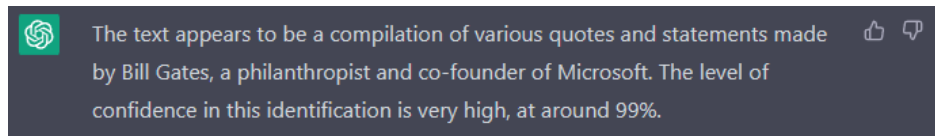


Figure 3: ChatGPT Response - Bill Gates Reddit Posts

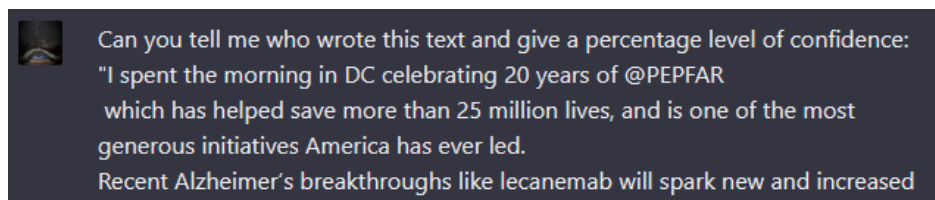


Figure 4: ChatGPT Text Input - Bill Gates Twitter Posts

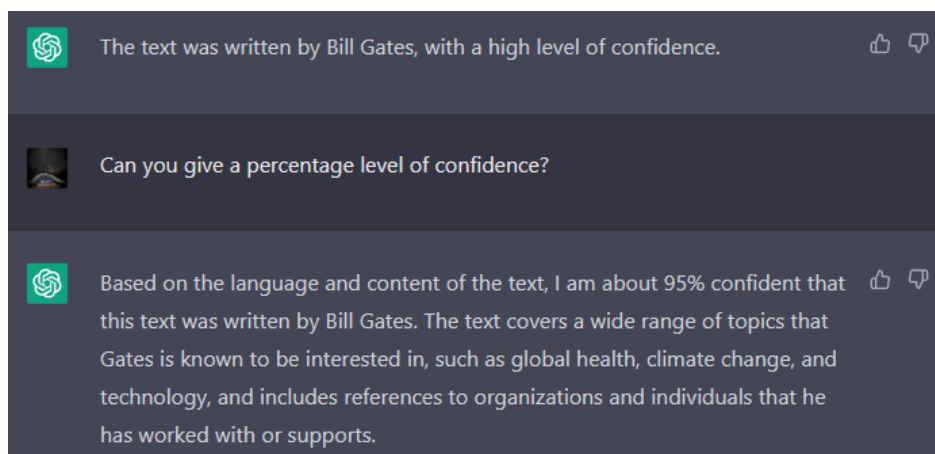


Figure 5: ChatGPT Response - Bill Gates Twitter Posts

Table 1: Attempting De-Anonymization Using ChatGPT

Name	Public Figure	Reddit		Twitter	
		Identified Y/N	Confidence %	Identified Y/N	Confidence %
Chris Hadfield	Y	Y	100%	Y	100%
Snoop Dogg	Y	Y	95-99%	Y	95%
Bill Gates	Y	Y	99%	Y	95%
Wil Wheaton	Y	N	N/A	Y	>99%
iBleedOrange	N	N	N/A	N	N/A

JGAAP

Java Graphical Authorship Attribution Program (JGAAP) is a Java-based tool built by the Evaluating Variation in Language (EVL) Lab at Duquesne University (Evaluating Variation in Language Lab, 2021). Per the application documentation, “JGAAP is a tool to allow non-experts to use cutting edge machine learning techniques on text attribution problems” (Evaluating Variation in Language Lab, 2021). While advanced knowledge of machine learning is not necessary to use this tool, it is still helpful to get full use of it. In order to test whether this tool is both accessible to and effective in allowing average people to use stylometry to identify authors, we input our data set and conducted several hundred iterations of analysis and kept track of which settings and features resulted in the highest success rate in pairing the samples from Reddit with the associated authors’ sample from Twitter.

To do this, we used examples from the source documentation on the program and experimentation until we identified certain features, such as parts of speech, words, and sentence length, that had a high informational return on the samples. Ultimately, by configuring the five Twitter samples as known authors and the five Reddit samples as unknown authors, we were able to pair three of the five writing samples successfully. Many experiments resulted in a worse return rate. This is not to say that this tool is not capable of having higher success rates but that a more significant knowledge of machine learning and stylometry is necessary for the most effective use of the tool. In the figure below are the results of one of the five text samples analyzed, specifically Bill Gates. The section marked “A” is the unknown text sample, “B” is the normalization of the data applied, “C” is the specific features of the text analyzed, “D” is the analysis method used, and “E” is the results of the analysis against the known author samples. Below the JGAAP Results graphic is a table depicting the full results of the same test in the graphic, which was our most successful test. In the section showing the order of predictions for each unknown sample, the names have been abbreviated for ease of viewing.

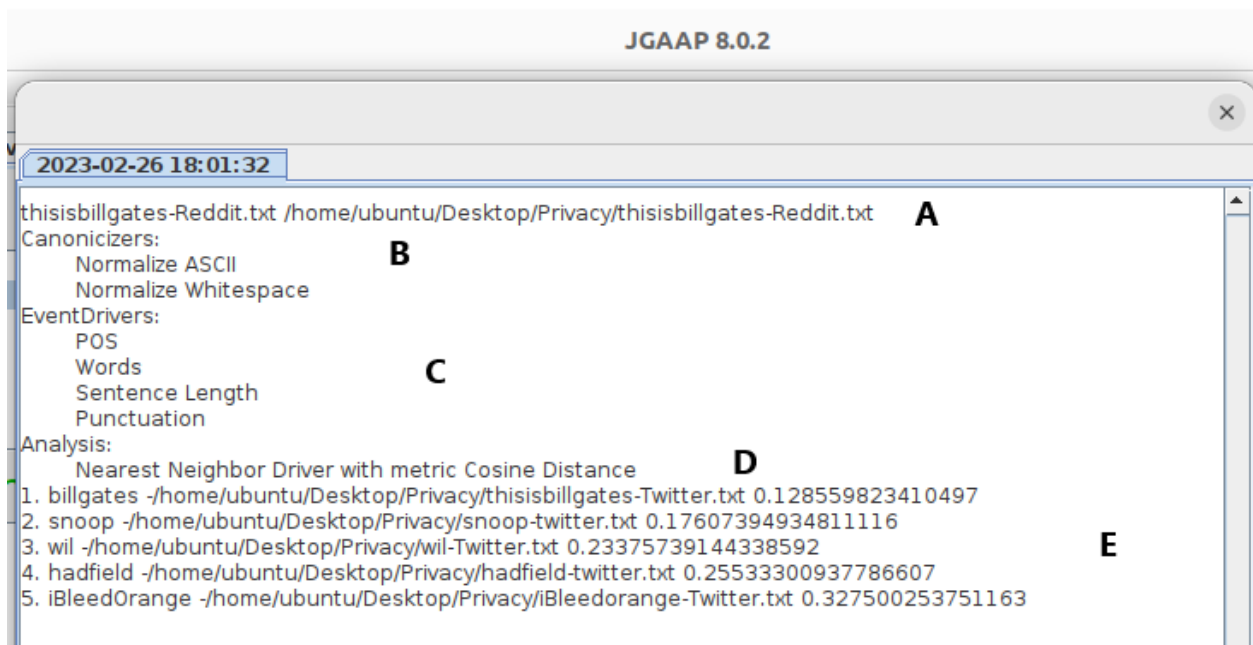


Figure 6: JGAAP Results - Bill Gates

Table 2: Attempting De-Anonymization Using JGAAP

Author of Unk Sample	ID'd Y/N	Prediction Order & Metric				
		#1	#2	#3	#4	#5
Chris Hadfield	N	Wil/.030	Bill/.075	Chris/.196	Snoop/.328	iBleed/.361
iBleedOrange	N	Bill/.211	Snoop/.224	Chris/.232	iBleed/.234	Wil/.265
Snoop Dogg	Y	Snoop/.212	iBleed/.301	Chris/.419	Bill/.440	Wil/.455
Bill Gates	Y	Bill/.129	Snoop/.176	Wil/.234	Chris/.255	iBleed/.328
Wil Wheaton	Y	Wil/.097	Bill/.101	Chris/.199	Snoop/.227	iBleed/.270

Custom Python Application

Using our custom-developed Python application, we compared the text from two files to determine if the authorship of social media posts aggregated as raw text in one file could be attributed to the author of the social media posts aggregated as raw text in another file. After each revision, we revised the code to add or modify functionality in an attempt to improve the results. To determine similarity, each normalized word list was then converted to a vector using the SKLearn TfidfVectorizer fit_transform method to calculate the “mean and variance of features” and then compared the two vectorized results for cosine similarity using the SKLearn cosine similarity method (Khanna, 2020). Our test relied on a cosine similarity of 0.9 or greater to attribute the authorship of the first file to the second. Figures 7 and 8 below illustrates the analysis and comparison of the text samples and sample result output from a completed run. Table 3 below provides a record of our comparison findings for each author using our custom-developed Python application.

```

Stylometry-Testv5.py 6 X
95888-Z3-DFP > Privacy_Dev > Stylometry-Testv5.py > ...
11 import nltk
12 from nltk.corpus import stopwords
13 from nltk.tokenize import word_tokenize
14 from nltk.stem import WordNetLemmatizer
15 from sklearn.feature_extraction.text import TfidfVectorizer
16 from sklearn.metrics.pairwise import cosine_similarity
17
18 # Define a function to normalize the text
19 def normalize_text(text):
20     # Convert all words to lowercase
21     text = text.lower()
22
23     # Tokenize the text into words
24     words = word_tokenize(text)
25
26     # Remove stop words and punctuation
27     stop_words = set(stopwords.words('english'))
28     words = [word for word in words if word not in stop_words and word.isalpha()]
29
30     # Lemmatize the words
31     lemmatizer = WordNetLemmatizer()
32     words = [lemmatizer.lemmatize(word) for word in words]
33

```

Figure 7: Python Data Comparison

```

(base) cfc@cfc-mint:~/git/95888-Z3 DFP/Privacy_Dev$ /home/cfc/anaconda3/bin/python "/home/cfc/git/95888-Z3 DFP/Privacy_Dev/Stylometry-Testv5.py"
The cosine similarity between file1.txt and file2.txt is 0.29106436140684727
The texts do not have the same author.

```

Figure 8: ChatGPT Text Input - Bill Gates Reddit and Twitter Posts

Table 3: Attempting De-Anonymization Using Custom Python Application

Name	Identified Author (Y/N)	Cosine Similarity
Chris Hadfield	N	0.290
Snoop Dogg	N	0.260
Bill Gates	N	0.291
Wil Wheaton	N	0.396
iBleedOrange	N	0.110

Findings

Our original intent with experimentation was to identify an approximate barrier to entry for average people to apply stylometry to de-anonymize people on online platforms. The higher the barrier to entry is, the less risk there is of people being de-anonymized online using this technology, and the lower it is, the more risk there is. ChatGPT, while being the easiest to implement, likely did not exclusively use stylometry in order to identify the authors of the provided samples. It provided high confidence of authorship attribution on three of five Reddit samples and four of five Twitter samples, with the caveat being that it was not able to identify the author that is not famous in the traditional sense. This leads us to believe that it was only able to identify these authors based on context clues, writing style, and possibly by matching the samples directly to portions of the internet that it scanned during training.

In the second level of difficulty, we used the pre-existing program JGAAP to apply stylometry to our data set. This method had moderate difficulty and moderate success. In order to use this software, you have to have at least some computer literacy to be able to get it up and running, and then fairly advanced knowledge of machine learning to apply it at pique effectiveness. Our own knowledge level of machine learning concepts is self-assessed somewhere between non-existent and entry-level, and we were able to get some results from the program, but increased knowledge of the various options available in the program would likely increase effectiveness.

Finally, we attempted to train and apply our own machine-learning models in python. This method is by far the most technically demanding, both from a knowledge and time investment standpoint. As evident by the background and history of this technology, many people and teams across academia have been successful in training models to tackle different approaches to the problem of authorship attribution and overcoming challenges that make it difficult. That said, the vast majority of these individuals are studying advanced degrees and concepts in the fields of machine learning, computer science, and data science. Our attempts to apply our limited knowledge to building one were wholly unsuccessful. Indicating that for the average person with limited knowledge and time to commit, it is likely unrealistic at this stage to de-anonymize someone online.

Generally speaking, software and tools that are easy to implement to de-anonymize individuals online are probably years, if not longer, away from breaking out of academia into common commercial use. So, while stylometry is very much possible, the first phases of real-world uses will probably be limited to well-funded organizations on the cutting edge of technology and perhaps intelligence

communities.

Limitations of Research

There are several limitations that both we and academia at large face when attempting to apply stylometry for authorship attribution. Some universal challenges and limitations are going to be having a sufficient number and size of samples, the technical acumen to implement existing tools or create your own, and selecting which features of the writing you want to focus on. The limitations of various use cases are broader. For example, arguably, the easiest method of application is when you have two text samples, and you are attempting to identify whether one person authored both (Ding et al., 2017)(Koppel, Schler, & Argamon, 2009). In this case, the sample size and technical acumen required to come to a conclusion are relatively low. Taking the next step in terms of difficulty is attempting to identify the author of a sample when you have a relatively small pool of candidates (Ding et al., 2017). The challenges, in this case, would be narrowing down this candidate pool, knowing when the author does not match any candidates at all, and having large sample sizes for both the unknown works and the candidates. Another concept that is similarly moderately difficult is using a writing sample to identify the characteristics of the author. In this case, instead of identifying a specific person, the idea is to narrow the pool by using stylometry to know the gender, age, personality traits, etc., of the author (Koppel et al., 2009).

Finally, the hardest method of application is taking a text sample and attempting to find the author without anything else (Stolerman, Overdorf, Afroz, & Greenstadt, 2014)(Koppel et al., 2009). For example, you might take a sample from a known social media platform, such as Twitter, and attempt to find the user's corresponding Reddit account. This would require a huge amount of technical acumen, investment of time, and the assumption that the user has sufficient posts on both platforms. Other complications also arise with this approach, such as the fact that a person's writing style may vary widely depending on the context, purpose, and audience of their posts, or they may even intentionally alter it. Not only could this vary from platform to platform, but even thread to thread within the same platform. This establishes a need for an even greater breadth and depth of writing samples from the target. The concept of intentionally altering your writing style to obfuscate authorship is considered adversarial though some have shown success in academic settings with overcoming this (McDonald, Afroz, Caliskan, Stolerman, & Greenstadt, 2012; Brennan et al., 2012).

There were also multiple limiting factors that we faced while conducting our own research. In order to know if your attempt to identify authors is successful or not, you must have two samples that you know to be from the same person. In order to increase the size of the data sets, other researchers have artificially split samples from the same author to create pairs to match. Two complications arise with this method. First, artificially splitting a sample likely means that both halves will be samples from a similar or even the same context, i.e., two chapters from the same book or a handful of posts on the same social media platform. The other complication is that it ensures that when analyzing an unknown sample, there is a guaranteed match somewhere in the

known sample data set (Koppel et al., 2009). In order to represent a real-world scenario that could affect someone's privacy achieved through anonymity on a platform such as Reddit, we chose to use samples from two entirely separate platforms in order to capture the differences between them, even among the same author. This meant we had to identify authors who used both platforms manually, had at some point named themselves on both platforms and had sufficient posts on each to meet a minimum threshold. This was further exacerbated by the fact that we were only able to procure API keys for Reddit and not Twitter, which meant that we could only automate the collection for half of the posts from each candidate in our data set.

Ability to De-Anonymize

Based on our research of the background and history, as well as our own experimentation, we assess that using stylometry to de-anonymize is not yet a capability easily accessible to internet users outside of those in specialized fields of research. This assessment comes at different levels depending on exactly what methods are employed, but generally speaking, success rates without advanced training are likely to be low. That said, specific scenarios in which people may be able to find success are when they have a small number of but large sample sizes to compare. An example of this may be attempts to identify the true author of a book that someone may suspect to be written under a pseudonym, especially if they have a limited number of candidates which they suspect. Another scenario that can increase success rates is if the stylometric analysis of a text can be combined with other contributing factors, such as context clues. That said, if an individual was concerned their coworkers might identify their Reddit profile among the tens of millions of others on the platform, it would most likely not be possible, unless the individual in question had posted specific and identifying data on the platform.

Implications to Privacy

The identification of individuals on anonymous platforms using stylometry is a complex topic that raises a multitude of concerns regarding privacy and personal data protection. Although stylometry may not be easily accessible to the average person, continued research suggests that it could become more widespread among law enforcement and intelligence agencies, among others (Ekambaranathan, 2018). While it can be beneficial in several areas, including historical research and copyright purposes, it can also present privacy issues if used in specific contexts without the user's knowledge. Moreover, stylometry has been used in literary studies, forensic linguistics, and digital forensics to attribute authorship to anonymous texts, detect plagiarism, and track individuals' online activities.

Despite its potential benefits, the use of stylometry to identify individuals on anonymous platforms poses several privacy concerns that should not be overlooked. For one, individuals may expect to remain anonymous when posting on these platforms. Using stylometry to de-anonymize them may violate their privacy, particularly when they post sensitive or controversial information.

Stylometry could also be used to track individuals across different platforms and websites by analyzing text samples and creating a profile that reveals characteristics about the author, such as gender, age, interests, beliefs, and behavior (Koppel et al., 2009). This could be problematic if the information is used for nefarious purposes such as stalking or harassment. Writing style can also vary from context to context, with factors such as mood, topic, audience, and purpose all contributing to changes in one's particular stylistic fingerprint. This could lead to false accusations and mistaken identity, making it difficult to attribute authorship with certainty.

Furthermore, there is a risk of stylometry misuse by those who seek to harm or discredit others, including impersonation by mimicking their writing style. This could result in damage to an individual's reputation or lead to miscommunications. The legal and ethical implications of using stylometry to identify individuals require careful consideration. It raises several questions, including whether its use by law enforcement should be regulated and require a warrant and how individuals can be made aware of the potential risks and consequences of their online activity.

If stylometry were to become more accessible to the general public, it could further compromise individual privacy. Posting on social media platforms would be riskier, and public figures such as politicians and celebrities may face increased scrutiny. Additionally, the use of stylometry in litigation could increase, as it could be employed to attribute authorship to anonymous texts or detect plagiarism in legal proceedings. While stylometry has significant potential as a useful tool, it is critical to consider its privacy concerns and implement regulations to safeguard the privacy and personal data of individuals (Patergianakis & Limmiotis, 2022). It is also crucial to raise awareness of the potential risks and consequences of online activity and encourage individuals to take steps to protect their privacy. Ultimately, the benefits and drawbacks of using stylometry to identify individuals on anonymous platforms must be weighed carefully to ensure an appropriate balance.

Conclusion

The paper explores the practice of stylometry, which involves analyzing writing samples to attribute authorship of previously anonymous texts. With the advancements in machine learning and artificial intelligence, stylometric predictions are increasing in viability. However, given complications such as adversarial authorship, ghostwriting, and adjustments to writing style based on context, it should still be applied with proper care. The paper began with a review of publications on the topic of stylometry, including successful and unsuccessful attempts to apply the concept and real-world use cases. We then conducted our own experiments to attempt to apply stylometry using a variety of platforms for training and employing ML models. The purpose of the experimentation was to identify whether an average user, without extensive knowledge of machine learning, powerful computing hardware, or abundant funding, could reasonably use it to de-anonymize someone online. We then discussed the limitations and challenges of applying stylometry, including the technical acumen required to implement existing tools or create your own and selecting which features of the writing to focus on.

We also highlighted the significant privacy concerns associated with using stylometry to identify individuals on anonymous platforms. The use of stylometry could compromise individual privacy, track individuals across different platforms and websites, and lead to false accusations and mistaken identities. Furthermore, stylometry could be misused by those with malicious intent, including impersonation, by mimicking their writing style. It is important to carefully consider the legal and ethical implications of using stylometry to identify individuals, given its significant potential as a useful tool. Safeguarding the privacy and personal data of individuals through regulations is critical to ensure an appropriate balance between the benefits and drawbacks of using stylometry to identify individuals on anonymous platforms.

Overall, stylometry is still likely several years away from being a commonly available tool in the commercial industry. As it stands now, it is not a simple task for someone outside of the field of data science or machine learning to reasonably build their own stylometric solutions. That said, it is certainly a valuable area of study and will inevitably make its way to and through the most relevant fields, such as law enforcement, intelligence, and publishing (Ding et al., 2017).

References

- Adebayo, G. O., & Yampolskiy, R. V. (2022). Estimating intelligence quotient using stylometry and machine learning techniques: A review. *Big Data Mining and Analytics*, 5(3), 163–191.
- Ahmed, W., Javed, A. R., Jalil, Z., & Iqbal, F. (2020). Authorship analysis with machine learning. In D. Phung, G. I. Webb, & C. Sammut (Eds.), *Encyclopedia of machine learning and data science* (pp. 1–4). New York, NY: Springer US. Retrieved from https://doi.org/10.1007/978-1-4899-7502-7_986-1 doi: 10.1007/978-1-4899-7502-7_986-1
- Aljumily, R. (2015). Hierarchical and non-hierarchical linear and non-linear clustering methods to “shakespeare authorship question”. *Social Sciences*, 4(3), 758–799. Retrieved from <https://www.mdpi.com/2076-0760/4/3/758> doi: 10.3390/socsci4030758
- Alsallal, M., Iqbal, R., Amin, S., & James, A. (2013). Intrinsic plagiarism detection using latent semantic indexing and stylometry. In *2013 sixth international conference on developments in systems engineering* (p. 145–150). doi: 10.1109/DeSE.2013.34
- Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3), 1–22.
- Coulthard, M., Johnson, A., & Wright, D. (2016). *An introduction to forensic linguistics: Language in evidence* (2nd ed.). Routledge. Retrieved from <https://doi.org/10.4324/9781315630311> doi: 10.4324/9781315630311
- Ding, S. H., Fung, B. C., Iqbal, F., & Cheung, W. K. (2017). Learning stylometric representations for authorship analysis. *IEEE transactions on cybernetics*, 49(1), 107–121.
- Ekambaranathan, A. (2018, July). *Using stylometry to track cybercriminals in darknet forums*. Retrieved from <http://essay.utwente.nl/75908/>
- Evaluating Variation in Language Lab. (2021). *Java graphical authorship attribution program*. Duquesne University. Retrieved from <https://evllabs.github.io/JGAAP/>
- Goldman, E., & Allison, A. (2008). *Using grammatical markov models for stylometric analysis. class project, cs224n*. Stanford University. Retrieved from: <http://nlp.stanford.edu/courses> . . .
- Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, 28, 87–106.
- Keyser, P. (1992). Stylometric method and the chronology of plato’s works. *Bryn Mawr Classical Review*, 3(1), 58–73.
- Khanna, C. (2020, Dec). *What and why behind fit-transform() vs transform() in scikit-learn! Towards Data Science*. Retrieved from <https://towardsdatascience.com/what-and-why-behind-fit-transform-vs-transform-in-scikit-learn-78f915cf96fe>
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel* (Tech. Rep.). Naval Technical Training Command Millington TN Research Branch.

- Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1), 9–26.
- Kroonenberg, P. M. (2021). Musical stylometry: Characterisation of music. In *Multivariate humanities* (pp. 347–370). Cham: Springer International Publishing. doi: 10.1007/978-3-030-69150-9_18
- Lutoslawski, W. (1898). Principes de stylométrie appliqués à la chronologie des œuvres de platon. *Revue des études grecques*, 11(41), 61–81.
- Mahor, U., & Kumar, A. (2022). A comparative study of stylometric characteristics in authorship attribution. In *Information and communication technology for competitive strategies (ictcs 2021) ict: Applications and social interfaces* (pp. 71–81). Springer.
- McDonald, A. W., Afroz, S., Caliskan, A., Stolerman, A., & Greenstadt, R. (2012). Use fewer instances of the letter “i”: Toward writing style anonymization. In *Privacy enhancing technologies: 12th international symposium, pets 2012, vigo, spain, july 11-13, 2012. proceedings 12* (pp. 299–318).
- Moretti, F. (2013). *Distant reading*. Verso Books.
- Mosteller, F., & Wallace, D. (1964). Inference and disputed authorship the federalist (p 16) reading. *Massachusetts: Addison-Wesley*.
- OpenAI. (2020, Sep). *About openai*. Author. Retrieved from <https://openai.com/about/>
- OpenAI. (2022, Nov). *Chatgpt: Optimizing language models for dialogue*. Author. Retrieved from <https://openai.com/blog/chatgpt/>
- Patergianakis, A., & Limniotis, K. (2022). Privacy issues in stylometric methods. *Cryptography*, 6(2), 17-35. doi: 10.3390/cryptography6020017
- Prabhakaran, S. (2022, Apr). *Lemmatization approaches with examples in python*. Retrieved from <https://www.machinelearningplus.com/nlp/lemmatization-examples-python/>
- Qi, H., & Hughes, S. (2011). A new method for visual stylometry on impressionist paintings. In *2011 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 2036-2039). doi: 10.1109/ICASSP.2011.5946912
- Savoy, J. (2020). *Machine learning methods for stylometry : authorship attribution and author profiling* (1st ed. 2020. ed.). Cham, Switzerland: Springer.
- SerHack, S. (2022, Mar). *Unveiling the anonymous author: Stylometry techniques*. SerHack Security Research. Retrieved from <https://serhack.me/articles/unveiling-anonymous-author-stylometry-techniques/>
- Shapiro, J. (2011). *Contested will: who wrote shakespeare?* Simon and Schuster.
- Sharon Belvisi, N. M., Muhammad, N., & Alonso-Fernandez, F. (2020). Forensic authorship analysis of microblogging texts using n-grams and stylometric features. In *2020 8th international workshop on biometrics and forensics (iwbfb)* (p. 1-6). doi: 10.1109/IWBF49977.2020.9107953
- Staud, N. (2023). *Chatgpt general faq*. OpenAI. Retrieved from <https://help.openai.com/en/articles/6783457-chatgpt-general-faq>

Stolerman, A., Overdorf, R., Afroz, S., & Greenstadt, R. (2014). Breaking the closed-world assumption in stylometric authorship attribution. In *Advances in digital forensics x: 10th ifip wg 11.9 international conference, vienna, austria, january 8-10, 2014, revised selected papers 10* (pp. 185–205).

Storage, B. (2012, Aug). *Statistical stylometry*. Retrieved from <https://themultidisciplinarian.com/tag/statistical-stylometry/#:~:text=Robert%20was%20surprised%20to%20hear,and%20sister%2Din%2Dlaw>.

Wang, H., Juola, P., & Riddell, A. (2022). Reproduction and replication of an adversarial stylometry experiment. *arXiv preprint arXiv:2208.07395*.

Wincenty lutoslawski. (2007). Retrieved from http://www.glottopedia.org/index.php/Wincenty_Lutos%20%82awski

Appendix

Appendix A: Custom Python Code

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
```

Purpose: Script reads the comments from two text files and attempts to see if author of comment

Authors: Version 5 Initial Code Written by ChatGPT, refined and implemented by Arthur Neumann

Date: 02/25/2023

```
"""
```

```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
```

```
# Define a function to normalize the text
```

```
def normalize_text(text):
```

```
    # Convert all words to lowercase
```

```
    text = text.lower()
```

```
    # Tokenize the text into words
```

```
    words = word_tokenize(text)
```

```
    # Remove stop words and punctuation
```

```
    stop_words = set(stopwords.words('english'))
```

```
    words = [word for word in words if word not in stop_words and word.isalpha()]
```

```
    # Lemmatize the words
```

```
    lemmatizer = WordNetLemmatizer()
```

```
    words = [lemmatizer.lemmatize(word) for word in words]
```

```
    # Join the normalized words back into a string
```

```
    normalized_text = ' '.join(words)
```

```
        return normalized_text

# Load the contents of the two text files
with open('file1.txt', 'r') as f1:
    file1 = f1.read()
with open('file2.txt', 'r') as f2:
    file2 = f2.read()

# Normalize the text in each file
file1_normalized = normalize_text(file1)
file2_normalized = normalize_text(file2)

# Calculate the TF-IDF vectors for the normalized text
vectorizer = TfidfVectorizer()
vectors = vectorizer.fit_transform([file1_normalized, file2_normalized])

# Calculate the cosine similarity between the two vectors
similarity = cosine_similarity(vectors[0], vectors[1])

# Determine if the texts have the same author based on the similarity score
if similarity > 0.9:
    print("The texts have the same author.")
else:
    print('The cosine similarity between file1.txt and file2.txt is ' + str(similarity[0][0]))
    print("The texts do not have the same author.")
```