cryptography

MDPI

Article

Privacy Issues in Stylometric Methods

Antonios Patergianakis [1] and Konstantinos Limniotis [1,2,*]

[1] School of Pure and Applied Sciences, Open University of Cyprus, Latsia, Nicosia 2220, Cyprus; antonios.patergianakis@st.ouc.ac.cy
[2] Hellenic Data Protection Authority, Kifissias 1-3, 11523 Athens, Greece
* Correspondence: konstantinos.limniotis@ouc.ac.cy or klimniotis@dpa.gr

Abstract: Stylometry is a well-known field, aiming to identify the author of a text, based only on the way she/he writes. Despite its obvious advantages in several areas, such as in historical research or for copyright purposes, it may also yield privacy and personal data protection issues if it is used in specific contexts, without the users being aware of it. It is, therefore, of importance to assess the potential use of stylometry methods, as well as the implications of their use for online privacy protection. This paper aims to present, through relevant experiments, the possibility of the automated identification of a person using stylometry. The ultimate goal is to analyse the risks regarding privacy and personal data protection stemming from the use of stylometric techniques to evaluate the effectiveness of a specific stylometric identification system, as well as to examine whether proper anonymisation techniques can be applied so as to ensure that the identity of an author of a text (e.g., a user in an anonymous social network) remains hidden, even if stylometric methods are to be applied for possible re-identification.

Keywords: anonymity; personal data protection; privacy; stylometry

Citation: Patergianakis, A.; Limniotis, K. Privacy Issues in Stylometric Methods. *Cryptography* **2022**, *6*, 17. https://doi.org/10.3390/cryptography6020017

Academic Editor: Cheng-Chi Lee

Received: 28 February 2022
Accepted: 2 April 2022
Published: 7 April 2022

1. Introduction

The anonymisation of personal data so as to ensure that it is not possible to re-identify an individual or even to infer some conclusions on an identified person (even if she/he cannot be explicitly singled out from a list that is deemed to be anonymised) has received much attention during last years (see, for example, [1–3]). However, it is well known that ensuring anonymity is a difficult task—especially in the era of big data, which allows for efficient data mining and further processing so as to correlate different datasets and establish connections between them (see, for example, [4–7]). The inherent difficulty in ensuring anonymity may lead to the violation of the rights and freedoms of an individual in cases that a person assumes erroneously that she/he remains anonymous—e.g., in an anonymous social network or in a whistleblowing system.

One such technology that could put privacy at risk in the case that it is not used in a fair manner is stylometry. Indeed, stylometry may be used to identify the author of a text, based only on the way she/he writes, through proper analysis of the content (based, for example, on frequently used words or phrases) and comparison with other texts of the same author. Stylometric techniques offer several benefits in fields such as reliable authorship attribution as well as in copyright investigation or in detecting harmful content in media—whereas the identification of specific writing types is also important for specific medical diagnoses [8]. However, at the same time, despite its obvious advantages, it may also yield privacy issues. For example, by applying stylometric techniques, a nickname may no longer be enough to ensure "anonymity" on the internet, even if it does not allow, by itself, the identification of its holder; although it may be already known that a user should not simply rely on a "meaningless" nickname to hide her/his identity, it is questionable whether the users are actually aware of the actual risks for re-identification stemming from the effectiveness of stylometric techniques.

This paper aims to study stylometry, focusing on its possible effects in cases that their use may be a threat for privacy and/or personal data protection; these are the cases for which data are considered anonymous and the relevant individuals have such expectations—i.e., they believe that they actually remain anonymous. A characteristic example is the case of anonymous social media. It is out of our scope to study and evaluate the benefits that stylometry offers that also rely on identifying the writer of the text; we assume only use cases in which such an identification could lead to the violation of fundamental human rights of privacy and/or personal data protection; to our knowledge, such an aspect of stylometry has not been studied to a great extent. Therefore, under this assumption, we investigate to what extent stylometry suffices to identify the author of a given text, under several hypotheses regarding the size and the type of the texts that are available for applying stylometric techniques. To achieve this goal, appropriate experiments are conducted, focusing on four types of texts: books, articles in blogs, emails and social media posts; by these means, it is also possible to identify whether specific types of texts are more "vulnerable" to re-identification attacks through stylometry than others. Moreover, the effectiveness of stylometry for several different sizes of input messages are also examined (taking into account that, on the Internet, messages may be of quite a small size—see, for example, so-called tweets). As a stylometric technique, we use, for all cases, statistical analysis of the so-called functional words between a set of authors, with known texts corresponding to these authors as the input.

This preliminary analysis illustrates that stylometry may indeed lead to the successful identification of the author of the text, especially in cases that large volumes of data are available. However, as it was actually expected, we notice that, for small texts, such a re-identification may not be an easy task. We also consider, as a safeguard against such re-identification attacks, possible techniques to thwart these attacks—i.e., to lower the effectiveness of stylometry; these techniques are based on the appropriate modification of texts so as to "mask" those elements that facilitate re-identification without changing their actual meaning. Such an "anonymisation" approach in the context of stylometry has not been investigated so far in the literature—at least not to a significant extent (one work in this direction is [9]).

The paper is organised as follows: First, Section 2 presents the necessary background, covering the notion of stylometry, as well as the notions of privacy and personal data protection; to this end, we use the European legal provisions for personal data protection, where the basic legal instrument is the General Data Protection Regulation (GDPR). Next, Section 3, being the main part of the paper, presents the methodology that we followed for our study, the testing environment, as well as the results obtained through our experiments, discussing the relevant outcomes. Section 4 presents some results from a preliminary analysis that we performed towards developing techniques to obfuscate the writing style of an individual so as to protect her anonymity with respect to re-identification attacks based on stylometric techniques. Next, a discussion of possible future research steps is given in Section 5. Finally, concluding remarks are given in Section 6.

2. Background

2.1. Introduction to Stylometry

Stylometry is the analysis of the way that a piece of work, such as a literary text, is produced. The goal of stylometry is usually to extract information, including the identity of the author, the genre of the work, their age, etc. A more specific definition describes stylometry as the statistical analysis of variations of literary style between one author, or one genre, and another [10]. As a technique, stylometry has been used for centuries and there have been recorded uses since the Renaissance. Its main principles are described in the book "Principes de stylométrie" written by the philosopher Wincenty Lutosławski in 1890. The term stylometry is attributed to him.

In recent years, with the development of technology, stylometry has become even more relevant. The frequency of the use of so-called function words, the average sentence length and other spelling information are used as indications of an author's identifiable style. It should be pointed out that, due to the effectiveness of stylometric identification methods, the FBI and the DARPA consider the way of writing as a biometric feature, in the sense of a "cognitive fingerprint", similar to the movement pattern of the mouse when using a computer [11].

One area in which stylometry was used from a very early age is the field of forensics, with the aim to de-anonymise written communications in order to infer the author, to prove a text's authenticity or to investigate plagiarism [11,12]. Another domain is that of linguistics, with applications including the linguistic analysis of texts and the analysis of vocabulary choices with regard to the author, period, genre, etc., as well as in the literary science and sociolinguistic research (see, for example, [8]). Research has also been done on the evolution of the writing style in terms of the writing period of the author [13,14]. Furthermore, stylometry has been used for creating user profiles regarding the gender, age, educational level, country of origin, personality traits, existing mental disorders, etc. [14–16]. The extraction of such profiles has, in turn, been used in several areas, such as psychology, sociology, medicine, marketing, etc. In addition, stylometry can be used to automatically detect extremist content [17]. Another possible application area of stylometry is to identify fake news on the internet—although it was recently shown that that the effectiveness of stylometry is limited against machine-generated misinformation [18]. Stylometry can be also used for user authentication [19], especially if combined with other biometric features, such as typing, mouse movement and so on. In the same way, stylometry can be used to detect the specific preferences, views, ideologies, voting intent and so on of individuals—this is actually an application area that significantly raises privacy and personal data protection concerns. Lastly, stylometry may be used to identify users on the internet that utilise "anonymous" profiles, such as in the case of anonymous social media, or to link profiles of the same person across different social networks [10,20]. Although such a processing may be allowed in specific cases (e.g., in the context of criminal investigation, under the provisions of an applicable law that includes appropriate safeguards for respecting the rights and freedoms of individuals), it becomes evident that, at the same time, stylometric techniques may violate fundamental human rights.

This paper focuses on the more traditional use of stylometry—namely, on the identification of the author of a text (i.e., authorship attribution) based on its style.

2.2. Legal Aspects

The rights to privacy and personal data protection are recognised as fundamental human rights by several international treaties. For example, in Europe, both of these rights are being recognised as such in the EU Charter of Fundamental Rights. Due to this, there are specific requirements that should be met when the processing of personal data is taking place.

The main legal instrument in Europe, with respect to the personal data protection, is the so-called General Data Protection Regulation (GDPR)—which can be considered a good model for all legislations throughout the world. According to the GDPR, personal data refers to any information relating to an identified or identifiable natural person. The GDPR codifies the principles that need to be guaranteed when personal data are being processed and sets specific obligations to those that process personal data. The basic principles include, amongst others, fairness and transparency (i.e., the personal data should be processed lawfully, fairly and in a transparent manner) as well as the purpose limitation (i.e., personal data should not further processed in a manner that is incompatible with the initial, well-determined and transparent, purposes).

The GDPR also describes the anonymous data as the data for which the relevant person is no longer identifiable—thus, anonymous data are not personal data. However, it is also explicitly described that one should be very cautious before characterising data as anonymous since all the possible means that can be reasonably used to re-identify a person should be taken into account. Hence, an erroneous characterisation of data as anonymous may in fact lead to the violation of personal data protection legislation and affect humans rights and freedoms. Indeed, if a user assumes that they remain anonymous in a specific context, but a stylometric technique suffices, by reasonable means, to identify them or to link him/her with another electronic account, then their data become erroneously considered anonymous; such a re-identification may allow further processing of their data for purposes that are fully different from the original purpose of the data processing, without the user being aware of this—thus, contradicting the principles of fairness and transparency as well as of the purpose limitation. The same risks occur even if the original data are considered pseudonymous (which are still considered as personal data, according to the GDPR) and not anonymous; since pseudonymisation is a privacy enhancing technique that can be used, for example, for hiding identities, such a re-identification from pseudonymous data may also violate the above data protection principles.

Therefore, from the above, at least in specific cases, the risks of re-identification through stylometry should be taken into account by any entity that share—or process by any other means—text data that correspond to individuals.

2.3. Contribution of This Work

The purpose of this work is to study the effectiveness of stylometry in person identification. Our aim is firstly to verify the results of the existing research on the subject which indicate that stylometric methods can indeed be used to successfully identify the authors of disputed texts, such as literary works, based solely on their style of writing. Moreover, as a subsequent step, we focus on assessing the effectiveness of stylometry not only for large literary works, but also for various types of documents, thus making the results much more relevant for the everyday user. Our ultimate goal is to study texts that are relevant to internet users in order to assess possible privacy threats for them that occur due to the use of stylometric techniques, taking also into account the relevant legal provisions.

More specifically, this work studies the effectiveness of stylometry in terms of identifying the authors of texts, such as emails, social media posts and blog posts—i.e., texts that are typically generated on a daily basis in the Internet in vast amounts. Moreover, apart from studying different types of texts, we also examine the impact of the length of the texts, as well as of the number of known authors, with respect to determining the main factors that play the most important role in the success of a stylometric system.

In addition, another aspect of our work is to investigate whether it is possible to appropriately utilise the principles of the stylometric methods in order to "anonymise" a text. More precisely, we examine whether, in a stylometric system that is based on the so-called function words to identify the user, it is possible to render the re-identification impossible by simply replacing few such words with their synonyms; indeed, taking into account the relevant legal requirements, anonymisation techniques to alleviate issues stemming from stylometry may be necessary in specific cases. It is shown that the proposed approach seems to be promising, being an an interesting topic for further research.

3. Assessing the Risks or Re-Identification through Stylometry

In this section, the main approach that we followed towards evaluating the effectiveness of stylometry, as well as the relevant results and conclusions that were derived, are presented. We first analyse our testing environment and, subsequently, the experiments that we performed.

3.1. Methodology—The Testing Environment

In order to draw realistic conclusions regarding the threat of re-identification of Internet users stemming from the use of stylometry, we developed a stylometric attribution system and used it to execute comprehensive experiments on real data. Towards achieving this, the following decision steps occurred: (i) First, we had to choose the metric that would represent the style of writing. (ii) Next, we had to select the algorithm that would analyse and compare the individual styles of the known authors, as well as the style of the unknown text, in order to achieve the attribution of texts to well-determined authors.

For the above, we chose to measure the style based on the variation in the frequency of use of the most common function words, while using John Burrows' delta algorithm for measuring and comparing the different styles. Our decision was based on the fact that these two methods were already proven to give reliable results, even for small data sets [21,22]. At the same time, we had to compile collections of real texts of different types that would be used as the input to our system. Lastly, we implemented the stylometric attribution system as a practical web application, being called `ShadowCloak`, so as to be able to import the compiled texts and run our various test cases.

3.1.1. The Representation of the Style

One of the biggest factors that determines the success of a stylometric system in authorship attribution is the metric upon which the style of each author is measured. The relevant research shows that one very effective metric is the variation of frequency in the use of the n most common words, which are usually the so-called *function words* of the text. The value of n depends on many factors, such as the richness of the vocabulary included in the texts, the length of the disputed text, the number of known texts per author, the number and the difference in style between the known authors, etc. In our test cases, the number $n = 50$ of most common words was chosen for the `ShadowCloak` system, since this is the value that we derived, through some preliminary tests, as being the most suitable to provide the best results. Hence, the change in the frequency of occurrence of the most common words represents the way of writing of each author; the difference between their frequencies allows determining the author of the disputed text.

3.1.2. The Identification of the Author

Having a list of the frequencies of the most common words for each author and a list of the frequencies of the most common words of the disputed text, the author can be identified by comparison. However, to achieve this, a normalisation must be performed so that the results are not disproportionately affected by the difference between the most used features, which will have a much higher frequency. Moreover, a method of evaluating such deviations must be used. To accomplish these, the Burrows delta method [23] was used, as mentioned above.

The process we followed includes the following steps:

- Collection of the corpus of texts belonging to a number of well-known authors. The texts are considered a "bag" of words (bag of words).
- Finding the n most common words that will constitute the features for the specific corpus.
- Finding the percentage of occurrence of each word of the above features in the subcorpus of each author.
- Calculating the total mean value and the standard deviation of each feature for the whole corpus, based on the values previously calculated for each individual subcorpus. These are the average mean value and standard deviation of this feature for the whole corpus.
- Calculation of the z-score of each feature for each subcorpus, i.e., for each author. By this way, the deviation of the specific author in the use of the specific word (feature) from the standard percentage of the whole corpus is calculated.

The z-score is a fundamental concept in the Burrows algorithm, as it allows the normalisation of frequency measurements and the avoidance of Zipf's law [24], according to which just a few words, but with great frequency, would greatly affect the results. The z-score for each feature F in each subcorpus is calculated by the following formula:

$$z_i(F) = \frac{f_i(F) - \mu_i}{\sigma_i}$$

where $f_i(F)$ is the percentage of occurrence of the feature in the subcorpus, μ_i is the total mean value of the feature and σ_i is its standard deviation.
Up to this step in the algorithm, the z-scores for each subcorpus were calculated, so taking each z-score as an entry of a vector, an n-dimensional vector can be extracted for each author.

- The next step similarly includes the calculation of the z-score of each feature for the anonymous text whose author is investigated. By this way, we create an n-dimensional vector that represents the style of the unknown author.
- As a last step, we calculate the delta score for each subcorpus by comparing the distance of its z-scores vector with the z-score vector of the anonymous text. The delta score is calculated based on the following formula:

$$\Delta_s = \sum_i \frac{|z_{A_i}(F) - z_{t_i}(F)|}{n}$$

where $z_{a_i}(F)$ is the z-score of the feature i for the author A and $z_{t_i}(F)$ is the z-score of the feature i for the anonymous text.

The selected author is the one whose subcorpus has the vector with the shortest distance from the vector of the anonymous text. In other words, the author with the lowest delta score is selected as the author whose way of writing is the least different from the way of writing of the unknown text.

3.1.3. Collection of Texts

According to the procedure described so far, the author of the anonymous text is selected among the authors whose texts are known and have been entered into the system. Therefore, the appropriate data had to be found that could support the usage scenarios required by this research.

To cover a wide range of sizes and structures, four categories of texts were selected—namely, books, emails, articles (blog posts) and social media posts. All the texts used were real data taken from real-life scenarios.

The source of the books is Project Gutenberg [25], which is the oldest digital library and contains mainly free full-text books in various formats, without copyright restrictions. The books used were downloaded in plain text and inserted into the `ShadowCloak` application through its user interface since the books, due to their large size, did not require a large number of texts for stylometric analysis. The texts used for the book category are excerpts from classic books, masterpieces of classic literature.

For the emails, the corpus of Enron emails was used. Enron was an energy company in Texas, USA, which went bankrupt in 2001 due to fraud. Some of the emails of its employees were made public by the competent service during the completion of the fraud investigation. These include 200,000 emails from 150 users in plaintext. These emails are now publicly available for research purposes (see [26]). A subset of these was used to create the `ShadowCloak` application corpus solely for our research purposes and with the appropriate data protection safeguards, as described below. Emails were inserted into the application per author (employee) in an automated way, using scripts. It was actually the sent emails that were selected to create our corpus, as they were unique to each employee; the emails containing forwards and replies were removed so as to exclude the included messages of other employees.

Blog posts were extracted from two, randomly chosen, well-known blogs, publicly available, which are related with the field of information security. We identified the same people (columnists) writing several articles in these two blogs—so our aim was to investigate whether, having some known texts from these known authors, it is possible to identify some of their other texts whose author is assumed to be unknown.

Finally, the social media posts were extracted from the Twitter U.S. Airline Sentiment (TUAS) dataset [27], which is a collection of real posts on the Twitter social networking platform, which were posted and collected in February 2015 and refer to airlines. This collection is licensed for use only for non-profit research purposes. A subset of TUAS was used for the corpus of `ShadowCloak` social media posts. Its input to the application was performed automatically using scripts, after removing unnecessary metadata and special characters, such as emoticons and hashtags which are repetitive and could affect the results of the measurements.

3.1.4. Data Processing

After the above data were collected, a filtering process was employed so that only the raw content of the texts and the absolutely necessary metadata, such as the author, the category, etc., were kept. A second filtering was performed on the content, as also mentioned above, to remove special characters, formatting characters, email headers, emoticons, hashtags, and others. Afterwards, where necessary, the data were restructured in order to be stored in the database in a uniform way. To extract the actual features from the texts, yet another process was employed on the words that make up the features of each corpus; more precisely, using Python's Natural Language Toolkit (NLTK) library, i.e., the large library of natural language processing tools, the texts were split into a collection of tokens consisting of lowercase text, without punctuation or other non-alphanumeric characters. The same library was used to find the frequency of occurrence of the most common words in each corpus.

3.1.5. Our Application for Stylometric Analysis

The `ShadowCloak` application is the stylometric analysis system created explicitly for the needs of this research. It consists of two web applications, the first is a web API that stores in a database the texts that the user uploads and performs the stylometric analysis, while the second is a graphical interface that runs in a web browser through which the user can access the API.

The first application is written in Python, allowing to utilise libraries, such as the aforementioned processing library NLTK for the pre-processing of the input texts, and is based on the Django framework. The second application provides a graphical interface for using the backend system; to this end, the Angular framework was used due its feasibility, allowing the use through a web browser from any device.

The user interface is divided into four pages. The first is the homepage of the app, which provides some general information about stylometry. The next page is called "Documents", being the management page of each user's library. Through this page, the user can enter different texts in the database, create authors or categories of texts and relate all of them in such a way as to create the corpus, on which the stylometric analysis algorithm will export the style of known authors per category and will identify the author of the unknown texts. The third page, being called "Find Author" is the page where the user can enter a text and try to identify its author among the authors contained in the library (see Figure 1). The category of the text should also be specified so that the system will analyse only the texts belonging to this category. This functionality is particularly useful, as it allows testing different scenarios, as described above.

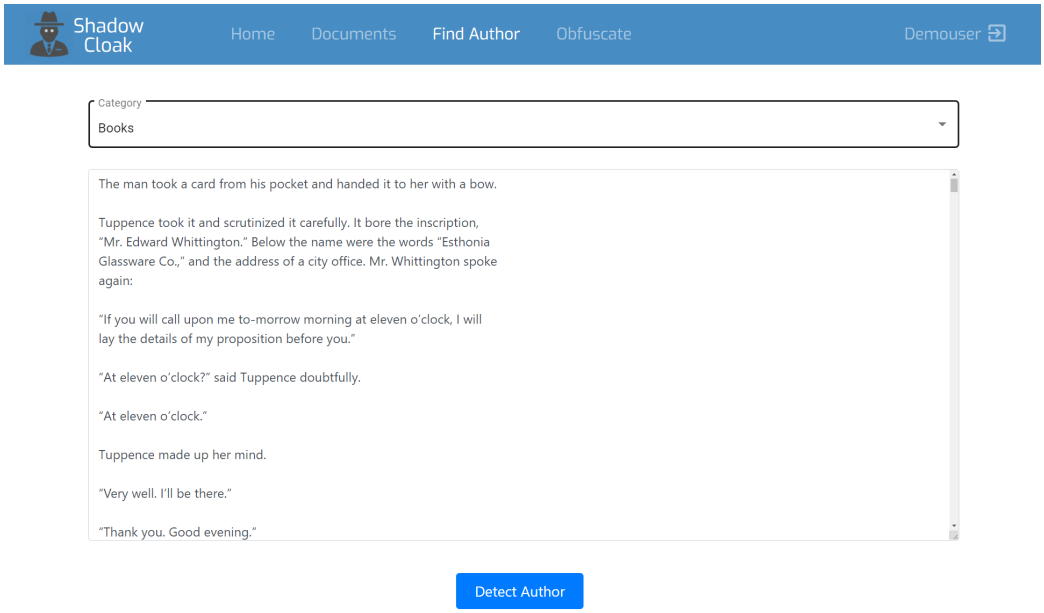

Figure 1. A screenshot from the `ShadowCloak` application, when trying to identify an author.

Once the user selects the "Find Author" button, the text and category are sent to the web API, where a number of procedures begin towards performing the analysis; these procedures include the transformation the texts into lists of words (tokens), which in turn consist of the content of the texts, after the punctuation marks and the non-alphanumeric characters have been removed, whilst they are also converted to lowercase letters. Since all the texts are converted into token lists, for each author, a list of the frequency of each token is calculated. This frequency, as well the frequency of each token for the entire corpus of texts, is then normalized in the manner previously described. Finally, the normalized frequencies of the most common tokens are compared to the frequencies for each author but also to the frequencies observed in the unknown text. From this comparison, we can infer which author has the most similar style with the style of the disputed text.

Once the described procedure is completed, the system informs the user by presenting them with the most probable author of the text.

In addition to identifying the author of a text, in this paper, we also investigate whether there is an efficient way to derive a method for avoiding the effectiveness of stylometrics in terms of identification so as to ensure anonymisation. In this way, the `ShadowCloak` application could be considered a tool providing anonymisation services, to thwart identification attacks based on stylometry. To achieve this, we identify the words that help most in the initial identification of the author and we subsequently gradually replace them with synonyms until a successful attribution to an author is not possible. This utility is available through the fourth page of the application, entitled "Obfuscate". By this page, the application displays to the user the tokens that had the most significant impact in identifying the author and at the same time they become underlined and coloured, as proper candidates to be replaced towards hiding the authorship.

In this page, the original author, i.e., the one whose style best matches the style of the unknown text, is listed at the top of the page, in the "Suggested Author" field. The user is able to choose which tokens (being coloured, as mentioned above) should be replaced (see Figure 2); this process can continue until the author of the text cannot be identified.

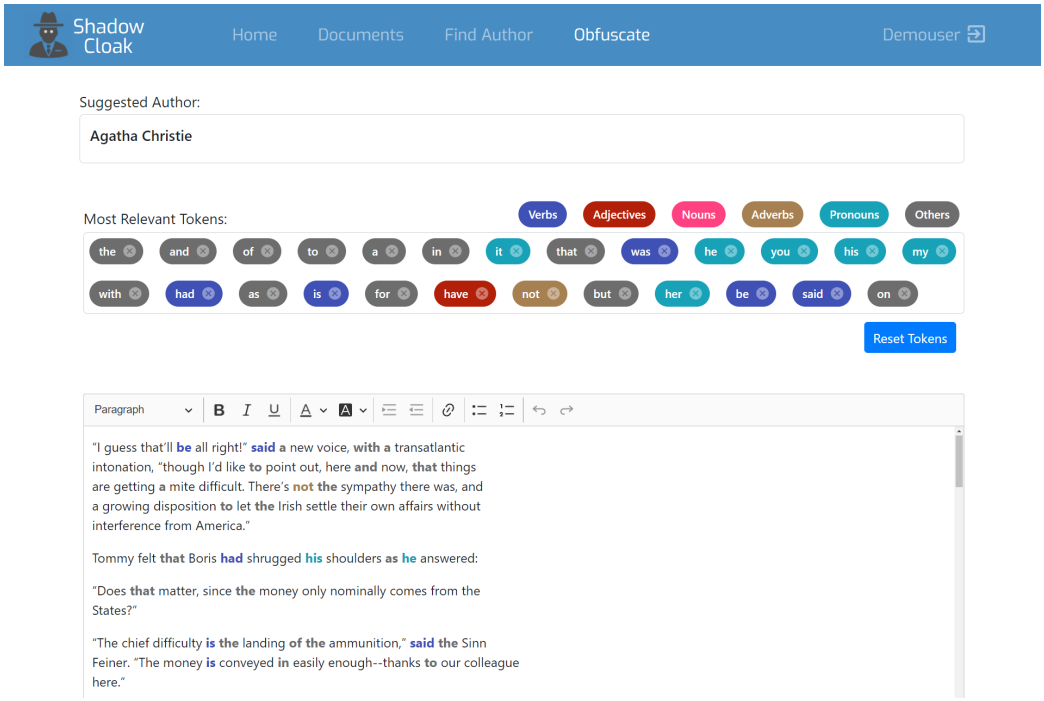

Figure 2. A screenshot from the `ShadowCloak` application, when trying to obfuscate the identification of the author.

3.2. The Case of Authors of Books

The first test scenario studies the case of identifying the authors of literary books. It is well known that the success rate that stylometry achieves in such texts is high, as there is quite a large amount of data to determine the style of each author and a large amount of data to determine the style of the disputed text; most importantly though, literary books are written in an expressive way that highly reflects the way the authors express themselves, whilst syntactic and grammatical errors are rare. In our experiment, the styles of three different authors were analysed. For each author, four books were entered into the application database to analyse their style (see Table 1), whilst four different books per author were used as anonymous texts in order to identify their authors (see Table 2).

Table 1. Books used to analyse the authors styles.

Agatha Christie	Arthur Conan Doyle	Charles Dickens
Poirot Investigates	*Adventures of Sherlock Holmes*	*A Christmas Carol*
The Man in The Brown Suit	*Tales of Terror and Mystery*	*Bleak House*
The Murder on the Links	*The Hound of the Baskervilles*	*David Copperfield*
The Mysterious Affair at Styles	*The Lost World*	*Hard Times*

Table 2. Books used to identify their author.

Agatha Christie	Arthur Conan Doyle	Charles Dickens
And Then There Were None	*The Return of Sherlock Holmes*	*Dombey and Son*
At Bertram's Hotel	*The Sign of the Four*	*The Mystery of Edwin Drood*
The Secret Adversary	*The Valley of Fear*	*The Old Curiosity Shop*
Third Girl	*The White Company*	*The Chimes*

The successful identification rate by the ShadowCloak application was 100%—i.e., a correct identification was achieved for each unknown book of the 12 that were tested in total, while the delta Scores' deviation was decisive. For the case of the Agatha Christie books, the relevant measurements with respect to the delta scores are shown in Figure 3; similarly, Figures 4 and 5 illustrate the delta scores for the books authored by Arthur Conan Doyle and Charles Dickens, respectively. Although the discovery of an unknown author of a literary book probably does not raise privacy concerns, these results illustrate the effectiveness of stylometry and it paves the way to further consider other types of texts, being smaller and less structured than literary books. Such test cases are studied next.

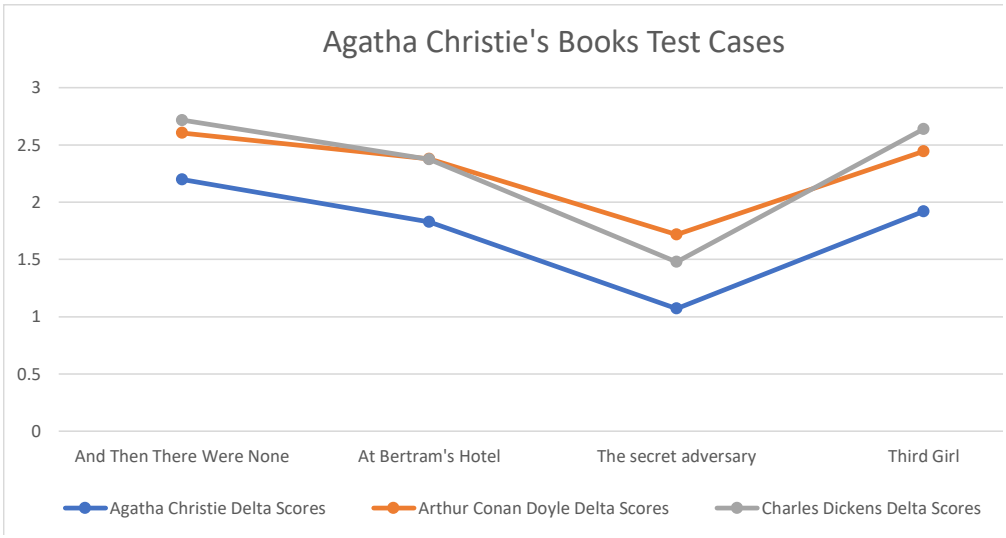

Figure 3. Delta scores, for all authors, for the books authored by Agatha Christie.

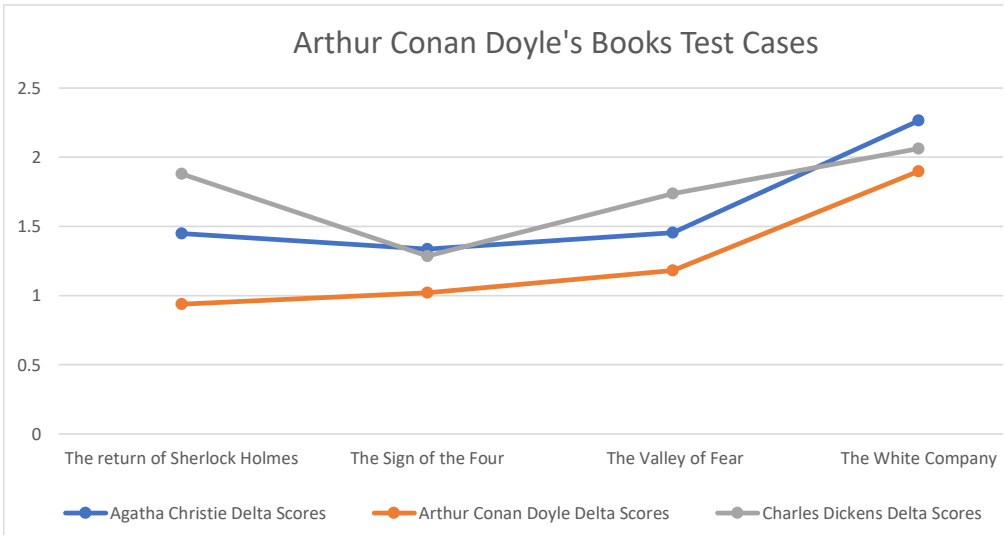

Figure 4. Delta scores, for all authors, for the books authored by Arthur Conan Doyle.

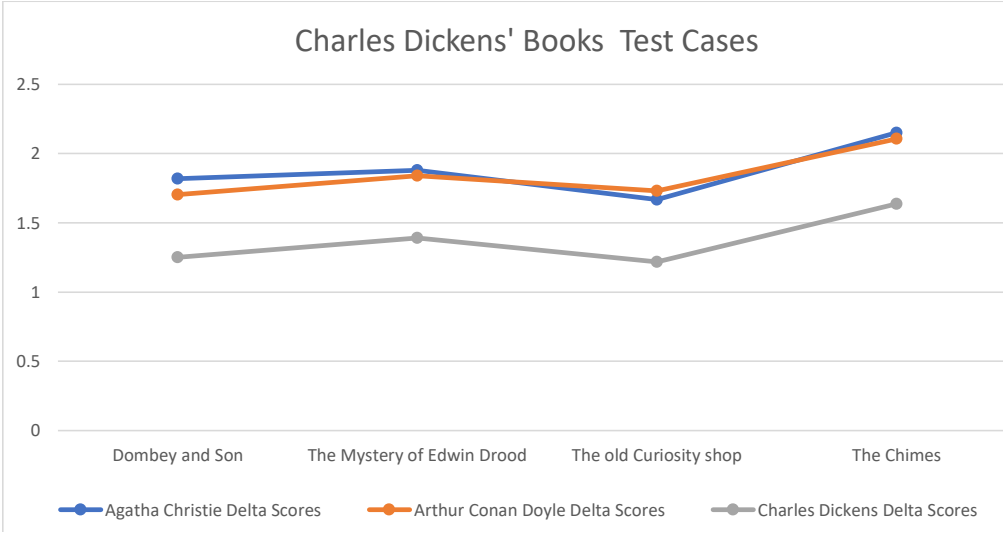

Figure 5. Delta scores, for all authors, for the books authored by Charles Dickens.

3.3. The Case of Authors of Articles in Blogs

The next test case focuses on blog posts. For our experiments, 20 articles were collected from three different columnists related to the field of information security; these articles were taken from two popular websites. The collection articles with similar topics clearly adds an extra degree of difficulty in the identification of the author—and that is why we proceeded in this way.

Figures 6–8 represent the delta scores for all texts (articles in blogs) corresponding to assumed unknown authors—each figure corresponds to the texts of a specific author (the names of the authors have been pseudonymised for this paper through replacing them with random strings and, therefore, we refer to them as the authors As25gC, 3Nb23FVAc and V2NH56A, respectively). Sixty (60) articles (blog posts) were used for the tests, consisting of twenty (20) for each of the three authors; fifteen (15) of them were used to construct the basis consisting of text from known authors so as to have their writing style, and the remaining five (5) were checked with respect to whether their author—assumed to be unknown—can be identified. The experiments illustrated, as shown in the aforementioned Figures, a success identification rate at 93.3%— actually, only for one of the blog posts we did not identify the correct author.

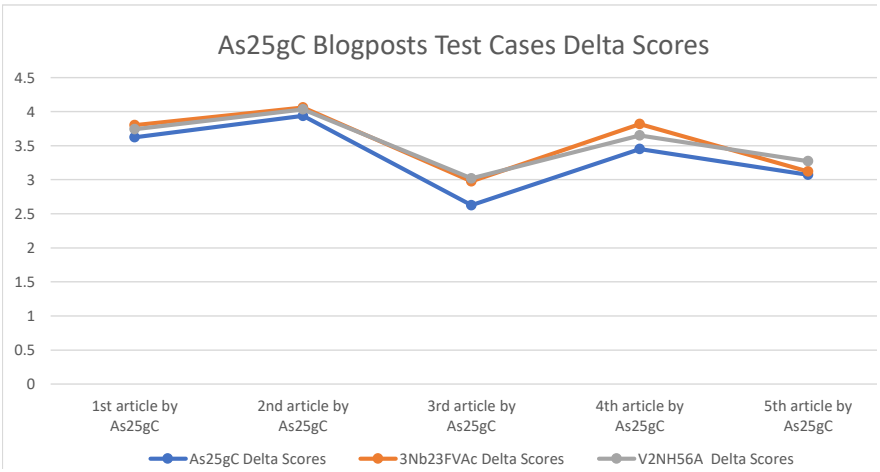

Figure 6. Delta scores, for all authors, for the articles authored by As25gC.

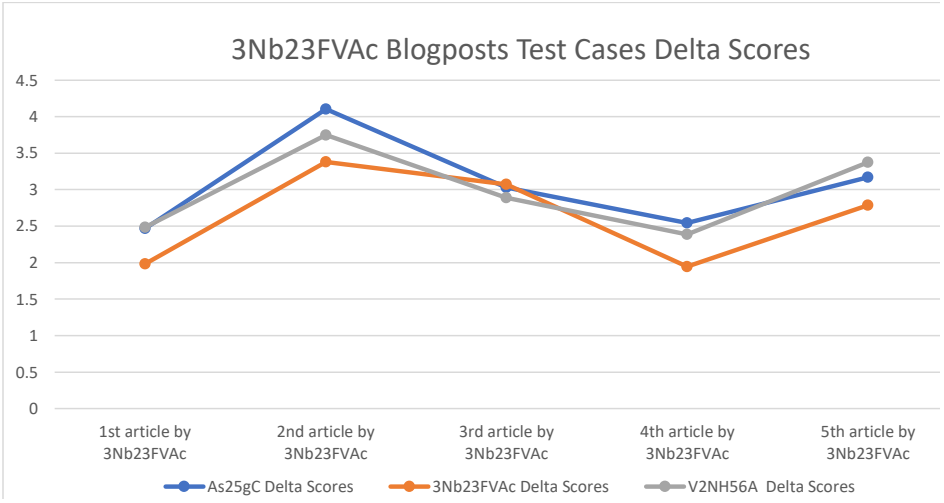

Figure 7. Delta scores, for all authors, for the articles authored by 3Nb23FVAc.

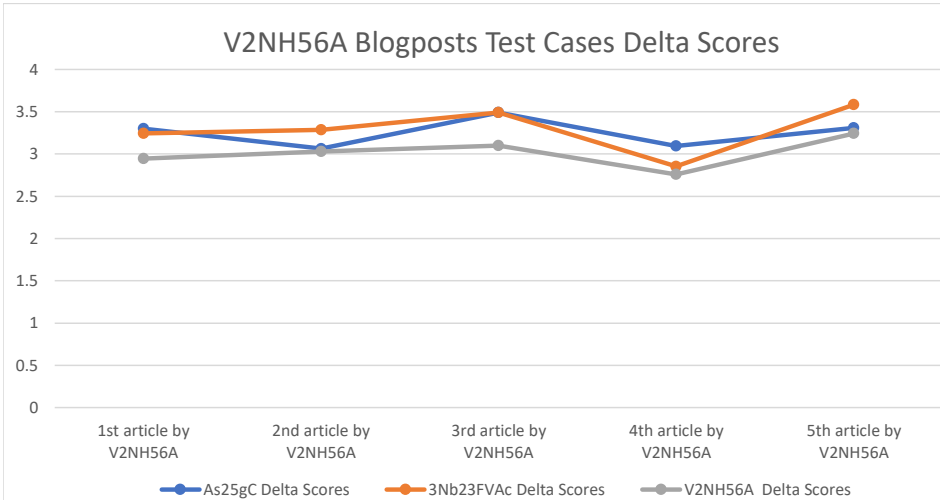

Figure 8. Delta scores, for all authors, for the articles authored by V2NH56A.

From the above results, we conclude that stylometry may be highly successful in the case of blog posts. It should be pointed out though that, by observing the values of the delta scores, they are close to each other (which was not the case for the scenario focusing on books). This may be attributed to the fact that there was a much smaller volume of known data, as well as to the fact that the texts examined had a much shorter length than those in the case of books. The selected articles were deliberately small in size, around 250–300 words so as to introduce a greater degree of difficulty. Another factor that affected the values of delta scores, leading them to divergence, is that the the topics of the articles studied were similar and very specific (related to information security) and, thus, the authors are not expected to fully exhibit their personal writing styles. Finally, since the articles were selected from two websites with similar content and vocabulary, it is not expected that differences in vocabulary have a crucial role in the differentiation of styles.

However, it can be said that stylometric identification methods may generally work very well in identifying anonymous blog posts, even when they are small in size and their content lies in a specific field with prescribed, let us say, vocabulary. It should be also pointed out that, even for the single case of failed identification of the author of the text, when this text was merged with another one written by the same author (thus constructing a larger text for authorship attribution), the identification was successful. Based on the above, it can be said that stylometry may be considered a possible threat to the privacy and personal data protection of, for example, journalists, whistleblowers, columnists, people in general who upload articles on the internet expressing their opinion freely under the assumption that they are anonymous (e.g., in anonymous social network users), etc.

3.4. The Case of Authors of E-Mails

The next case study concerns e-mails, being a medium that most people use on a daily basis. Using a subset of the Enron corpus, containing 32 authors and hundreds of emails, different scenarios were considered towards evaluating the effectiveness of stylometry with respect to the identify of the senders of e-mails. In all these scenarios, there are inherent difficulties due to nature of the texts (i.e., e-mails) considered; more precisely, emails, especially those used in internal communications between the executives of a company, do not allow much freedom of expression, whereas they are usually very small in size. Indeed, in the corpus of Enron's emails, most messages consisted of a few sentences or a few words, or even a single syllable. This makes the extraction of the mode of expression very difficult and for this reason, it is of particular interest to investigate the capabilities of the present stylometric system to identify the author.

To handle the situation of having a small number of words in the emails, a different measurement method was used. Three authors were randomly selected, for each of whom three texts were compiled. The first text contained 5 of the author's emails, the second 10 and the third 15. By this way, we sought to estimate the percentage of successful identification in relation with the size of the unknown text (number of emails). Additionally, through this testing scenario, we investigated whether increasing the size of the text suffices to overcome the aforementioned limitation with respect to the lack of a clear personal style in writing. Finally, in order to evaluate the effect of the number of known authors on the success rate of the system, the same exact same tests were performed but in a subset of the previous corpus, which contained only the three authors to whom the test texts belonged.

Hundreds of different tests, with different text corpora and sub-corpora and different sets of known authors, were performed. As an outcome, we concluded that the increase in text size does not lead to a significant increase in the identification rate; namely, from the delta Scores that were produced, it became obvious that it remained difficult to extract a clear style from the available texts, regardless of their size. At the same time, the reduction in the total number of the set of known authors from 32 to 3 had no positive effect on the calculation of the delta scores and, consequently, on the success of the stylometric system to identify the author.

Therefore, it becomes evident that the determining factor affecting the potential of identification seems to be the ability to extract a defining and recognizable style, which depends primarily on the type of text and not on their size or the number of authors. Hence, stylometry, in the form used in our stylometric system, is not sufficient to identify a person through a set of realistic professional emails, such as the Enron email dataset. However, further research is needed with additional datasets, including data of better quality from a stylometric point of view, such as personal emails sent by users outside of a business environment.

3.5. The Case of Authors of Social Media Posts

Similar to the case of professional emails (studied above) which do not have the necessary structure to allow deriving a well-determined authorship style, it is expected that, somehow, similar observations will also occur for the social media posts (tweets in our

case). In fact, in this category the texts are even shorter and "poorer" in terms of expression, reflecting the real way in which most people nowadays express themselves online. For our case study, the Twitter platform was chosen because its messages have restrictions on the maximum number of characters. In the dataset used, even though the number of authors is very large, the posts per user are very few, usually five or six. This makes extracting a clear style per author even more difficult. Nevertheless, the constraints that stem from the poor stylistic quality of the actual posts, as well as the lack in the amount of data per author, are very important when trying to extract realistic conclusions in the context of possible privacy issues.

Due to the above characteristics of our dataset, we proceeded as follows for our measurements. The first measurement aims to identify the authors of 10 tweets from 10 different authors, among a large number of authors (>100). A second measurement was also conducted, for which a different corpus was created that included only the texts of the ten authors whose texts were to be examined in terms of authorship attribution and, subsequently, the texts of each author were tested again to examine whether the number of known authors affects the performance of the stylometric system.

For the first case of identifying the author (i.e., the Twitter user) of 10 tweets amongst a hundred Twitter users, our stylometric technique correctly identified three of them—i.e., achieving a success identification rate of 30%. For the second case, which was based on a subcorpus of 10 authors, the results were a little better since the successful identification rate was increased from 30% to 50% (i.e., successful identification of the author of 5, out of 10, tweets). However, the delta scores calculated did not differ much among authors, which indicates that identification was not decisive, even if the identification rate increased due to the smaller number of potential authors.

Again, a general conclusion is that stylometry, in its somehow traditional form, as implemented in the `ShadowCloak` stylometric system, does not provide much reliability in the identification of people for small social media posts, such as tweets (unless, possibly, a very large volume of known data per user is available). Apparently, much more research is also needed to investigate this scenario (e.g., examining other datasets with larger volumes of data, not only from Twitter, but from other social networks.

4. Anonymising Texts

Having identified the potential risks with respect to the privacy and personal data protection stemming from the effectiveness of stylometry, it is natural to consider how to alleviate these issues. Since a determining factor for identifying an unknown author of a text is its writing style, which could be determined in several cases simply by the usage of specific function words, we performed—through our `ShadowCloak` system—some preliminary experiments towards examining whether the replacement of some functional words with their synonyms suffices to obfuscate the author's writing style and, thus, to render the stylometry ineffective. This analysis, although not extensive and not automated, illustrated that such an anonymisation is indeed possible for some cases.

More precisely, let us consider the case shown in Figure 9, which corresponds to an e-mail from the aforementioned Enron's corpus whose author (denoted by L. J.) was correctly identified, whilst the relevant function words are emphasised. Having these function words as a starting point, we started manually to replace them with synonyms, until our stylometric system returned a wrong output with respect to author attribution. In this case, this occurred very easily; more precisely, for this text consisting of 76 words, if we changed the last sentence from "Kay, can you please try to organise ASAP" to "Kay, would you try organizing it ASAP" (i.e., removing the words "can", "you", "please", "to" from the last sentence), the identification of the author through our technique was not possible.

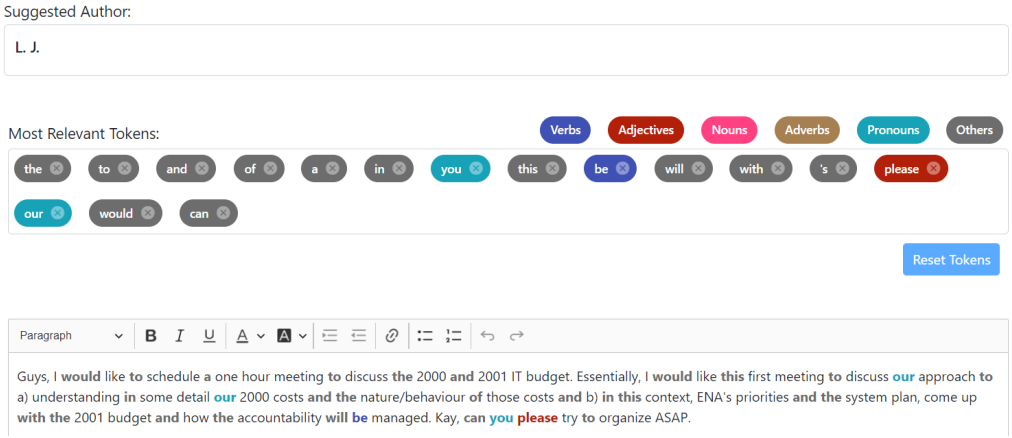

Figure 9. An e-mail whose author was correctly identified, according to specific function words.

Although such preliminary examples do not allow deriving sound conclusions, they clearly indicate that it is of importance to further examine this field—i.e., how to hide identities of authors of texts, under the assumption that effective stylometric techniques can be applied.

5. Future Work

This work opens many directions for further research. Indeed, despite the fact that this research put effort on compiling realistic data of different categories to perform a large number of measurements for various scenarios, our tests are clearly far from exhaustive, and several other parameters should be further investigated.

First, since this paper is based on the frequency of use of the most common function words as a metric of the author's style, it is evident that other metrics should be also examined. Such metrics could be the length of sentences, paragraphs or texts, the capitalisation, the spelling or grammatical errors, the structure of the text in terms of the alignment and spaces, the frequency of use of special characters such as hashtags, reference symbols "@", or even combinations of the above. Such a stylometric system would probably be more efficient than the one described in our work for the types of texts, such as social media posts and email, that were studied in this work.

Moreover, apart from the quantification of the writing style, there is also room for experimentation with different algorithms for document classification. The selection of the Burrows delta algorithm in the ShadowCloak system was based on its satisfactory performance, even for small sets of input data. However, there are other algorithms based on machine learning methods whose capabilities in this context should be explored. It is highly possible that the rapid development of machine learning methods will enhance stylometric analysis by providing more efficient algorithms than the existing ones, making stylometry an even more useful method in the future. Of course, a stylometric system could be based on more than one algorithm, a case that could also be investigated.

In addition, for the cases that the effectiveness of the stylometry seems to have room for improvement, it is of interest to consider—regardless of the underlying chosen metric—additional datasets and corpora for the analysis so as to have more data per author. The question that naturally arises is how much data are needed for each case; this is highly related with the evaluation of the privacy risks for each case (for example, if a huge volume of data is needed, then stylometry may not constitute a threat).

Finally, a very interesting ground for research with important practical applications is related to the concept of anonymisation, or otherwise the avoidance of identification through stylometric methods. A first attempt was already made in the present paper, with a gradual replacement of the most important function words, illustrating that such techniques could indeed be effective. In this regard, an approach of automatic text transformation into an "anonymous" version, through replacing function words with synonyms, multiple

translations of the text or other techniques, would be very interesting. This could possibly result in an anonymisation tool that could be used to address privacy threats, especially by individuals whose privacy is directly threatened and whose protection is particularly critical, such as journalists, whistle blowers, political dissidents and others. To this goal, the proper exploitation of machine learning techniques is an interesting direction to be explored.

6. Conclusions

This paper studied stylometry from a personal data protection point of view, i.e., it examined whether it is feasible—and under which prerequisites—to find out the author of a text who is assumed to be anonymous. Stylometry actually identifies the writing style of known authors and subsequently aims to check, for a given text from an unknown author, whether the writing style reflects the identity of the author. This could be highly intrusive in terms of privacy and personal data protection if anonymity needs to be ensured (e.g., in whistleblowing systems). A large number of experiments were performed based on a research-oriented software that was developed for our study, illustrating that there exist cases in which stylometry can be very effective. These cases mainly include texts of a particularly large length, including articles in blogs. However, it seems that some limitations exist in the effectiveness of stylometry for small texts, such as tweets or professional e-mails, when the stylometric technique is solely based on the identification of function words. However, much research is still needed so as to examine even more sophisticated stylometric techniques, as well as a wider pool of texts with several types of structures.

A main outcome of the above analysis is that, when there is a need to anonymise personal data related to texts of various types (for fulfilling data protection and privacy requirements), the stylometry as a threat should be also taken into account, under a risk-based approach—i.e., one should examine whether an actual risk of re-identification does occur due to the existence of stylometric techniques. To this end, questions such as how much data are available for analysis or whether the type of text reflects a specific writing style of the author should be addressed at an early stage. Clearly, there is much research still to be done on this aspect; it is also of importance to consider finding systematic ways for text anonymisation, under the assumption that re-identification attacks based on stylometry should be avoided.

Author Contributions: Conceptualization, A.P.; methodology, A.P.; software, A.P.; writing—original draft preparation, A.P. and K.L.; supervision, K.L.; project administration, K.L. All authors have read and agreed to the published version of the manuscript. .

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available in the article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
DARPA	Defence Advanced Research Projects Agency
EU	European Union
FBI	Federal Bureau of Investigation
GDPR	General Data Protection Regulation
NLTK	Natural Language Toolkit

References

1. Fung, B.C.M.; Wang, K.; Chen, R.; Yu, P.S. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.* **2010**, *42*, 14:1–14:53. [CrossRef]
2. Chatzistefanou, V.; Limniotis, K. On the (Non-)anonymity of Anonymous Social Networks. In *E-Democracy—Privacy-Preserving, Secure, Intelligent E-Government Services, Proceedings of the 7th International Conference, E-Democracy 2017, Athens, Greece, 14–15 December 2017, Proceedings*; Katsikas, S.K., Zorkadis, V., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; Volume 792, Communications in Computer and Information Science, pp. 153–168.
3. Finck, M.; Pallas, F. They who must not be identified—Distinguishing personal from non-personal data under the GDPR. *Int. Data Privacy Law* **2020**, *10*, 11–36. [CrossRef]
4. Narayanan, A.; Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (sp 2008), Oakland, CA, USA, 18–21 May 2008; pp. 111–125.
5. Calandrino, J.A.; Kilzer, A.; Narayanan, A.; Felten, E.W.; Shmatikov, V. "You Might Also Like:" Privacy Risks of Collaborative Filtering. In Proceedings of the 32nd IEEE Symposium on Security and Privacy, S&P 2011, Berkeley, CA, USA, 22–25 May 2011; IEEE Computer Society: Washington, DC, USA, 2011; pp. 231–246.
6. Soria-Comas, J.; Domingo-Ferrer, J. Big Data Privacy: Challenges to Privacy Principles and Models. *Data Sci. Eng.* **2016**, *1*, 21–28. [CrossRef]
7. Carlini, N.; Liu, C.; Erlingsson, Ú.; Kos, J.; Song, D. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In Proceedings of the 28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, 14–16 August 2019; USENIX Association: Berkeley, CA, USA, 2019; pp. 267–284.
8. Daelemans, W. Explanation in Computational Stylometry. In *Computational Linguistics and Intelligent Text Processing, Proceedings of the 14th International Conference, CICLing 2013, Samos, Greece, 24–30 March 2013*; Proceedings, Part II; Lecture Notes in Computer Science; Gelbukh, A.F., Ed.; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7817, pp. 451–462.
9. Brennan, M.; Afroz, S.; Greenstadt, R. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur.* **2012**, *15*, 12:1–12:22. [CrossRef]
10. Vosoughi, S.; Zhou, H.; Roy, D. Digital Stylometry: Linking Profiles Across Social Networks. In *Social Informatics, Proceedings of the 7th International Conference, SocInfo 2015, Beijing, China, 9–12 December 2015*; Lecture Notes in Computer Science; Liu, T., Scollon, C.N., Zhu, W., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; Volume 9471, pp. 164–177.
11. Davis, R.C. Obfuscating Authorship: Results of a User Study on Nondescript, a Digital Privacy Tool. 2019. Available online: https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1273&context=jj_pubs (accessed on 31 January 2021).
12. Stolerman, A.; Overdorf, R.; Afroz, S.; Greenstadt, R. Breaking the Closed-World Assumption in Stylometric Authorship Attribution. In *Advances in Digital Forensics X, Proceedings of the 10th IFIP WG 11.9 International Conference, Vienna, Austria, 8–10 January 2014*; Revised Selected Papers; IFIP Advances in Information and Communication Technology; Peterson, G.L., Shenoi, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 433, pp. 185–205.
13. Gómez-Adorno, H.; Ríos-Toledo, G.; Posadas-Durán, J.P.; Sidorov, G.; Sierra, G. Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts. *Comput. Sist.* **2018**, *22*, 47–53. [CrossRef]
14. Piasecki, M.; Walkowiak, T.; Eder, M. Open Stylometric System WebSty: Integrated Language Processing, Analysis and Visualisation. *Comput. Methods Sci. Technol.* **2018**, *24*, 43–58. [CrossRef]
15. Grivas, A.; Krithara, A.; Giannakopoulos, G. Author Profiling using Stylometric and Structural Feature Groupings. In Proceedings of the Working Notes of CLEF 2015—Conference and Labs of the Evaluation Forum, Toulouse, France, 8–11 September 2015; Cappellato, L., Ferro, N., Jones, G.J.F., SanJuan, E., Eds.; CEUR-WS.org: Aachen, Germany 2015; Volume 1391.
16. Mikros, G.K.; Perifanos, K. Gender Identification in Modern Greek Tweets. In *Recent Contributions to Quantitative Linguistics*; Quantitative Linguistics; Tuzzi, A., Benesová, M., Macutek, J., Eds.; DeGruyter: Berlin, Germany, 2015; Volume 70, pp. 75–88.
17. de Pablo, Á.; Araque, O.; Iglesias, C.A. Radical Text Detection based on Stylometry. In Proceedings of the 6th International Conference on Information Systems Security and Privacy, ICISSP 2020, Valletta, Malta, 25–27 February 2020; Furnell, S., Mori, P., Weippl, E.R., Camp, O., Eds.; SCITEPRESS: Setúbal, Portugal, 2020; pp. 524–531.
18. Schuster, T.; Schuster, R.; Shah, D.J.; Barzilay, R. The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Comput. Linguist.* **2020**, *46*, 499–510. [CrossRef]
19. Sadman, N.; Datta Gupta, K.; Haque, M.A.; Sen, S.; Poudyal, S. Stylometry as a Reliable Method for Fallback Authentication. In Proceedings of the 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Phuket, Thailand, 24–27 June 2020; pp. 660–664.
20. Chatzakou, D.; Soler Company, J.; Tsikrika, T.; Wanner, L.; Vrochidis, S.; Kompatsiaris, I. User Identity Linkage in Social Media Using Linguistic and Social Interaction Features. In Proceedings of the WebSci'20: 12th ACM Conference on Web Science, Southampton, UK, 6–10 July 2020; Ferrara, E., Leonard, P., Hall, W., Eds.; ACM: New York, NY, USA, 2020; pp. 295–304.
21. Evert, S.; Proisl, T.; Jannidis, F.; Reger, I.; Pielström, S.; Schöch, C.; Vitt, T. Understanding and explaining Delta measures for authorship attribution. *Digit. Scholarsh. Humanit.* **2017**, *32*, ii4–ii16. [CrossRef]
22. Plechác, P. Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns. *Digit. Scholarsh. Humanit.* **2021**, *36*, 430–438. [CrossRef]
23. Burrows, J. 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. *Lit. Linguist. Comput.* **2002**, *17*, 267–287. [CrossRef]

24. Zipf, G.K. *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*; Addison-Wesley: Boston, MA, USA, 1949.
25. Project Gutenberg. Available online: www.gutenberg.org (accessed on 20 February 2022).
26. Enron Email Dataset. Available online: https://www.cs.cmu.edu/~enron/ (accessed on 20 February 2022).
27. Twitter US Airline Sentiment364 (TUAS) Dataset. Available online: https://www.kaggle.com/crowdflower/twitter-airline-sentiment (accessed on 20 February 2022).