Textwash - automated open-source text anonymisation

Bennett Kleinberg* Toby Davies Maximilian Mozes

Version: 27 August, 2022

Abstract

The increased use of text data in social science research has benefited from easy-to-access data (e.g., Twitter). That trend comes at the cost of research requiring sensitive but hard-to-share data (e.g., interview data, police reports, electronic health records). We introduce a solution to that stalemate with the open-source text anonymisation software *Textwash*. This paper presents the empirical evaluation of the tool using the TILD criteria: a technical evaluation (how accurate is the tool?), an information loss evaluation (how much information is lost in the anonymisation process?) and a de-anonymisation test (can humans identify individuals from anonymised text data?). The findings suggest that Textwash performs similar to state-of-the-art entity recognition models and introduces a negligible information loss of 0.84%. For the de-anonymisation test, we tasked humans to identify individuals by name from a dataset of crowdsourced person descriptions of very famous, semi-famous and non-existing individuals. The de-anonymisation rate ranged from 1.01-2.01% for the realistic use cases of the tool. We replicated the findings in a second study and concluded that Textwash succeeds in removing potentially sensitive information that renders detailed person descriptions practically anonymous.

1 Introduction

With the increasing digitisation of society and human communication, text data are becoming more important for research in the social and behavioural sciences (Gentzkow, Kelly, and Taddy 2019; Salganik 2019). Advances made in natural language processing (NLP) in particular have led to exciting insights derived from text data (e.g., on emotional responses to the pandemic (Kleinberg, Vegt, and Mozes 2020) or on the rhetoric around immigration in political speeches (Card et al. 2022); for an overview, see (Boyd and Schwartz 2021)). Importantly, the use of computational techniques to quantify and analyse text data has triggered a demand, especially for large datasets (often of several tens of thousands of documents) that can be harnessed for machine learning approaches (e.g., (Socher et al. 2013; Lewis et al. 2020)). That status quo of a need for larger datasets and an appetite to use text data for the study of social science phenomena has resulted in a dilemma: many of the important questions require targeted, primary data collection or access to potentially sensitive data. However, such data are hard to obtain, not because they do not exist but because sharing them is constrained by data protection regulations and ethical concerns. One potential consequence is that research activity may be biased toward topics for which suitable data is more readily available rather than those most important.

One of the few viable solutions to this dilemma is automated text anonymisation; that is, the large-scale processing of text data so that individuals cannot be identified from the resulting output. Such a method would allow for the flow of sensitive data so that the staggering potential of text data can be exploited for scientific progress. With this paper and the tool it introduces, we seek to enable researchers to work with such sensitive data in a way that protects the privacy of individuals whilst retaining the usefulness of anonymised data for computational text analysis.

*bennett.kleinberg@tilburguniversity.edu

1.1 Anonymising text data

Text anonymisation refers to redacting - and potentially replacing - personally identifiable information (PII) within text data. Since such information is the crucial concern of data protection regulations – data is protected if and only if it can be associated with a living individual – its removal means that such data can be freely shared. Text anonymisation aims to facilitate the sharing of data whilst protecting the identities of individuals that are the subject of text data. Text anonymisation comes with two critical challenges.

First, for subsequent content-based text analyses, an anonymisation tool should anonymise text so that the anonymised text remains readable, meaningful, and useful for syntax- and content-based analyses after anonymisation (i.e. it should preserve the semantic context of the text). Consider, for example, the sentence "Joe Biden is the current president of the United States." If we replace all PII with a generic term (e.g., XXX), we obtain the sentence "XXX XXX is the current XXX of the XXX XXX," rendering the sentence semantically meaningless and context-free. Any subsequent text analysis would fail to capture the context, semantic roles or other relevant linguistic constructs.

Second, evaluating the performance of text anonymisation tools is challenging. Even if we assume an anonymisation tool to identify almost all PII, just a few unidentified sensitive words can reveal an individual's identity and jeopardise the whole anonymisation procedure. For example, if the sequence above were anonymised as "XXX XXX is the current president of the United States," the reader would be able to infer with an educated guess that the person mentioned in the text is Joe Biden.

In an attempt to pave the way for privacy-preserving data sharing, in this paper, we introduce *Textwash*, an open-source text anonymisation tool that anonymises text data in a fully automated way. Textwash anonymises texts in a semantically-meaningful manner, ensuring that the anonymised documents remain usable for downstream text analyses with respect to a document's syntactic properties and content. The tool achieves this in a two-stage process, consisting of i) the automatic identification of relevant information from input documents and ii) the subsequent replacement of the detected information with meaning-preserving tokens. Rather than replacing PII with XXX, we use category-specific replacement tokens for each identified word. For example, Joe would be replaced with [firstname1], Biden with [lastname1], president with [occupation1], and United States with [location1], resulting in "[firstname1] [lastname1] is the current [occupation1] of the [location1]."

Textwash is based on supervised machine learning techniques, leveraging pre-trained contextualised word representations as provided by the BERT language model (Devlin et al. 2018). In order to train a model capable of automatically identifying relevant information in the input text, we first annotate a large corpus of text data sourced from the British National Corpus (BNC; Consortium and others (2007)) as well as the Enron email dataset.[1] We then fine-tune a pre-trained BERT model on the collected data.

1.2 Existing approaches to text anonymisation

Machine learning-based text anonymisation approaches have been proposed in various languages. Mamede, Baptista, and Dias (2016), for example, propose an approach based on named entity recognition (NER) and coreference resolution to anonymise texts in Portuguese. Their model provides three modes of anonymisation: suppression (i.e., each PII is replaced with a generic token such as XXX), tagging (i.e., each PII is replaced with a category-specific and indexed token, such as [***ORGANIZATION123***]), and random substitution (i.e., each PII is replaced with a random PII of the same category). The authors evaluate the performance of their model with respect to an automated precision- and recall-based method as well as through human evaluation. However, in contrast to Textwash, their proposed model only focuses on the entities person, location, and organisation, and hence might miss crucial PII corresponding to other categories, as we show in this paper.

Elsewhere, NETANOS (named entity-based text anonymization for open science) has been proposed as an anonymisation tool for the English language (Kleinberg and Mozes 2017). NETANOS utilises an available named entity recognition tool, the Stanford Named Entity Tagger (Finkel, Grenager, and Manning 2005), to

[1] https://www.cs.cmu.edu/~enron/

identify PII from input text data. Unlike Textwash, NETANOS does not require annotated training data but, as a consequence, only identifies persons, locations, organisations, and dates.

More recently, Francopoulo and Schaub (2020) approached text anonymisation from the context of customer relationship management (CRM). The authors suggest a method comprising an NER-based module, an entity linker, and a substitution module evaluated on a collection of French legal and administrative documents. In contrast to Textwash, an off-the-shelf NER module, Tagparser (Francopoulo 2007), is used.

Adams et al. (2019) proposed AnonyMate, which classifies identifiable information as either PII or CII (corporate identifiable information) and removes these. Their method is based on an analysis of historical chat logs, from which 24 entity types of interest were extracted. Based on these entities, the authors propose an annotated NER data set comprising the six languages English, German, Swedish, Spanish, Italian, and Swedish. They then trained two NER models, one based on conditional random fields (CRF) and one based on recurrent neural networks, and combined the NER model with a coreference resolution model. The proposed approach is evaluated against automatic performance metrics (precision, recall, F1-score), leaving the question of how well their method would work when tested against a human benchmark.

While the tools above rely on identifying to-be-removed words, Hassan et al. (2019) devised an anonymisation method based on word embeddings. Their approach, however, is limited in that documents can only be anonymised if a specific entity is to be removed, and hence does not generalise to the task described in this paper. Various related methods have been proposed recently (Di Cerbo and Trabelsi 2018; Berg, Chomutare, and Dalianis 2019; Romanov et al. 2019).

The tool that comes closest to Textwash's aim is scrubadub – a Python package using various existing off-the-shelf packages (e.g., spaCy, Stanford NER detector) to detect PII in text, albeit with fewer categories than Textwash and without retaining context between replaced entities. Unfortunately, similar to AnonyMate, the tool is not evaluated against de-anonymisation or information loss, leaving its usefulness for data-sharing activities unclear.

In addition to these published examples, some commercial tools claim to perform text anonymisation. However, these tools typically do not provide empirical evidence of their performance and are closed-source. This lack of transparency and the absence of evidence of their validity essentially disqualifies them from scientific applications.

The present paper provides a novel perspective on text anonymisation that addresses one or several weaknesses of previous approaches. Most immediately, we evaluate Textwash according to rigorous empirical criteria that surpass those used in previous work: as well as measuring its technical performance, we evaluate its performance on the core task of anonymisation via a realistic scenario involving human participants. In addition, the software is developed and tested[2] Specifically, with a focus on scientific research purposes: Textwash adheres to the principles of open science (i.e., open source code, transparent processes, free non-commercial use) and is usable without an internet connection (i.e., not imposing any vulnerabilities when processing sensitive data).

1.3 Aims

This paper has two aims: introducing the Textwash tool for automated text anonymisation and providing an empirical evaluation of the tool. We first detail the software and then report two validation studies that put the tool to various tests.

1.4 Transparency statement

The Textwash software is an open-source Python project and is available and documented at https://github.com/maximilianmozes/textwash. All data collected for the reported studies and the material used are publicly available in that repository. All participants in the studies where data were collected provided informed consent, and the procedures were approved by the IRB at University College London.

[2]Note that some commercial tools exist that fail to provide any information about the validity of their software and the process of building it and thereby disqualify from an application for scientific purposes.

2 The Textwash tool

2.1 Requirements for modern text anonymisation

For the development of Textwash, we identified a set of requirements enumerated in this section.

- A text anonymisation tool operating on sensitive data should be usable offline, on a regular computer (e.g., laptop), without sending any data through external APIs for processing and anonymisation. This is to ensure that Textwash adheres to potential privacy regulations that users have in place and to remove any requirement to trust a third party.
- Anonymised text data need to retain value for secondary analysis in NLP (e.g., topic modelling, sentiment analysis or coreference resolution) since one of the main goals of this software is to provide researchers and practitioners with a tool to share anonymised datasets for further data analysis.
- A text anonymisation tool should be generalisable to new contexts and domains. While rule-based approaches might be capable of anonymising textual documents corresponding to a specific topic, Textwash aims to be as adaptable as possible and should be able to anonymise out-of-domain texts. In order to do this, Textwash is based on machine learning-based methods and uses linguistic patterns extracted from contextual information to predict whether individual words and phrases contain sensitive information.
- For the software to be trustworthy and auditable, its mechanisms should be transparent, and for users to fully exhaust its capabilities, it should be customisable. The software should therefore be open source and available for public use.
- A tool that deals with the sensitive context of anonymisation needs to be validated appropriately using empirical experiments. These should go beyond the automated metrics typically used and should involve testing by a set of human judges.

2.2 Potentially sensitive information

The often-used definition of personally identifiable information rests on the assumption that a few pre-defined categories capture all that could lead to the identification of an individual. But, as this paper will show, there are pieces of information that - in combination with other information - could reveal an identity even if they would not count as personally identifiable information in themselves. These include references to hyper-specific attributes of a person (e.g., a mention of a specific pronunciation error that someone makes) and context-dependent sensitive information (e.g., mentioning that an actor performed in the last film of a famous series). Put differently: the information that can reveal an individual's identity is not limited to categories such as dates, names and locations. Therefore, we propose a concept encompassing the full spectrum of information that could reveal an identity: **potentially sensitive information (PSI)**. PSI is hereafter used to describe any piece of information that could directly or indirectly (e.g., through the combination with other information) be useful to identify an individual.

With that broadening of identification risk in mind, the task for a text anonymisation system can be summarised as identifying PSI in an input sequence and replacing that information with generic terms that do not reveal any information about individuals mentioned in the text.

2.3 Identifying potentially sensitive information

The first problem of identifying PSI input words can be described as a named entity recognition (NER) task using supervised learning techniques. Given a corpus of text sequences where PSI words are annotated as such, we can train a machine learning model to classify tokens in input sequences according to whether they represent PSI or not.

Textwash realises this by utilising pre-trained contextualised word representations using BERT. Specifically, we fine-tune pre-trained BERT representations on an annotated dataset using the BertForTokenClassification module from the HuggingFace library (Wolf et al. 2019).

2.4 Replacing identifiable tokens

Once we have identified PSI, we use a rule-based approach to replace them with generic terms. To do this, we enumerate instances of a specific class and replace all occurrences of that instance in the text with a generic identifier representing the class to which the sensitive phrase belongs. For example, assuming we identify the word "London" as "LOCATION" in the text, we replace it with the term "LOCATION_N." Here, N is a number that uniquely identifies "London" from all other identified LOCATION phrases and ensures that subsequent mentions of London in the same document are also mapped to the identical replacement (e.g., multiple mentions of London in one document all become LOCATION_1).

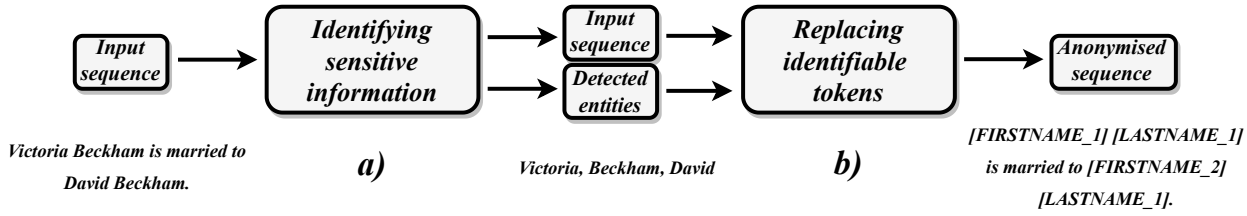

Figure 1: Illustration of the end-to-end anonymisation process of Textwash.

Figure 1 shows how, given the input sequence "Victoria Beckham is married to David Beckham," Textwash identifies sensitive information (a). The tool then inputs the original sequence and the detected entities into the second module, which replaces the identifiable tokens in a meaning-preserving way (b).

2.5 Dataset

The dataset used to train Textwash consists of 3,717 articles where each phrase (i.e., single or multiple words in a sequence) is annotated according to whether it represents PSI. It contains 417 articles from the British National Corpus (BNC; Consortium and others (2007)), 1,800 emails (email body only) from the Enron email dataset (https://www.cs.cmu.edu/~enron/) and 1,500 Wikipedia articles. For the latter, we randomly sampled Wikipedia articles about persons that contain at least 100 words. We excluded documents with less than 20 words from voth datasets. For the Enron dataset, we furthermore excluded documents with more than 500 words, and for the BNC, we truncated all documents to the first 500 words. We then sampled documents from both datasets according to the highest named entity ratios.[3] However, for Enron, we only select documents with a named entity ratio of less than 20% since we observed that otherwise a substantial proportion of emails consist of simple lists of names rather than continuous text. Choosing a high named entity ratio ensured that the documents were sufficiently rich in PSI to allow the machine learning model to learn to identify its various categories.

Two domain experts annotated the dataset with the following tags:

- PERSON_FIRSTNAME: a person's first name (e.g., Jane)
- PERSON_LASTNAME: a person's last name (e.g., Doe)
- OCCUPATION: an occupation (e.g., nurse, carpenter)
- LOCATION: a location (e.g., London, Berlin, France)
- TIME: a time (e.g., 12 pm, afternoon)
- ORGANIZATION: an organisation (e.g., Google, NHS)
- DATE: a reference to a specific day (e.g., 12/10/2021, yesterday)
- ADDRESS: an address (e.g., 42 London Road)
- PHONE_NUMBER: a phone number
- EMAIL_ADDRESS: an email address
- OTHER_IDENTIFYING_ATTRIBUTE: an identifying attribute that cannot be categorised into the above but is still considered PSI
- NONE: all other tokens in the input sequence

[3]The named entity ratio was the number of named entities - irrespective of their category - divided by the number of words in the document.

To train Textwash, we randomly split the dataset into a training, validation and test set by using 80% for training and 10% each for validation and testing.

The following sections detail two studies on the empirical validation of the Textwash tool, including automated performance metrics and human benchmarking.

3 Study 1

3.1 Method

We evaluate Textwash using the TILD criteria (Mozes and Kleinberg 2021). These go beyond a technical evaluation (i.e., how many entities are correctly identified per category) and include a test of the information loss and a human de-anonymisation test. Information loss refers to the difference in text analysis outcomes attributable to the anonymisation procedure and is further divided into utility loss (i.e., the difference between original and anonymised texts in prediction tasks) and construct loss (i.e., the difference in linguistic variables or text statistics between original and anonymised version). On the other hand, the de-anonymisation evaluation examines whether an individual's identity is leaking from anonymised text data. Since that test is hard to conduct automatically, we adopt the 'motivated intruder' principle and task human participants with the de-anonymisation of text data.

3.1.1 Technical evaluation

Desirable for a good anonymisation tool is a high detection accuracy on the sub-categories of phrases that are identified for replacement in a subsequent step (here: first names, locations, etc.). We report the detection results of anonymisation categories on the test set (i.e., unseen data). Each token in the test set's input sequences is labelled according to the aforementioned categories, and we assess the model's performance in predicting the respective label for each word. Specifically, we report the precision, recall and F1-score for each category, as well as macro and weighted averages.[4]

3.1.2 Information loss

The two sub-categories of information loss are utility loss (i.e., the difference in some performance metric on a classification task between original and anonymised data) and construct loss (i.e., the difference in variables between original and anonymised data). Higher loss values imply a decreased performance (utility loss) or variable value (construct loss) in the anonymised text compared to its original counterparts. Small loss values are desirable and would suggest that the anonymisation procedure retains the usefulness of the text data for downstream analyses.

3.1.2.1 Utility loss For the utility loss, we investigated to what extent the anonymisation of text data affected text classification performance. We used the IMDb movie reviews dataset (Maas et al. 2011), a dataset widely used in NLP research for sentiment analysis. The dataset consists of 50,000 movie reviews, each annotated with a positive or negative sentiment label. For training and testing, the dataset has a pre-defined split of 25,000 samples (balanced across sentiment).

RoBERTa is a neural network-based machine learning model based on the Transformer architecture (Vaswani et al. 2017), widely used in NLP research (Xia, Wu, and Van Durme 2020). The model has been pre-trained on large corpora of text in a self-supervised fashion by learning to predict masked tokens in textual input sequences. At the model's core lies an attention mechanism capable of learning to capture words in different contexts. The pre-trained layers of the RoBERTa model are then fine-tuned for a specific downstream task, which is classification in our case. To evaluate the utility loss in classification results, we first used the original, non-anonymised data and fine-tuned a pretrained RoBERTa model (Liu et al. 2019) on the training set, holding out 1,000 randomly selected training set sequences for validation. After training, we tested the model

[4]The macro accuracy expresses the average irrespective of each class's support (i.e., how many occurrences does this category have in the test set?), whereas the weighted accuracy weighs the score for each category based on its support.

on the test set to measure the performance on unseen data. We report the model accuracy in predicting the correct sentiment label for each test set sequence.

We then anonymised the entire dataset using Textwash and repeated the procedure from above (training, validation, testing). The utility loss is the difference in model prediction accuracy on the test set before and after anonymisation.

3.1.2.2 Construct loss Since researchers may also be interested in linguistic variables, we tested how much the values obtained from anonymised text data deviated from those derived from the original text data. The construct loss was assessed by testing whether the frequencies of part-of-speech tags differed between original and anonymised texts[5]. Since we wish to quantify statistical evidence of the absence of a difference (i.e., supporting the assertion that both texts result in the same values), we use Bayesian hypothesis testing for this (Rouder et al. 2009)[6] We report the Bayes factor in either direction: BF_{01} for evidence for the null and BF_{10} for evidence for the alternative hypothesis.

3.1.3 De-anonymisation

As the litmus test for any anonymisation tool, we investigated how well human participants can identify individuals from anonymised text data. In these terms, a tool's good performance corresponds to low re-identification rates. From a data protection point of view, the ability to sufficiently anonymise data is essential for the tool's usefulness and supersedes the other criteria. We operationalised this assessment through a motivated intruder test.

We first instructed a group of participants to write descriptions of persons similar in style to the introductory paragraphs of a Wikipedia entry.[7] In order to submit the anonymisation tool to tests of varying degrees of difficulty, we elicited person descriptions of very famous, semifamous and non-existing individuals. The rationale behind that decision was to provide a complete picture of what the tool can achieve in different contexts.

Very famous individuals are the most challenging test for an anonymisation tool because de-anonymisation can often hinge on specific and seemingly-innocuous details that reveal an identity (i.e., the range of PSI is exceptionally broad). For example - and as shown further below - the UK singer Adele could be easily identified through mentions of "singer" (for which she is famous) and "weight loss" (which was widely discussed in her case). Notably, the identification of Adele in this case only works because this privileged information is in the public domain and a human de-anonymiser could combine these pieces of information to make an educated guess. As such, very famous persons may be identifiable through properties that no longer meet the definitions of anonymisation as set out elsewhere (Information Commissioner's Office 2012) and are somewhat removed from research use cases (e.g., anonymising interview data from participants who are not widely known to the general public). Thus, very famous people present the lower bounds of what the tool can achieve (i.e., less famous persons would be more anonymisable).

At the other end of the difficulty spectrum, we use person descriptions for individuals who do not exist. These persons cannot be Googled or searched for in any other capacity, so identifying them by name is only possible if the name itself has leaked. In between these two extremes (very famous celebrities and non-existing persons), we use "semifamous" individuals who all have a Wikipedia entry but are not well known generally. For applications of the Textwash tool on research or business data, the performance on semifamous and non-existing persons is thus the most meaningful benchmark.

3.1.3.1 Gathering person descriptions

[5]The frequency of part-of-speech tags was extracted with the *spacyr* package (Benoit and Matsuo 2017) using the Universal POS tags from https://universaldependencies.org/u/pos/

[6]Bayesian t-tests allow us to quantify evidence for the null hypothesis (i.e., original = anonymised) as well as for an alternative hypothesis (i.e., original > anonymised).

[7]We did not use actual Wikipedia entries because the sentence structure - even with anonymisation - could be used in search engines and quickly lead to the source text. This would not represent de-anonymisation in a meaningful sense.

- Task: After providing informed consent, the participants were instructed to write a person description of at least 500 characters for three individuals. They wrote (in that order) one description of a very famous person, one of a semifamous person, and one of a non-existing, fictitious person. The person subjects were chosen at random from a pre-defined item pool. The data collection was done in Qualtrics.
- Item pool: The following items were used as pointers for the participants - each category consisted of 10 items:
  - Very famous persons: the participants were presented with the name of a very well-known person in the UK (e.g., Emma Watson, Benedict Cumberbatch).
  - Semifamous persons: the participants were presented with the name of a lesser-known person in the UK (e.g., Irvin Brooks, Kenny Kramm) and encouraged to look up additional information on Wikipedia. The URL to the person's Wikipedia page was provided.
  - Fictitious persons: we used the *charlatan* R package (Chamberlain and Voytovich 2020) to generate profiles consisting of a full name (e.g., Amelie Crooks), a relationship status (e.g., married), a job (e.g., software engineer), an age and a country of residence (e.g., Belgium).
- Participants: A total of $n = 401$ participants wrote three person descriptions each, resulting in a corpus of 1202 person descriptions (famous: 401, semifamous: 401, fictitious: 400)[8]. The participants who wrote these person descriptions had a mean age of 34.79 years ($SD = 11.99$), with 65.58% females. The data were collected through Prolific Academic and each participant was paid GBP 1.25 per set of three descriptions.
- Corpus details: The person descriptions were, on average, 130.49 tokens long ($SD = 35.17$), which did not differ between the type of person description, $F(2, 1199) = 2.25, p = .106$, ($M_{famous} = 128.54$, $SD_{famous} = 29.07$; $M_{semifamous} = 129.44$, $SD_{semifamous} = 41.94$; $M_{fictitious} = 133.48$, $SD_{fictitious} = 33.16$). The paragraph below shows a verbatim example of a description from the 'very famous' category (person described: actress Emma Watson).

Emma Watson is most well known for starring as Hermione Granger in Harry Potter, she was in all of the films. She is also known for playing Belle in beauty and the beast. She has strong feminist views. Emma Watson was born in Paris and bought up in Oxfordshire. Emma Watson is currently 30 years old, her date of birth being 15th April 1990. Emma Watson has 4 siblings, she attended Brown University. Both Emma Watsons's parents are lawyers. Emma Watson speaks French but claims not as well as she used to. Emma Watson has also ventured into modelling, she became the face of Burberry and resulted in her earning a 6 figure sum from the campaign.

3.1.3.2 Motivated intruder testing

- Task: We created three tasks, one for each difficulty level (famous, semifamous, fictitious). Each participant in the motivated intruder task judged a random selection of ten person descriptions in the respective group, and we aimed to collect three motivated intruder judgments for each person description. After providing informed consent, the participants were informed that they had to "play an adversary" whose task it was to find out who a given person's description was about. After reading a description, they had to state i) whether they could identify the person (yes/no), ii) who they think it is (if they could identify the person) or the extent of knowledge that they could establish about the person (if they could not identify the person), and iii) which part of the text revealed the identity. After making these judgments for ten person descriptions, each participant was debriefed, automatically reimbursed for their time (1.50 GBP) and redirected to Prolific Academic. The data collection task was conducted in a custom-made web interface.
- Participants: A total of $n = 366$ participants completed this part of the study. Of these participants, $n = 122$ judged descriptions of famous persons, $n = 123$ those of semifamous persons, and $n = 121$ those of fictitious persons. The mean age of participants was 27.31 years ($SD = 9.37$), with 40.98% female participants. In total, the dataset consisted of 3660 judgments. Each person description was judged, on average, 3.06 times ($SD = 1.09$).
- Measuring de-anonymisation: We examined the de-anonymisation of individuals by participants by calculating the string similarity between the correct solution (e.g., Sam Smith) and the participant-

[8]Note that one description of a fictitious person was missing.

indicated person. Specifically, we calculated the cosine similarity between the vectors of individual characters (e.g., s, a, m, s, m, i, t, h). The resulting similarity is a score between -1.00 and +1.00, with values closer to +1.00 indicating higher similarity. The multiple judgments per item were averaged to obtain a single (average) similarity per item. To allow for fuzziness (e.g., typos), we chose a cosine similarity of 0.75 as the cut-off above which we deemed a person successfully identified, resulting in a binary outcome (identified vs not identified). In addition to that binarisation, we used cosine similarity as a continuous variable to assess the degree of similarity.

3.2 Results

3.2.1 Technical evaluation

When assessing the tagging performance of the entity-detection model, we observe that the model accurately predicts most categories, with F1-scores above 80% for 10 out of 12 categories (Table 1). However, we also observe that the model performs poorly on OCCUPATION (F1-score of 52%) and OTHER_IDENTIFYING_ATTRIBUTE (F1-score of 69%). To be clear, a confusion of categories does not necessarily mean that a phrase is not removed during anonymisation; just that it may not have been replaced with the correct context-preserving placeholder. The high F1-score for the NONE category (96%) implies that sub-categories notwithstanding, PSI is successfully identified.

Table 1: Performance results of the trained Textwash model on the test set.

Entity tag	Precision	Recall	F1-score	Support
ADDRESS	0.84	0.93	0.88	556
DATE	0.95	0.95	0.95	3301
EMAIL_ADDRESS	0.96	0.98	0.97	1815
LOCATION	0.77	0.83	0.80	2111
OCCUPATION	0.43	0.65	0.52	307
ORGANIZATION	0.85	0.79	0.82	7300
PERSON_FIRSTNAME	0.91	0.89	0.90	4278
PERSON_LASTNAME	0.94	0.91	0.92	6434
PHONE_NUMBER	0.96	0.95	0.95	874
TIME	0.90	0.91	0.90	934
OTHER_IDENTIFYING_ATTRIBUTE	0.64	0.74	0.69	3292
NONE	0.97	0.96	0.96	55134
accuracy			0.93	86336
macro avg	0.84	0.87	0.86	86336
weighted avg	0.93	0.93	0.93	86336

3.2.2 Information loss

3.2.2.1 Utility loss On the original dataset, the model achieves a test set accuracy of 92.82% (precision: 92.69%, recall: 92.97%, F1: 92.83%), which is comparable to existing work (Mozes et al. 2021). On the anonymised dataset, the trained RoBERTa model achieves an accuracy of 91.98% (precision: 91.41%, recall: 92.67%, F1: 92.04%) on the test set, resulting in a utility loss of 0.84%. These findings indicate that anonymising documents has a negligible influence on the performance of the sentiment classification task, thereby retaining the usefulness of the anonymised dataset for downstream text classification tasks.

3.2.2.2 Construct loss The findings for the construct loss analysis indicate that the information is preserved for the POS tags where that was expected (Table 2). We see substantial deviations from the original text data for adjectives, adverbs, nouns, numerals, pronouns and proper nouns - all of these are detected in the Textwash tool and replaced with placeholders that are not captured by POS tagging algorithms. If one

were to map the replacements (e.g., PERSON_X) to desired POS tags (here: noun), the POS information would be fully recoverable. As a whole, these findings suggest that anonymisation retains the usefulness of linguistic variables.

Table 2: Bayes factors for the construct loss test on POS frequencies (original vs anonymised text data).

POS	Description	BF_{10}	BF_{01}
X	Other	0.05	21.03
INTJ	Interjections	0.05	20.85
CCONJ	Coordinating conjunction	0.05	19.72
DET	Determiners	0.05	19.66
VERB	Verbs	0.06	16.26
SYM	Symbols	0.06	16.17
PART	Particles	0.07	13.71
ADP	Adposition	0.09	11.66
ntok	No. of words	0.19	5.17
PUNCT	Punctuation	0.20	4.98
ADJ	Adjectives	1.277946e+35	0.00
ADV	Adverbs	5.712415e+05	0.00
NOUN	Nouns	5.369620e+182	0.00
NUM	Numerals	1.192332e+223	0.00
PRON	Pronouns	Inf	0.00
PROPN	Proper nouns	1.984156e+28	0.00

3.2.3 De-anonymisation

3.2.3.1 Overall de-anonymisation Table 3 shows the de-anonymisation rates for each person description level.[9] As expected, very famous people are identified more often than semifamous or fictitious people. Almost 19% of famous persons are identified in the motivated intruder test, while merely 2% and 1% of the semifamous and fictitious persons could be identified by name, respectively.

Table 3: Cosine similarities between the true person name and the participant choice (M, SD) and (un)successful de-anonymisations per type.

Item type	M	SD	% identified	SE % identified
famous	0.41	0.36	18.25	1.93
fictitious	0.04	0.13	1.01	0.50
semifamous	0.13	0.20	2.01	0.70

3.2.3.2 Person analysis and information leakage There was considerable variation among the items in the de-anonymisation rate and cosine string similarity. To avoid flooring effects, we only looked at the famous persons' descriptions (Table 4). While some famous persons were only identified in less than 10% of the cases (Cumberbatch, Jagger, Bale), some had a de-anonymisation rate of over 25% (John, Radcliffe, Smith).

[9]The participants' self-reported success in de-anonymisation - "Could you identify the person?" - had high agreement with the actual de-anonymisation of 92.43%. We use the actual de-anonymisation as a criterion throughout the paper.

Table 4: Mean (SD) cosine similarities per famous person and de-anonymisation rate (%) with standard error (SE) for Study 1.

Name	*M*	*SD*	% identified	SE % identified
benedict cumberbatch	0.25	0.29	4.65	3.25
sam smith	0.50	0.40	32.50	7.50
ed sheeran	0.56	0.34	28.95	7.46
emma watson	0.39	0.34	16.67	5.82
elton john	0.47	0.37	25.00	6.60
mick jagger	0.30	0.31	7.32	4.12
adele	0.46	0.35	18.42	6.37
daniel radcliffe	0.48	0.37	30.00	7.34
christian bale	0.36	0.34	8.33	4.67
hugh grant	0.29	0.32	10.53	5.05

To understand how participants identified these individuals, we looked at the information provided in the motivated intruder test when asked "what revealed the identity for you?" We analysed the answers that participants provided when they were successful by looking at the n-grams[10] (unigrams, bigrams, trigrams) most telling for each item. Table 5 shows the top 10 n-grams per item. We see that the persons that had relatively high de-anonymisation rates were identified through the leakage of very specific details: "glasses" and "gay" were mentioned for Elton John and presumably allowed participants to combine information about the person being a singer and these attributes in a "guesstimate" that it might be Elton John. Daniel Radcliffe was identified through being an actor playing a role in Harry Potter, and for Sam Smith, the participants mentioned song titles and "gay" as revealing attributes.

Table 5: Top-10 n-grams that revealed the person's identity in the motivated intruder test in Study 1.

Name	Top-10 n-grams
adele	weight, song, songs, loss, weight_loss, names, famous, fact, titles, name
benedict cumberbatch	strange, doctor, doctor_strange, fact, roles, sherlock, movie, played, role, description
christian bale	batman, role, movie, movies, american, psycho, oscar, dark, american_psycho, film
daniel radcliffe	harry, potter, harry_potter, young, black, boy, j.k, actor, woman, woman_black
ed sheeran	hair, ginger, songs, ginger_hair, red, name, musician, names, music, red_hair
elton john	songs, song, singer, glasses, gay, name, music, names, first, description
emma watson	harry, potter, harry_potter, beauty, beast, beauty_beast, role, actress, feminist, activist
hugh grant	movies, bridget, jones, bridget_jones, film, name, hill, movie, actor, played
mick jagger	rolling, stones, rolling_stones, singer, song, name, singer_rolling, singer_rolling_stones, band, fact
sam smith	song, songs, name, title, la, wall, singer, song_title, name_song, gay

3.2.3.3 Non-identifying information leakage In addition to the *revealing leakage* that helped intruders identify a person, we also asked those who did not identify the person what they could still recover as information from the anonymised text data. Table 6 shows the most mentioned n-grams per person and suggests that in cases *when the person was not identified*, it was mainly generic information. It is possible that in some cases, the intruder was not able to combine these pieces effectively: for example, Ed Sheeran could have been identified with a combination of the attributes "ginger," "hair," "pop," and "music."

[10] An n-gram is a sequence of n tokens. We removed stopwords before obtaining n-grams.

Table 6: Top-10 n-grams that were leaked but did not allow the motivated intruder to identify the person in Study 1.

Name	Top-10 n-grams
adele	singer, famous, person, female, musician, know, music, pop, awards, famous_singer
benedict cumberbatch	actor, person, famous, know, movies, male, movie, married, tv, man
christian bale	actor, person, male, movies, famous, nothing, won, roles, many, movie
daniel radcliffe	actor, person, hair, know, famous, films, brown, brown_hair, movie, male
ed sheeran	ginger, know, singer, hair, daughter, ginger_hair, person, music, pop, lot
elton john	singer, famous, gay, person, musician, married, song, songs, children, name
emma watson	actress, model, actor, person, female, feminist, famous, know, actress_model, nothing
hugh grant	actor, film, producer, romantic, movies, person, plays, know, english, one
mick jagger	person, singer, band, famous, know, rock, married, male, well, musician
sam smith	singer, gay, person, songwriter, won, voice, non-binary, musician, childhood, lot

3.2.3.4 Item-level analysis The findings presented above are aggregations of items per person described. But since each person has been described in multiple texts, there may be variation in the de-anonymisability even across items relating to the same individual. On average, each person has been described in 40 unique person descriptions ($SD = 2.54$), of which each was subject to the motivated intruder test on average 3.06 times ($SD = 1.09$). We explored how - for one person - the de-anonymisability differed. Below we show examples for two persons (Ed Sheeran and Sam Smith) - one for each that was always identified and one that was never identified.

- Ed Sheeran (never identified): *PERSON_FIRSTNAME_2 PERSON_LASTNAME_1 is a famous LOCATION_1 musician - singer and songwriter. Born on DATE_4 in Halifax, UK, under the full name PERSON_FIRSTNAME_1 PERSON_LASTNAME_1 . PRONOUN estimated net worth as of DATE_3 is NUMERIC_5 NUMERIC_4 NUMERIC_4. PRONOUN is also known as a record producer, as well as actor PRONOUN played PRONOUN in a LOCATION_4 soap opera OTHER_IDENTIFYING_ATTRIBUTE_1, which was filmed while PRONOUN was in the country in DATE_2 for a NUMERIC_1-off performance). PRONOUN career began somewhere in DATE_5, but PRONOUN shot to international fame in DATE_1. and had an immense success ever since. Currently, PRONOUN is said to be the NUMERIC_7th richest musician in the LOCATION_2.*
- Ed Sheeran (always identified): *PERSON_FIRSTNAME_2 is a famous musician,singer and songwritter. PRONOUN has ginger hair and wears glasses. PRONOUN has just had a baby i think. PRONOUN is english and has written songs for lots of other artists PRONOUN songs are very popular all around the world PRONOUN wrote a really good song PRONOUN did with OTHER_IDENTIFYING_ATTRIBUTE_1 swift.i like PRONOUN album shape of you . its really good PRONOUN is also a record producer and an actor PRONOUN has sold more than NUMERIC_3million records worldwide PRONOUN has won all sorts of awards e.g.honoary degrees from universities,the MbEFOR SERVICES TO MUSIC PRONOUN lives in LOCATION_1. i really dont know what else to write. im sorry but that is all i know. it pretty much covers everything you would want to know about PERSON_FIRSTNAME_2.*
- Sam Smith (never identified): *PERSON_FIRSTNAME_4 PERSON_LASTNAME_3 is an internationally successful singer-songwriter. They were born on DATE_2 in LOCATION_1. They became famous in DATE_1 and have been nominated and won multiple musical awards. PERSON_FIRSTNAME_6 announced NUMERIC_1 years ago that they were genderqueer and preferred the pronoun, 'they.' PERSON_FIRSTNAME_6 has recorded NUMERIC_2 albums: OTHER_IDENTIFYING_ATTRIBUTE_1 and OTHER_IDENTIFYING_ATTRIBUTE_2. PERSON_FIRSTNAME_6's networth is estimated at NUMERIC_4 million. Their main musical genres are R&B, pop and soul and they are signed for ORGANIZATION_1. Famous relatives include their third cousins: singer PERSON_FIRSTNAME_3 PERSON_LASTNAME_1 and actor PERSON_FIRSTNAME_5 PERSON_LASTNAME_1. PERSON_FIRSTNAME_6 is gay and has previously dated actor and model PERSON_FIRSTNAME_1 PERSON_LASTNAME_4 and actor PERSON_FIRSTNAME_2 PERSON_LASTNAME_2.*

- Sam Smith (always identified): *PERSON_FIRSTNAME_1 PERSON_LASTNAME_2 was born on DATE_2. PRONOUN is a singer/songwriter. PRONOUN was born in LOCATION_1 and has net worth of NUMERIC_4 million . PRONOUN hits include Money on my Mind, Writings on the Wall and OTHER_IDENTIFYING_ATTRIBUTE_3. PRONOUN debut album was OTHER_IDENTIFYING_ATTRIBUTE_2. PRONOUN has won several awards including NUMERIC_5 Grammy Awards and a Golden Globe award. PRONOUN parents are called PERSON_FIRSTNAME_2 PERSON_LASTNAME_2 and PERSON_FIRSTNAME_3 PERSON_LASTNAME_1. PRONOUN had liposuction when PRONOUN only NUMERIC_3 years old and reported to be bullied badly as a child. PRONOUN came out as gay in DATE_3 and then said PRONOUN was genderqueer in DATE_1. PRONOUN is quoted as saying I feel as much a woman as a man.*

These examples offer a glimpse at the difficulty in text anonymisation attributable to tiny details that, when combined, reveal an identity. In Study 2, we seek to learn more about identifiable and non-identifiable text data properties.

3.3 Discussion

Study 1 shows that i) the Textwash anonymisation can anonymise practically all person descriptions of individuals who are not famous and about whom intruders did not, therefore, have access to "privileged information." The most challenging test (anonymising the most famous individuals in the UK) was successful in 81.75% of the cases. We further found that ii) anonymisation is very hard if even minute details are leaked. In the case of very famous persons, a song title, reference to recent weight loss (Adele) or movie role (Christine Bale, Benedict Cumberbatch) can be sufficient for de-anonymisation. We also found that iii) considerable variation in identifiability exists even in person descriptions of the same individual.

To further improve the anonymisation, it would thus be helpful to understand how identified text data differ from fully anonymised texts. In Study 2, we extend and replicate the findings for the famous persons group and examine the properties of (non-)identified person descriptions.

4 Study 2

4.1 Method

Since the aim of Study 2 was to replicate findings from Study 1 and understand the properties of unsuccessful and successful anonymisation, we sought to avoid flooring effects and used only very famous individuals. The list of individuals was extended to a total of 20 famous people. We gathered a new sample of person descriptions for all of them.

4.1.1 Person descriptions

The task was identical to the person description task from Study 1 with three exceptions: the item pool was extended to 20 person descriptions, and each participant was paid GBP 3.75 for the task and wrote five descriptions. A total of $n = 200$ participants ($M_{age} = 31.07$ years, $SD = 8.32$, 73.00% female) wrote 1080 person descriptions with an average length of 112.13 tokens ($SD = 24.68$). Each person was described in - on average - 54 person descriptions ($SD = 4.09$).

4.1.2 Motivated intruder test

We recruited $n = 222$ participants from Prolific Academic for the motivated intruder task ($M_{age} = 32.19$, $SD = 10.09$, 68.92% female). The task instructions were identical to those from Study 1. Each participant was assigned a random selection of ten texts and was paid GBP 1.75. We aimed to collect two judgments per text and obtained 2.06 judgments on average per item ($SD = 0.73$).

4.1.3 Examining text anonymisability

By choosing the most difficult text anonymisation task (i.e., very famous people), we expect some texts to result in re-identifications of the individual. That allows us to test whether texts that led to de-anonymisation differ statistically from those that were not de-anonymised on the following variables.

4.1.3.1 Proportion of anonymised text We define the proportion of anonymised text as $P_{removed} = 1 - \frac{ntok_{anonymised}}{ntok_{original}}$, with *ntok* being the number of tokens in the anonymised or original text. We removed all anonymisation tags (e.g., [Person_1]) from the anonymised documents.

4.1.3.2 Global frequency ranks of anonymised texts A look at the leaking information from Study 1 (Tables 5 and 6) suggests that very specific pieces of information lead intruders to identify individuals. We test whether the words in the de-anonymised documents differ from those in the documents that did not result in re-identification regarding their global occurrence frequency. The frequency is operationalised as the average frequency rank of the words in each document based on a list of the most frequent 10k words based on the Google Trillion Word Corpus[11]. Higher rank scores imply a low frequency. If highly specific - and hence less frequent - words reveal an identity, we expect a higher global frequency rank score for identified person descriptions than non-identified ones.

4.1.3.3 Perplexity of anonymised texts Another way to measure the "unusualness" of the information left in the anonymised documents is perplexity. Perplexity is the inverse exponentialised probability of observing a sequence of words as computed by a language model. Low perplexity implies a higher probability assigned to the input sequence and, therefore, a lower unusualness of the text. We compute the perplexity of input text using a pre-trained GPT language model (Radford et al. 2018) as provided by the HuggingFace Transformers library (Wolf et al. 2019).

4.2 Results

4.2.1 De-anonymisation[12]

The mean de-anonymisation rate was 26.39% ($SE = 1.34$) with an average cosine similarity of 0.42 ($SD = 0.39$) between the actual person's name and the participant's input. Table 7 shows the variation in de-anonymisation among the items in the augmented stimuli set of Study 2. Some newly added person descriptions showed particularly high de-anonymisation rates (e.g., Beckham, Hamilton, Middleton). When we only consider the persons that were also among the famous people in Study 1, we obtain a comparable de-anonymisation rate as in Study 1 ($M = 22.63\%$, $SE = 1.79$).

Table 7: Mean (SD) cosine similarities per famous person and de-anonymisation rate (%) with standard error (SE) for Study 2.

Name	M	SD	% identified	SE % identified
adele	0.53	0.41	38.98	6.40
christian bale	0.24	0.31	9.80	4.21
david beckham	0.68	0.39	55.10	7.18
naomi campbell	0.41	0.34	19.61	5.61
daniel craig	0.35	0.38	22.22	5.71
benedict cumberbatch	0.18	0.29	7.41	3.60
cara delevigne	0.36	0.40	23.53	6.00
judi dench	0.14	0.25	5.36	3.04
ricky gervais	0.47	0.39	28.26	6.71
hugh grant	0.39	0.40	22.92	6.13

[11] https://github.com/first20hours/google-10000-english

[12] The agreement between participant-indicated identification and actual identification was 78.38%. As in Study 1, we used the actual identification for analysis.

Name	M	SD	% identified	SE % identified
lewis hamilton	0.69	0.35	50.82	6.45
mick jagger	0.26	0.32	9.43	4.05
elton john	0.57	0.37	38.89	6.70
kate middleton	0.60	0.34	44.64	6.70
kate moss	0.39	0.40	22.81	5.61
daniel radcliffe	0.25	0.33	10.17	3.97
jk rowling	0.49	0.35	29.41	6.44
ed sheeran	0.60	0.40	45.00	6.48
sam smith	0.40	0.35	16.07	4.95
emma watson	0.36	0.41	24.07	5.87

4.2.2 Information leakage

When we look at the leakage that led to successful identifications of the ten most identified persons in Study 2, we observe that similar to Study 1, very specific features revealed the identities of individuals (Table 8). For example, Lewis Hamilton was identified by the combination of "formula 1" and "black," while David Beckham was identified through the mention of being married to "posh spice"[13].

Table 8: Top-10 n-grams that revealed the person's identity in the motivated intruder test in Study 2.

Name	Top-10 n-grams
adele	song, weight, loss, weight_loss, titles, songs, lost, song_titles, singer, age
cara delevigne	model, eyebrows, dyspraxia, book, singer, bisexual, comic, comic_book, actor, know
david beckham	spice, married, girl, spice_girl, footballer, posh, posh_spice, married_spice, married_spice_girl, football
ed sheeran	ginger, hair, song, guitar, singer, married, team, ginger_hair, names, album
elton john	pianist, song, gay, piano, singer, married, charity, outfits, costumes, funeral
emma watson	harry, potter, harry_potter, rights, actress, womens, film, womens_rights, feminist, activist
jk rowling	books, author, harry, series, potter, harry_potter, book, name, female, writer
kate middleton	royal, family, prince, throne, married, royal_family, line, line_throne, married_prince, future
lewis hamilton	driver, racing, formula, racing_driver, 1, black, f1, formula_1, race, car
ricky gervais	office, comedian, comedy, shows, afterlife, reference, life, actor, show, office_afterlife

4.2.3 Statistical differences in text anonymisability

The comparisons between de-anonymised and anonymised documents on linguistic variables (Table 9) suggest that the only effect observed was for the proportion anonymised (effect size Cohen's $d = 0.19$): documents that were de-anonymised had a marginally lower percentage removed from the original text than documents which remained anonymous. The other variables (frequency ranks and perplexity) did not indicate a significant difference.

[13]The name his wife, Victoria Beckham, used in the band Spice Girls.

Table 9: Means (SDs) for anonymised (0) and de-anonymised documents (1) for i) the proportion of anonymised text, global rank frequency, and perplexity. The effect size Cohen's d (with 95% CI) represents the magnitude of the difference between anonymised and de-anonymised documents per variable.

Variable	M_0	SD_0	M_1	SD_1	d
Proportion anonymised	22.94	8.77	21.45	7.09	0.19 [0.03; 0.35]
Frequency rank (original)	1131.21	237.74	1109.18	241.19	0.09 [-0.07; 0.25]
Frequency rank (anonymised)	1012.19	231.01	1003.55	224.06	0.04 [-0.12; 0.20]
Perplexity (original)	66.68	47.96	69.29	35.98	-0.06 [-0.23; 0.10]
Perplexity (anonymised)	145.12	180.61	149.32	81.34	-0.03 [-0.20; 0.13]

The absence of evidence that successful anonymisation is detectable through quantifiable linguistic indices suggests, again, that more subtleties are at play. As a last exploratory analysis, we now look in detail at three examples, their leaked information and the use thereof by intruders for raw data.

4.2.4 Case studies to understand de-anonymisation

The intruder task can be abstracted as follows: they read the anonymised text, combine potentially leaking information to de-anonymise the text and provide a brief explanation about what they used to de-anonymise the individual. We look at three examples that shed light on that decision-making in practice:

Example 1: Kate Middleton

Original text: Kate Middleton is the wife of Prince William. She is a mother of 3 children; 2 boys and a girl. Kate is educated to university level and that is where she met her future husband. Kate dresses elegantly and is often seen carrying out charity work. However, she is a mum first and foremost and the interactions we see with her children are adorable. Kate's sister, Pippa, has followed Kate into the public eye. She was born in 1982 and will soon turn 40. When pregnant, Kate suffers from a debilitating illness called Hyperemesis Gravidarum, which was little known about until it was reported that Kate had it.

Anonymised text: [firstname1] [lastname1] is the wife of [occupation1] [lastname2]. [pronoun] is a mother of [numeric] children; [numeric] boys and a girl. [firstname1] is educated to university level and that is where [pronoun] met [pronoun] future husband. [firstname1] dresses elegantly and is often seen carrying out charity work. However, [pronoun] is a mum first and foremost and the interactions we see with [pronoun] children are adorable. [firstname1]'s sister, [firstname2], has followed [firstname1] into the public eye. [pronoun] was born in [date1] and will soon turn [numeric]. When pregnant, [firstname1] suffers from a debilitating illness called [otherattribute1], which was little known about until it was reported that [firstname1] had it.

Information mentioned by intruder: "Suffered from an illness in pregnancy and has a famous sister."

Example 2: Lewis Hamilton

Original text: Lewis Hamilton is a British racing driver. He currently competes in Formula One for Mercedes, having previously driven for McLaren. In Formula One, Hamilton has won a joint-record seven World Drivers' Championship titles (tied with Michael Schumacher), and holds the records for the most wins, pole positions, and podium finishes. Hamilton is an advocate against racism and for increased diversity in motorsport. More recently, he took the knee before every race he entered in the 2020 Formula One season in support of the Black Lives Matter movement and wore t-shirts bearing the Black Lives Matter slogan. Following the murder of George Floyd, he criticised prominent figures in Formula One for their silence on the issue.

Anonymised text: [firstname1] [lastname1] is a [location1] racing driver. [pronoun] currently competes in [organisation1] for [organisation2], having previously driven for [organisation3]. In [organisation1], [lastname1] has won a joint-record [numeric] World Drivers' Championship titles (tied with [firstname2] [lastname2] , and holds the records for the most wins, pole positions, and podium finishes. [lastname1] is an advocate against

racism and for increased diversity in motorsport. More recently, [pronoun] took the knee before every race [pronoun] entered in the [date1] [organisation1] season in support of the [otherattribute1] movement and wore t-shirts bearing the [otherattribute1] slogan. Following the murder of [firstname3] [lastname3], [pronoun] criticised prominent figures in [organisation1] for their silence on the issue.

Information mentioned by intruder: "A top racing driver. advocate against racism and recently took the knee against racism. I believe he criticised prominent people for their silence on the murder of George Floyd."

Example 3: Daniel Craig

Original text: He is an English film actor known for playing James Bond in the 007 series of films. Since 2005, he has been playing the character but he confirmed that No Time to Die would be his last James Bond film. He was born in Chester on 2nd of March in 1968. He moved to Liverpool when his parents divorced and lived there until he was sixteen years old. He auditioned and was accepted into the National Youth Theatre and moved down to London. He studied at Guildhall School of Music and Drama. He has appeared in many films.

Anonymised text: [pronoun] is an [location1] film actor known for playing [otherattribute1] in the [otherattribute2] series of films. Since[date1], [pronoun] has been playing the character but [pronoun] confirmed that [otherattribute3] would be [pronoun] last [otherattribute1] film. [pronoun] was born in [location2] on [date2] of [date3] in [date4]. [pronoun] moved to [location3] when [pronoun] parents divorced and lived there until [pronoun] was [numeric] years old. [pronoun] auditioned and was accepted into the [organisation1] and moved down to [location4]. [pronoun] studied at [organisation2]. [pronoun] has appeared in many films.

Information mentioned by intruder: "The comment about it being a last film."

These examples illustrate the difficulty of anonymising text data. The first example led to de-anonymisation through the mention of the person having had an illness during pregnancy (not the actual name of the illness, which was tagged and removed) and a famous sister. Lewis Hamilton was identified by combining the information about being a driver and having a firm stance against racism. At no place was the name of the team he races for mentioned and references to the Black Lives Matter movement or George Floyd were anonymised. Lastly, in the third example, Daniel Craig was identified not through leakage of a name or film title but through the note that "[No Time To Die] would be [his] last [James Bond] film."

Note that in none of the cases was the name, a date of birth or occupation leaked. Instead, it was "soft" identifiers that the intruders could combine and connect to public knowledge to make well educated guesses. All three cases highlight the importance of normally "privileged information": these individuals can be identified only because very minute details about them are in the public domain, or because the information included only applies to a very narrow range of individuals.

4.3 Discussion

The second study replicated the findings from Study 1 for very famous individuals. Only negligible differences in text variables emerged between successfully and unsuccessfully anonymised texts on the proportion of tagged tokens. A qualitative analysis of the decision-making process suggests that minute details, combined with public information, allow for the identification of famous persons. Overall, the data show that text anonymisation works well and does not leak information that - without external public knowledge - can lead to the identification of an individual.

5 General discussion

Text anonymisation is one of the current hurdles in moving advances in text analysis and NLP research closer to practice and enabling sharing of text data. With an increased scale of datasets available and needed for computational text analysis, automated text anonymisation tools offer a promising approach to solving these issues. In this paper, we introduced and validated Textwash - an open-source software that uses machine learning and natural language processing to identify potentially sensitive information in unstructured text data and remove it. Importantly, the tool works with two end-users in mind: (1) for researchers, the removal of information must be done so that the anonymised documents retain the usefulness and utility of the

original documents. Put differently: researchers – for whom the PSI itself is often of no interest - can just as well work with the anonymised data and still obtain the same results in statistical analyses and prediction tasks. (2) For organisations and data-owners of sensitive data, it is important that the process of anonymising data is successful and does not itself introduce additional risks to the data. To mitigate any concerns on that level, Textwash works fully offline and the data never leave the user's system.

5.1 Core findings

This paper evaluated the text anonymisation software Textwash along three criteria: a technical, an information loss, and a de-anonymisation evaluation (see (Mozes and Kleinberg 2021)).

5.1.1 Technical evaluation

We assessed how many of the categories of phrase deemed meaningful to anonymise text data were correctly identified by the underlying model on unseen data. The weighted average of the F1 score over all categories was high (0.93), suggesting that the model is able to identify the categories correctly and makes the correct decision in the vast majority of cases (weighted recall and precision: 0.93). There is some variation across categories with "occupation" ($F1 = 0.52$) and "other_identifying_attribute" ($F1 = 0.69$) being at the lower end of the per-category performances. Although this results in some mismatching (e.g., an organisation identified as a location), it has no adverse effect on the anonymisation since even misclassified categories are removed. As a whole, the technical evaluation indicated that the tool is comparable to state-of-the-art entity recognition systems.[14]

5.1.2 Information loss

We used two common NLP tasks to test the information loss criterion, which is further subdivided into utility loss and construct loss. First, for utility loss, we compared the prediction performances for a popular movie review dataset with those for the same but anonymised dataset. The difference in performance can be attributed to the loss of information due to the anonymisation procedure. Here, the data suggest negligible loss ($< 1\%$). It is worth noting, however, that we focused on sentiment analysis as a downstream task which may be less affected by the removal of potentially sensitive information than other tasks. Future work could examine various downstream tasks for a more comprehensive picture. Second, we evaluated the construct loss by testing for statistical differences in part-of-speech tag frequencies between original and anonymised data. We used the person description data collected for the motivated intruder testing and found statistical evidence in favour of the hypothesis that the anonymisation procedure does not introduce frequency differences in relevant POS tags. We do observe considerable differences for POS tags that are directly removed by the anonymisation model (e.g., pronouns). These differences vanish when the replacements are mapped back to their corresponding POS tag (e.g., [PRONOUN_1] –> POS tag 'PRON').

In summary, the information loss evaluation suggested that, for popular research tasks (sentiment, POS tagging), the difference between raw and anonymised data is negligible, so the conclusions that can be derived from anonymised text data are identical to those from the raw, original data.

5.1.3 De-anonymisation

De-anonymisation was introduced as the litmus test for the tool and assessed with a motivated intruder procedure. We collected a set of highly detailed person descriptions for individuals of varying levels of fame. These descriptions were then presented in an anonymised form to human intruders tasked with the identification of identities. The results show that even the world's most famous individuals' descriptions are rendered anonymous in 82% (Study 1) and 74% (Study 2) of the cases. This finding is noteworthy because what would normally be considered privileged information (e.g., where someone was born, their partners) is in the public domain and often common knowledge for all these very famous individuals. This is not the case for the typical research context (e.g., interview transcripts, diary data) or envisioned data sharing (e.g.,

[14] See https://paperswithcode.com/sota/named-entity-recognition-ner-on-conll-2003.

police reports, health records), so the more realistic benchmarks are the findings for semifamous and, even more so, fictitious persons. For these, the descriptions are rendered anonymous in 98%-99% of the cases.

We also explored how, in the cases where an intruder succeeded, they were able to identify a very famous person. The findings here illuminate that it is highly specific details that can be used only in combination with public, privileged information for an educated guess. That kind of information leakage shows one of the challenges of anonymising qualitative data but does not threaten the anonymisation quality for the typical use case.

5.2 The need for transparent, empirical evaluation

To date, we are not aware of any text anonymisation tools evaluated to the same rigour as we did in this paper. We encourage others who develop anonymisation tools to follow suit and present more comprehensive empirical validation; this paper can serve as a guideline for an extensive procedure. At a minimum, a user of text anonymisation software should insist on an adequate empirical validation of the software they intend to use in terms of both anonymisation success and the characteristics of the output text. The datasets produced in the studies of the current paper are publicly available and can be used as a benchmark for other anonymisation tools. Similarly, a mere technical evaluation is insufficient: a tool could perform well on technical performance metrics but fail with the essential task of protecting the identity of individuals. The proposed motivated intruder testing as part of the TILD criteria (Mozes and Kleinberg 2021) is a way to assess the de-anonymisability of text data.

5.3 Implications

The availability of a validated text anonymisation tool has implications on various levels.

Research: With the rise of hybrid disciplines such as computational social science, there is an increased need to break down disciplinary silos. For text data, methodological advances typically come from natural language processing and machine learning research requiring large datasets (e.g., to train language models). Consequently, many of the advancements in these areas are based on and applied to easy-to-get large-scale data (e.g., Twitter data with all its known problems (Morstatter et al. 2013; Pfeffer, Mayer, and Morstatter 2018)). This may mean that research activity is biased towards topics that are most amenable for study rather than those that are most meaningful or important. With automated anonymisation, large datasets currently not in use due to the hurdle of manual anonymisation can become public datasets. Ideally, this would bring topical datasets (e.g., police reports, health records) closer to the computational methods that have advanced what we can learn from text data.

Data-sharing: The dilemma between data sharing and the privacy protection of research participants has long been considered an either-or question, often decided in favour of not sharing data. With the anonymisation tool introduced in this paper, that dilemma can be resolved efficiently and effectively removing the conflict between these two desirable open science practices (data sharing and privacy protection). Researchers and organisations can now share data without violating privacy or data protection guidelines. Likewise, science data archives can now fulfil their mission of making data publicly available – particularly for studies funded by public research councils - without risking privacy violations.

Open science: The umbrella of all of the above-listed ways in which automated, fast, and validated text anonymisation improves procedures is open science. On the one hand, anonymisation tools enable researchers who were, until now, understandably hesitant about sharing qualitative text data to make their data public and thereby meet the mandates of an increasing number of journals and research funding agencies. It is also beneficial for the reliability of science because by sharing more data, researchers can pool datasets and thereby increase the sizes thereof, which can ameliorate the controversy of small sample sizes (Yarkoni and Westfall 2017) and statistically under-powered studies (Maxwell 2004; Card et al. 2020). Ultimately, having more data available and large sample sizes will improve the reproducibility of research.

5.4 Limitations and outlook

Despite the promise of our evaluation results, a few points merit attention. First, even though the human intruder testing showed that practically no individual that is not very famous could be identified after detailed person descriptions are presented in anonymised form, the question of 'how good is good enough?' remains. We argue that in most contexts, the performances obtained with the current Textwash tool are sufficient to justify it as anonymisation, comparable to what a human anonymiser would be able to do in considerably more time. Here, future work could experimentally compare automated anonymisation approaches, including Textwash, with manual human anonymisation. Improving the tool is desirable regardless of the promising findings presented here. Future work in this area could include an active learning module into the Textwash pipeline: users are presented with a small number of samples of documents after anonymisation and provide feedback on the quality of the anonymised document. That way, the model can actively learn (and hence optimise for) what desirable anonymisation is.

Second, the current procedure assumed that all documents are equally anonymisable and that one degree of anonymisation fits all contexts. To further improve the tool's performance, future work could seek to identify an a priori document risk score. Using supervised machine learning, one could train a higher-level model to estimate the degree to which Textwash can anonymise data successfully, and then decide to submit 'difficult-to-anonymise' documents to human review. A related future improvement could lie in adjusting the level of anonymisation required: the stricter the anonymisation need, the lower one could set the probability threshold of the underlying machine learning model to remove a token. While the approach taken here resulted in little information loss, the degree of removed tokens and the usefulness and utility of anonymised data are inevitably a trade-off, and the user could tailor this to their needs.

Third, one option not explored in the current paper is a token-consistent replacement approach. Rather than replacing "London" with [LOCATION_1], we could replace it with "Madrid." Such an approach would likely improve the information loss criteria and possibly the de-anonymisation rate. However, we desisted from that procedure as it introduces additional complexity because it requires retaining semantic relations between tokens. For example, an original sequence of "[. . .] lived in the capital of Spain, Madrid, and plays for the local team Atletico Madrid" would require a token-consistent replacement where the relationship of country, capital city and local football club is retained. That complexity could be solved with external knowledge bases (Ji et al. 2022) in the future but is out of the scope of this current work.

6 Conclusion

Text anonymisation is an important pillar of open science efforts and has until now been insufficiently addressed for large-scale purposes. This paper introduced the Textwash tool that allows researchers, data-owners and individuals to automatically remove potentially sensitive information from text data, thereby enabling them to share data without compromising data privacy.

7 Acknowledgments

This work was supported by a Concept Grant from SAGE Ocean.

8 Full author affiliations

- Bennett Kleinberg, corresponding author: Department of Methodology & Statistics, Tilburg University, The Netherlands; Department of Security and Crime Science, University College London, UK.Contact: bennett.kleinberg@tilburguniversity.edu
- Toby Davies: Department of Security and Crime Science, University College London, UK. Contact: toby.davies@ucl.ac.uk
- Maximilian Mozes: Department of Computer Science & Dawes Centre for Future Crime, University College London, UK. Contact: maximilian.mozes@ucl.ac.uk

References

Adams, Allison, Eric Aili, Daniel Aioanei, Rebecca Jonsson, Lina Mickelsson, Dagmar Mikmekova, Fred Roberts, Javier Fernandez Valencia, and Roger Wechsler. 2019. "AnonyMate: A Toolkit for Anonymizing Unstructured Chat Data." In *Proceedings of the Workshop on NLP and Pseudonymisation*, 1–7.

Benoit, Kenneth, and Akitaka Matsuo. 2017. "Spacyr: R Wrapper to the Spacy NLP Library." https://CRAN.R-project.org/package=spacyr.

Berg, Hanna, Taridzo Chomutare, and Hercules Dalianis. 2019. "Building a de-Identification System for Real Swedish Clinical Text Using Pseudonymised Clinical Text." In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, 118–25. Hong Kong: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-6215.

Boyd, Ryan L., and H. Andrew Schwartz. 2021. "Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field." *Journal of Language and Social Psychology* 40 (1): 21–41. https://doi.org/10.1177/0261927X20967028.

Card, Dallas, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. "Computational Analysis of 140 Years of US Political Speeches Reveals More Positive but Increasingly Polarized Framing of Immigration." *Proceedings of the National Academy of Sciences* 119 (31): e2120510119.

Card, Dallas, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. "With Little Power Comes Great Responsibility." *arXiv Preprint arXiv:2010.06595*.

Chamberlain, Scott, and Kyle Voytovich. 2020. *Charlatan: Make Fake Data*. https://CRAN.R-project.org/package=charlatan.

Consortium, BNC, and others. 2007. "British National Corpus." *Oxford Text Archive Core Collection*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv Preprint arXiv:1810.04805*.

Di Cerbo, Francesco, and Slim Trabelsi. 2018. "Towards Personal Data Identification and Anonymization Using Machine Learning Techniques." In *European Conference on Advances in Databases and Information Systems*, 118–26. Springer.

Finkel, Jenny Rose, Trond Grenager, and Christopher D Manning. 2005. "Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 363–70.

Francopoulo, Gil. 2007. "TagParser: Well on the Way to ISO-Tc37 Conformance."

Francopoulo, Gil, and Léon-Paul Schaub. 2020. "Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP." In *Workshop on Legal and Ethical Issues (Legal2020)*, 9–14. ELRA.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3): 535–74.

Hassan, Fadi, David Sánchez, Jordi Soria-Comas, and Josep Domingo-Ferrer. 2019. "Automatic Anonymization of Textual Documents: Detecting Sensitive Information via Word Embeddings." In *2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE)*, 358–65. IEEE.

Information Commissioner's Office. 2012. "Anonymisation: Managing Data Protection Risk Code of Practice." https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf.

Ji, Shaoxiong, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications." *IEEE Transactions on Neural Networks and Learning Systems* 33 (2): 494–514. https://doi.org/10.1109/TNNLS.2021.3070843.

Kleinberg, Bennett, and Maximilian Mozes. 2017. "Web-Based Text Anonymization with Node. Js: Introducing NETANOS (Named Entity-Based Text Anonymization for Open Science)." *Journal of Open Source Software* 2 (14): 293.

Kleinberg, Bennett, Isabelle van der Vegt, and Maximilian Mozes. 2020. "Measuring Emotions in the COVID-19 Real World Worry Dataset." In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Lewis, Patrick, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. "Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art." In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, 146–57.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Roberta: A Robustly Optimized Bert Pretraining Approach." *arXiv Preprint arXiv:1907.11692*.

Maas, Andrew, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. "Learning Word Vectors for Sentiment Analysis." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–50.

Mamede, Nuno, Jorge Baptista, and Francisco Dias. 2016. "Automated Anonymization of Text Documents." In *2016 IEEE Congress on Evolutionary Computation (CEC)*, 1287–94. IEEE.

Maxwell, Scott E. 2004. "The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies." *Psychological Methods* 9 (2): 147.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose." *arXiv:1306.5204 [Physics]*, June. http://arxiv.org/abs/1306.5204.

Mozes, Maximilian, and Bennett Kleinberg. 2021. "No Intruder, No Validity: Evaluation Criteria for Privacy-Preserving Text Anonymization." *arXiv Preprint arXiv:2103.09263*.

Mozes, Maximilian, Pontus Stenetorp, Bennett Kleinberg, and Lewis Griffin. 2021. "Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 171–86.

Pfeffer, Jürgen, Katja Mayer, and Fred Morstatter. 2018. "Tampering with Twitter's Sample API." *EPJ Data Science* 7 (1): 50. https://doi.org/10.1140/epjds/s13688-018-0178-0.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. "Improving Language Understanding by Generative Pre-Training."

Romanov, Aleksandr, Anna Kurtukova, Anastasia Fedotova, and Roman Meshcheryakov. 2019. "Natural Text Anonymization Using Universal Transformer with a Self-Attention." In *Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), Saint Petersburg, Russia*, 22–37.

Rouder, Jeffrey N, Paul L Speckman, Dongchu Sun, Richard D Morey, and Geoffrey Iverson. 2009. "Bayesian t Tests for Accepting and Rejecting the Null Hypothesis." *Psychonomic Bulletin & Review* 16 (2): 225–37.

Salganik, Matthew J. 2019. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–42.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R.

Garnett. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2019. "Huggingface's Transformers: State-of-the-Art Natural Language Processing." *arXiv Preprint arXiv:1910.03771*.

Xia, Patrick, Shijie Wu, and Benjamin Van Durme. 2020. "Which *BERT? A Survey Organizing Contextualized Encoders." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7516–33. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.608.

Yarkoni, Tal, and Jacob Westfall. 2017. "Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning." *Perspectives on Psychological Science* 12 (6): 1100–1122. https://doi.org/10.1177/1745691617693393.