Silencing the Risk, Not the Whistle: A Semi-automated Text Sanitization Tool for Mitigating the Risk of Whistleblower Re-Identification

Dimitri Staufer
Technische Universität Berlin
Berlin, Germany
staufer@tu-berlin.de

Frank Pallas
Paris Lodron University Salzburg
Salzburg, Austria
frank.pallas@plus.ac.at

Bettina Berendt
TU Berlin, Weizenbaum Institute, and KU Leuven
Berlin and Leuven, Germany and Belgium
berendt@tu-berlin.de

ABSTRACT

Whistleblowing is essential for ensuring transparency and accountability in both public and private sectors. However, (potential) whistleblowers often fear or face retaliation, even when reporting anonymously. The specific content of their disclosures and their distinct writing style may re-identify them as the source. Legal measures, such as the EU Whistleblower Directive, are limited in their scope and effectiveness. Therefore, computational methods to prevent re-identification are important complementary tools for encouraging whistleblowers to come forward. However, current text sanitization tools follow a one-size-fits-all approach and take an overly limited view of anonymity. They aim to mitigate identification risk by replacing typical high-risk words (such as person names and other labels of named entities) and combinations thereof with placeholders. Such an approach, however, is inadequate for the whistleblowing scenario since it neglects further re-identification potential in textual features, including the whistleblower's writing style. Therefore, we propose, implement, and evaluate a novel classification and mitigation strategy for rewriting texts that involves the whistleblower in the assessment of the risk and utility. Our prototypical tool semi-automatically evaluates risk at the word/term level and applies risk-adapted anonymization techniques to produce a grammatically disjointed yet appropriately sanitized text. We then use a Large Language Model (LLM) that we fine-tuned for paraphrasing to render this text coherent and style-neutral. We evaluate our tool's effectiveness using court cases from the European Court of Human Rights (ECHR) and excerpts from a real-world whistleblower testimony and measure the protection against authorship attribution attacks and utility loss statistically using the popular IMDb62 movie reviews dataset, which consists of 62 individuals. Our method can significantly reduce authorship attribution accuracy from 98.81% to 31.22%, while preserving up to 73.1% of the original content's semantics, as measured by the established cosine similarity of sentence embeddings.

CCS CONCEPTS

• **Security and privacy** → **Privacy protections**; **Pseudonymity, anonymity and untraceability**; **Human and societal aspects of security and privacy**; • **Computing methodologies** → **Natural language processing**; • **Information systems** → **Retrieval tasks and goals**; *Web applications*; Data mining; • **Human-centered computing** → **Collaborative and social computing systems and tools**; • **Applied computing** → *Law, social and behavioral sciences.*

KEYWORDS

Text Sanitization, Whistleblower Anonymity, Authorship Obfuscation, Fine-tuning Language Models, LLM-based Rephrasing

1 INTRODUCTION

In recent years, whistleblowers have become "a powerful force" for transparency and accountability, not just in the field of AI [9], but also in other technological domains and across both private- and public-sector organizations. Institutions such as the AI Now Institute [9] or the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [22] have emphasized the key role of whistleblower protection for societal well-being and often also the organizations' own interests [21]. However, whistleblowing may be a threat for the organizations whose malfeasance is being revealed; thus (potential) whistleblowers often fear or face retaliation. Computationally-supported anonymous reporting seems to be a way forward, but even if reporting frameworks are sufficiently secure system- and network-wise, the report itself may allow inferences towards the whistleblower's identity due to its content and the whistleblower's writing style. Non-partisan organizations such as Whistleblower-Netzwerk e.V. (WBN) provide guidance on concise writing. Our interactions with WBN confirm that whistleblower testimonies often include unnecessary personal details.

Existing approaches modifying the texts of such reports appear promising, but they take an overly limited view of anonymity and – like whistleblower protection laws – address only parts of the problem. This is detailed in Section 2. To improve on these approaches, we propose, implement, and evaluate a novel classification and mitigation strategy for rewriting texts that puts the whistleblower into the loop of assessing risk and utility.

Our contributions are threefold. First (Section 3), we analyse the interleaved contributions of different types of identifiers in texts to derive a description of the problem for anonymous whistleblowing in terms of a trade-off between risk (identifiability of the whistleblower) and utility (of the rewritten text retaining sufficient information on the specific event details). We derive a strategy for assigning re-identification risk levels of concern to textual features composed of an automated mapping and an interactive adjustment of concern levels. Second (Section 4), we describe our toolwhich implements this strategy. It applies (i) the word/term-to-concern mapping using natural language processing to produce a sanitized but possibly ungrammatical intermediate text version, (ii) a Large Language Model (LLM) that we fine-tuned for paraphrasing to render this text coherent and style-neutral, and (iii) interactivity to draw on the user's context knowledge. Third (Section 5), we evaluate the resulting risk-utility trade-off. We measure the protection against authorship attribution attacks and utility loss statistically using an established benchmark dataset and show that it can significantly reduce authorship attribution accuracy while retaining utility. We also evaluate our our tool's effectiveness in masking direct and quasi-identifiers using the Text Anonymization Benchmark [48] and demonstrate its effectiveness on excerpts from a real-world whistleblower testimony. Section 6 sketches current limitations and future work. Section 7 describes ethical considerations and researchers' positionality, and it discusses possible adverse impacts.

2 BACKGROUND AND RELATED WORK

This section describes the importance of, and threats to, whistleblowing (Section 2.1) and the promises and conceptual and practical challenges of "anonymity" in reporting (Section 2.2). We survey related work on the anonymization/de-identification of text and argue why it falls short in supporting whistleblowing (Section 2.3).

2.1 Challenges of Safeguarding Whistleblowers

Whistleblowers play a crucial role in exposing wrongdoings like injustice, corruption, and discrimination in organizations [6, 41]. However, their courageous acts often lead to negative consequences, such as subtle harassment and rumors, job loss and blacklisting, and, in extreme cases, even death threats [34, 37, 58]. In Western nations, whistleblowing is largely viewed as beneficial to society [66], leading to protective laws like the US Sarbanes-Oxley Act of 2002 and the European Union's "Whistleblowing Directive" (Directive 2019/1937). The latter, for example, mandates the establishment of safe reporting channels and protection against retaliation. It also requires EU member states to provide whistleblowers with legal, financial, and psychological support. However, the directive faces criticism for its limitations. Notably, it does not cover all public-sector entities [63, p. 3] and leaves key decisions to member states' discretion [1, p. 652]. This discretion extends to the absence of mandatory anonymous reporting channels and permits states to disregard cases they consider "clearly minor", leaving whistleblowers without comprehensive protection for non-material harms like workplace bullying [63, p. 3]. Furthermore, according to White [70], the directive's sectoral approach and reliance on a list of specific EU laws causes a patchwork of provisions, creating a complex and possibly confusing legal environment, particularly for those sectors impacting human rights and life-and-death situations.

Last but not least, organizations often react negatively to whistleblowing due to the stigma of errors, even though recognizing these mistakes would be key to building a culture of responsibility [5, p. 12] and improving organizations and society [69]. The reality for whistleblowers is thus fraught with challenges, from navigating legal uncertainties to dealing with public perception [26, 51, 52], leaving many whistleblowers with no option but to report their findings anonymously [50]. However, "anonymous" reporting channels alone do not guarantee anonymity [5].

2.2 Anonymity, (De-)anonymization, and (De-/Re-)Identification

Anonymity is not an alternative between being identified uniquely or not at all, but "the state of being not identifiable within a set of subjects [with potentially the same attributes], the anonymity set" [46, p.9]. Of the manifold possible approaches towards this goal, state-of-the-art whistleblowing-support software as well as legal protections (where existing) focus on *anonymous communications* [5]. This, however, does not guarantee *anonymous reports.* Instead, a whistleblower's anonymity may still be at risk due to several factors, including: (i) surveillance technology, such as browser cookies, security mechanisms otherwise useful to prevent unauthenticated uses, cameras, or access logs, (ii) the author's unique writing style, and (iii) the specific content of the message [33]. Berendt and Schiffner [5] refer to the latter as "epistemic non-anonymizability", i.e., the risk of being identified based on the unique information in a report, particularly when the information is known to only a few individuals. In some cases, this may identify the whistleblower uniquely.

Terms and their understanding in the domain of anonymity vary. We use the following nomenclature: *anonymization* is a modification of data that increases the size of the anonymity set of the person (or other entity) of interest; conversely, *de-anonymization* decreases it (to some number $k \geq 1$). De-anonymization to $k = 1$, which includes the provision of an identifier (e.g., a proper name), is called *re-identification.* The removal of some identifying information (e.g., proper names), called *de-identification*, often but not necessarily leads to anonymization [4, 68].

In structured data, direct identifiers (e.g., names or social security numbers) are unique to an individual, whereas quasi-identifiers like age, gender, or zip code, though not unique on their own, can be combined to form unique patterns. Established mathematical frameworks for quantifying anonymity, such as Differential Privacy (DP) [16], and metrics such as k-anonymity [53], along with their refinements [27, 31], can be used when anonymizing datasets.

Unstructured data such as text, which constitutes a vast majority of the world's data, requires its own safeguarding methods, which fall into two broader categories [28]. The first, NLP-based text sanitization, focuses on linguistic patterns to reduce (re-)identification risk. The second, privacy-preserving data publishing (PPDP), involves methods like noise addition or generalization to comply with pre-defined privacy requirements [15].

2.3 Related Work: Text De-Identification and Anonymization, Privacy Models, and Adversarial Stylometry

De-identification methods in text sanitization mask identifiers, primarily using named entity recognition (NER) techniques. These methods, largely domain-specific, have been particularly influential in clinical data de-identification, as evidenced, for instance, by the 2014 i2b2/UTHealth shared task [62]. However, they do not or only partially address the risk of indirect re-identification [4, 38]. For example, Sánchez et al. [55, 56, 57] make the simplifying assumption that replacing noun phrases which are rare in domain-specific corpora or on the web with more general ones offers sufficient protection. Others use recurrent neural networks [12, 30], reinforcement learning [71], support vector machines [65], or pre-trained language models [23] to identify and remove entities that fall into pre-defined categories. However, all of these approaches ignore or significantly underestimate the actual risks of context-based re-identification.

More advanced anonymization methods, in turn, also aim to detect and remove identifiers that do not fit into the usual categories of named entities or are hidden within context. For example, Reddy and Knight [49] detect and obfuscate gender, and Adams et al. [2] introduce a human-annotated multilingual corpus containing 24 entity types and a pipeline consisting of NER and co-reference resolution to mask these entities. In a more nuanced approach, Papadopoulou et al. [44] developed a "privacy-enhanced entity recognizer" that identifies 240 Wikidata properties linked to personal identification. Their approach includes three key measures to evaluate if a noun phrase needs to be masked or replaced by a more general one [43]. The first measure uses RoBERTa [29] to assess how "surprising" an entity is in its context, assuming that more unique entities carry higher privacy risks. The second measure checks if web search results for entity combinations mention the individual in question, indicating potential re-identification risk. Lastly, they use a classifier trained with the Text Anonymization Benchmark (TAB) corpus [48] to predict masking needs based on human annotations.

Kleinberg et al.'s [24] "Textwash" employs the BERT model, fine-tuned on a dataset of 3717 articles from the British National Corpus, Enron emails, and Wikipedia. The dataset was annotated with entity tags such as "PERSON_FIRSTNAME", "LOCATION", and an "OTHER_IDENTIFYING_ATTRIBUTE" category for indirect re-identification risks, along with a "NONE" category for tokens that are non-re-identifying. A quantitative evaluation (0.93 F1 score for detection accuracy, minimal utility loss in sentiment analysis, and part-of-speech tagging) and its qualitative assessment (82% / 98% success in anonymizing famous / semi-famous individuals) show promise. However, the more recent gpt-3.5-turbo can re-identify 72.6% of the celebrities from Textwash's qualitative study on the first attempt, highlighting the evolving complexity of mitigating the risk of re-identification in texts [45].

In PPDP, several privacy models for structured data have been adapted for privacy guarantees in text. While most are theoretical [28], "C-sanitise" [54] determines the disclosure risk of a certain term t on a set of entities to protect (C), given background knowledge K, which by default is the probability of an entity co-occurring with a term t in the web. Additionally, DP techniques have been adapted to text, either for generating synthetic texts [20] or for obscuring authorship in text documents [68]. This involves converting text into word embeddings, altering these vectors with DP techniques, and then realigning them to the nearest words in the embedding model [73, 74]. However, "word-level differential privacy" [35] faces challenges: it maintains the original sentence length, limiting variation, and can cause grammatical errors, such as replacing nouns with unrelated adjectives, due to not considering word types.

Authorship attribution (AA) systems use stylistic features such as vocabulary, syntax, and grammar to identify an author. State-of-the-art approaches involve using Support Vector Machines [64, 72], and more recently, fine-tuned LLMs like BertAA [3, 18, 64]. The "Valla" benchmark and software package standardizes evaluation methods and includes fifteen diverse datasets [64]. Contrasting this, adversarial stylometry modifies an author's writing style to reduce AA systems' effectiveness [61]. Advancements in machine translation [67] have also introduced new methods based on adversarial training [60], though they sometimes struggle with preserving the original text's meaning. Semi-automated tools, such as "Anonymouth" [36], propose modifications for anonymity in a user's writing, requiring a significant corpus of the user's own texts. Moreover, recent advances in automatic paraphrasing using fine-tuned LLMs demonstrated a notable reduction in authorship attribution, but primarily for shorter texts [35].

To the best of our knowledge, there is no – and maybe there can be no – complete list of textual features contributing to the re-identification of individuals in text. As Narayanan and Shmatikov [40] highlight, "any attribute can be identifying in combination with others" [p. 3]. In text, we encounter elements like characters, words, and phrases, each carrying varying levels of meaning [19]. Single words convey explicit lexical meaning as defined by a vocabulary (e.g. "employee"), while multiple words are bound by syntactic rules to express more complex thoughts implicitly in phrases ("youngest employee") and sentences ("She is the youngest employee").

In addition, the European Data Protection Supervisor (EDPS) and Spanish Data Protection Agency (AEPD) [17] state that anonymization can never be fully automated and needs to be "tailored to the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons" [p. 7].

To take these insights and limitations into account, our semi-automated text sanitization tool leverages insights on the removal of identifying information but involves the whistleblower (the user) in the decision-making process.

3 RISK MODELLING AND RISK MITIGATION APPROACH

In this section, we derive the problem statement (Section 3.2) from an analysis of different identifier types (Section 3.1). Following an overview of our approach (Section 3.3), we detail the anonymization operations for textual features (Section 3.4) and the automatic assignment of default concern levels (Section 3.5).

3.1 Identifier Types, Author Identifiability, and Event Details in the Whistleblowing Setting

Whistleblowing reports convey information about persons, locations, and other entities. At least some of them need to be *identified* in order for the report to make any sense. The following fictitious example consists of three possible versions of a report in order to illustrate how different types of identifiers may contribute to the re-identification of the anonymously reporting employee Jane Doe, a member of the Colours and Lacquer group in the company COLOURIFICS.

V1 On 24 January 2023, John Smith poured polyurethane resin into the clover-leaf-shaped sink of room R23.

V2 After our group meeting on the fourth Tuesday of January 2023, the head of the Colours and Lacquer Group poured a toxin into the sink of room R23.

V3 Somebody poured a liquid into a recepticle on some date in a room of the company.

In V1, "John Smith" is the *lexical identifier*[1] of the COLOURIFICS manager John Smith, as is "24 January 2023" of that date. Like John Smith, room R23 is a unique named entity in the context of the company and also identified lexically. "Polyurethane resin" is the lexical identifiers of a toxin (both are common nouns rather than names of individual instances of their category). The "clover-leaf-shaped" serves as a descriptive identifier of the sink. In V2, John Smith is still identifiable via the *descriptive identifier* "head of the Colours and Lacquer Group", at least on 24 January 2023 (reconstructed with the help of a calendar and COLOURIFIC's personnel files). "Our" group meeting is an *indexical identifier* that signals that the whistleblower is one of the, say five employees in the Colours and Lacquer Group.

The indexical information is explicit in V2 given the background knowledge that only employees in this group were co-present (for example, in the company's key-card logfiles). The same information may be implicit in V1 (if it can be seen from the company's organigram who John Smith is and who works in his group). Both versions provide for the inference that Jane Doe or any of her four colleagues must have been the whistleblower. If, in addition, only Jane Doe stayed behind "after the meeting", that detail in V2 descriptively identifies her uniquely[2]. V3 contains only identifiers of very general categories. Many other variants are possible (for example, referencing, in a V4, "the head of our group", which would enlarge the search space to all groups that had a meeting in R23 that day).

The example illustrates the threats (i)-(iii) of Section 2.2. It also shows that the whistleblower's "anonymity" (or lack thereof) is only one aspect of a more general and graded picture of who and what can be identified directly, indirectly, or not at all – and what this implies for the whistleblower's safety as well as for the report's effectiveness.

Inspired by Domingo-Ferrer's [14] three types of (data) privacy, we distinguish between the identifiability of the whistleblower Jane Doe (*author*[3] *identifiability* A_{id}) and descriptions of the event or other wrongdoing, including other actors (*event details* E_{dt}). Given the stated context knowledge, we obtain an anonymity set of size $k = 1$ for John Smith in V1 and V2. Jane Doe is in an anonymity set of size $k = 5$ or even $k = 1$ in V2. In V1, that set may be of size $k = 5$ (if people routinely work only within their group) or larger (if they may also join other groups). Thus, the presence of a name does not necessarily entail a larger risk. Both are in an anonymity set containing all the company's employees at the reported date in V3 (assuming no outsiders have access to company premises). The toxin and the sink may be in a smaller anonymity set in V1 than in V2 or V3, and they could increase further (for example, if only certain employees have access to certain substances). Importantly, the identifiability of people and other entities in E_{dt} can increase the identifiability of the whistleblower.

V3 illustrates a further challenge: the misspelled receptacle may be a typical error of a specific employee, and the incorrect placement of the temporal before the spatial information suggests that the writer may be a German or Dutch native speaker. In addition to errors, also correct variants carry information that stylometry can use for authorship attribution, which obviously can have a large effect on A_{id}.

The whistleblower would, on the one hand, want to reduce all such identifiabilities as much as possible. On the other hand, the extreme generalization of V3 creates a meaningless report that neither the company nor a court would follow up on. This general problem can be framed in terms of risk and utility, which will be described next.

3.2 The Whistleblowing Text-Writing Problem: Risk, Utility, And Many Unknowns

A potential whistleblower faces the following problem: "make A_{id} as small as possible while retaining as much E_{dt} as necessary". We propose to address this problem by examining the text and possibly rewriting it.

In principle, this is an instance of the oft-claimed trade-off between privacy (or other risk) and utility. In a simple world of known repositories of structured data, one could aim at determining the identifying problem (e.g., by database joins to identify the whistleblower due to some attributive information they reveal about themselves and by multiple joins for dependencies such as managers and teams) and compute how large the resulting anonymity set (or A_{id} as its inverse) is. Given a well-defined measure of information utility, different points on the trade-off curve would then be well-defined and automatically derivable solutions to a mathematical optimization problem.

However, texts offer a myriad of ways to express a given relational information. The space of information that could be cross-referenced, sometimes in multiple steps, is huge and often unknown to the individual. Consequently, in many cases, it is not possible

[1] The classification of identifiers is due to Phillips [47]. Note that all types of identifiers can give rise to *personal data.*. in the sense of the EU's General Data Protection Regulation (GDPR), Article 4(1): "any information which is related to an identified or identifiable natural person", or *personally identifiable data* in the senses used in different US regulations. See [11] for legal aspects in the context of whistleblowing.

[2] If John Smith knows that only she observed him, she is also uniquely identified in V1, but for the sake of the analysis, we assume that only recorded data/text constitute the available knowledge.

[3] We assume that the potential whistleblower is also the author of the report. This is the standard setting. Modifications for the situation in which a trusted third party writes the report on their behalf are the subject of future work.

to determine the anonymity set size with any mathematical certainty. In addition, setting a threshold could be dangerous: even if the anonymity set is $k > 1$, protection is not guaranteed – for example, the whole department of five people could be fired in retaliation. At the same time, exactly how specific a re-written text needs to be about A_{id} and E_{dt} in order to make the report legally viable [4] cannot be decided without much more context knowledge. For example, the shape of the sink into which a toxic substance is poured probably makes no difference to the illegality, whereas the identity of the substance may affect it.

These unknowns have repercussions both for tool design (Section 3.3) and for evaluation design (Section 5.1.1).

3.3 Risk Mitigation Approach and Tool Design: Overview

Potential whistleblowers would be ill-served by any fully automated tool that claims to be able to deliver a certain mathematically guaranteed anonymization. Instead, we propose to provide them with a semi-automated tool that does have some "anonymity-enhancing defaults" that illustrate with the concrete material how textual elements can be identifying and how they can be rendered less identifying. Our tool starts with the heuristic default assumption that identifiability is potentially *always* problematic and then lets the user steer our tool by specifying how "concerning" *specific* individual elements are and choosing, interactively, the treatment of each of them that appears to give the best combination of A_{id} and E_{dt}. By letting the author/user assign these final risk scores in the situated context of the evolving text, we enable them to draw on a maximum of implicit context knowledge.

Our approach and tool proceed through several steps. We first determine typical textual elements that can constitute or be part of the different types of identifiers. As can be seen in Table 1, most of them can affect A_{id} and E_{dt}.

Since identification by name (or, by extension, pronouns that co-reference names) does not even need additional background knowledge and since individuals are more at risk than generics, we classify some textual features as "highly concerning", others as having "medium concern", and the remainder as "potentially concerning". We differentiate between two types of proper nouns. Some names refer to typical "named entities", which include, in particular, specific people, places, and organizations, as well as individual dates and currency amounts. These pose particular person-identification risk in whistleblowing scenarios.[5] "Other proper nouns", such as titles of music pieces, books and artworks generally only pose medium risk. For stylometric features, we explicitly categorize out-of-vocabulary words, misspelled words, and words that are surprising given the overall topic of the text. Other low-level stylometric features, such as punctuation patterns, average word and sentence length, or word and phrase repetition, are not (and in many cases, such as with character n-gram pattern, cannot be [25]) explicitly identified. Instead, we implicitly/indirectly account for them as a byproduct of the LLM-based rephrasing. For all other parts of speech, we propose to use replacement strategies based on data-anonymization operations that are proportional to the risk (Table 2). Given the complexities of natural language and potential context information, the latter two operations are necessarily heuristic; thus, our tool applies the classification and the risk mitigation strategy as a default which can then be adapted by the user.

Table 1: Overview of the approach from identifier types to default risk.

Identifier Type	Textual Feature	A_{id}/E_{dt}	Default Risk
Lexical	Names of named entities	A_{id},E_{dt}	High
Lexical	Other proper nouns	E_{dt}	Medium
Indexical	Pronouns	A_{id},E_{dt}	High
Descriptive	Common nouns	E_{dt},(A_{id})	Potential
Descriptive	Modifiers	E_{dt},(A_{id})	Potential
Descriptive (via pragmatic inferences)	Out-of-vocabulary words[a]	A_{id}, (E_{dt})	Medium
	Misspelled words[a]	A_{id}	Medium
	Surprising words[b]	A_{id}	Medium
	Other stylometric features	A_{id}	N/A[c]

[a]Treated as noun. [b]Nouns or proper nouns. [c]Not explicitly specified. Indirectly accounted for through rephrasing.

Table 2: Mitigation strategies based on assigned risk (LvC = level of concern, NaNEs = names of named entities, OPNs = other proper nouns, CNs = common nouns, Mods = modifiers, PNs = proper nouns, OSFs = other stylometric features).

LvC	NaNEs	OPNs	CNs	Mods	PNs	OSFs
High	Suppr.	Suppr.	Suppr.	Suppr.	Suppr.	Pert.
Medium	Pert.	Generl.	Generl.	Pert.	Suppr.	Pert.

3.4 Anonymization Operations for Words and Phrases

In our sanitization pipeline, we conduct various token removal and replacement operations based on each token's *POS tag* and its assigned level of concern (LvC), which can be "potentially concerning", "medium concerning", or "highly concerning". Initially, we consider all common nouns, proper nouns, adjectives, adverbs, pronouns, and named entities[6] as potentially concerning. Should the user or our automatic LvC estimation (see subsection 3.5) elevate the concern to either medium or high, we apply anonymization operations that are categorized into generalization, perturbation, and suppression. Specific implementation details are elaborated on in section 4.

[4] "a situation in which a plan, contract, or proposal is able to be legally enforced", https://ludwig.guru/s/legally+viable, retrieved 2024-01-02

[5] PERSON, GPE (region), LOC (location), EVENT, LAW, LANGUAGE, DATE, TIME, PERCENT, MONEY, QUANTITY, and ORDINAL

[6] By this, we mean names of named entities, e.g. "Berlin" for GPE, but we use `named entities` instead for consistency with other literature.

3.4.1 Generalization.

The least severe type of operation targets common nouns and other proper nouns marked as **medium concerning**. We assume their specificity (not necessarily their general meaning) poses re-identification risks. Thus, more general terms can be used to preserve meaning while mitigating the risk of re-identification.

- Common nouns like "car" are replaced with hypernyms from WordNet, such as "vehicle".
- Other proper nouns become broader Wikidata terms, e.g. "political slogan" for "Make America Great Again".

3.4.2 Perturbation.

This applies to modifiers[7] and named entities annotated as **medium concerning**. In this process, original words are retained but are assigned zero weight in the paraphrase generation, along with their synonyms and inflections. This approach relies on the LLM to either (a) find similar but non-synonymous replacement words or (b) completely rephrase the sentence to exclude these words. For example, "*Microsoft*, the *giant* tech company, ..." could be paraphrased as "A leading corporation in the technology sector...".

3.4.3 Suppression.

The most severe type of operation is applied to common nouns, other proper nouns, modifiers and named entities annotated as **highly concerning**, and to pronouns that are either **medium concerning** or **highly concerning**. We assume these words are either too unique or cannot be generalized.

- For common nouns and other proper nouns, dependent phrases are omitted (e.g., "We traveled *to the London Bridge* in a bus." becomes "We traveled in a bus.").
- Modifiers are removed (e.g., "He used to be the *principal* dancer" becomes "He used to be a dancer").
- Named entities are replaced with nondescript phrases (e.g., "Barack Obama" becomes "certain person").
- Pronouns are replaced with "somebody" (e.g., "*He* drove the bus." becomes "Somebody drove the bus.").

3.5 Automatic Level of Concern (LvC) Estimation

In our whistleblowing context, we deem the detection of outside-document LvC via search engine queries, as proposed by Papadopoulou et al. [44] (refer to related work in 2.3), impractical. This is because whistleblowers are typically not well-known, and the information they disclose is often novel, not commonly found on the internet. Therefore, instead of relying on external data, we focus on inner-document LvC, setting up a rule-based system and allowing users to adjust the LvC based on their contextual knowledge. Further, we assume that this pre-annotation of default concern levels raises awareness for potential sources of re-identification.

- Common nouns and modifiers, by default, are **potentially concerning**. As fundamental elements in constructing a text's semantic understanding, they could inadvertently reveal re-identifying details like profession or location. However, without additional context, their LvC is not definitive.
- Other proper nouns, unexpected words, misspelled words and out-of-vocabulary words default to **medium concerning**. Unlike categorized named entities, other proper nouns only indirectly link to individuals, places, or organizations. Unexpected words may diminish anonymity, according to Papadopoulou et al. [44], while misspelled or out-of-vocabulary words can be strong stylometric indicators.
- Named entities are considered **highly concerning** by default, as they directly refer to specific entities in the world, like people, organizations, or locations, posing a significant re-identification risk.

4 IMPLEMENTATION

Our semi-automated text sanitization tool consists of a sanitization pipeline (Sections 4.1 and 4.2) and a user interface (Section 4.3). The pipeline uses off-the-shelf Python NLP libraries (*spaCy*, *nltk*, *lemminflect*, *constituent_treelib*, *sentence-transformers*) and our paraphrasing-tuned FLAN T5 language model. FLAN T5's error-correcting capabilities [39, 42] aid in reconstructing sentence fragments after words or phrases with elevated levels of concern have been removed. The user interface is built with standard HTML, CSS, and JavaScript. Both components are open source and on GitHub[8].

4.1 Anonymization Operations for Words and Phrases

4.1.1 Generalization.

Common nouns undergo generalization by first retrieving their synsets and hypernyms from WordNet, followed by calculating the cosine similarity of their sentence embeddings with those of the hypernyms. This calculation ranks the hypernyms by semantic similarity to the original word, enabling the selection of the most suitable replacement. By default, we select the closest hypernym. **Other proper nouns** are generalized as follows: We first query Wikipedia to identify the term, using the *all-mpnet-base-v2* sentence transformer to disambiguate its meaning through cosine similarity. Next, we find the most relevant Wikidata QID and its associated hierarchy. We then flatten these relationships and replace the entity with the next higher-level term in the hierarchy.

4.1.2 Perturbation.

We add randomness to **modifiers** and **named entities** through LLM-based paraphrasing, specifically, by using the FLAN-T5 language model, which we fine-tuned for paraphrase generation (Section 4.2). To achieve perturbation[9], we give the tokens in question and their synonyms and inflections zero weight during next token prediction. This forces the model to either use a less probable word (controlled by the *temperature* hyperparameter) or rephrase the sentence to omit the token. Using a LLM for paraphrase generation has the added benefit that it mends fragmented sentences caused by token suppression and yields a neutral writing style, adjustable through the *no_repeat_ngram_size* hyperparameter.

[7] The current version of our tool considers only adjectives and adverbs as modifiers.

[8] https://github.com/dimitristaufer/Semi-Automated-Text-Sanitization

[9] The strategies "suppression" and "generalization" are straightforward adaptations of the classical methods for structured data. Perturbation "replaces original values with new ones by interchanging, adding noise or creating synthetic data" [7]. Interchanging would create ungrammatical texts, and noise can only be added to certain data. We, therefore, generate synthetic data via LLM-Rephrasing, disallowing the highly specific words / terms and their synonyms while producing a new but grammatical text.

4.1.3 Suppression. `Common nouns` and `other proper nouns` are suppressed by removing the longest phrase containing them with the *constituent_treelib* library. Sentences with just one noun or proper noun are entirely removed. Otherwise, the longest phrase, be it a main clause, verb phrase, prepositional phrase, or noun phrase, is identified, removed, and replaced with an empty string. `Modifiers` are removed (e.g., "He is their principal dancer" → "He is their · dancer"). `Pronouns` are replaced with the static string "somebody". For example, "His apple" → "Somebody apple" (after replacement) → "Somebody's apple" (after paraphrase generation). `Named entities` are replaced with static phrases based on their type. For example, "John Smith sent her 2 Million Euros from his account in Switzerland" → "certain person sent somebody certain money from somebody account in certain location" (after suppressing pronouns and named entities) → "A certain individual sent a specific amount of money to whoever's account in some particular place" (after paraphrase generation).

4.2 Paraphrase Generation

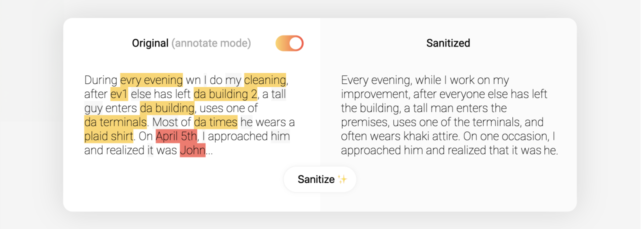
We fine-tuned two variants of the FLAN T5 language models, FLAN T5$_{\text{Base}}$ and FLAN T5$_{\text{XL}}$, using the "chatgpt-paraphrases" dataset, which uniquely combines three large paraphrasing datasets for varied topics and sentence types. It includes question paraphrasing from the "Quora Question Pairs" dataset, context-based paraphrasing from "SQuAD2.0", and summarization-based paraphrases from the "CNN-DailyMail News Text Summarization" dataset. Furthermore, it was enriched with five diverse paraphrase variants for each sentence pair generated by the *gpt-3.5-turbo* model, resulting in 6.3 million unique pairs. This diversity enhances our model's paraphrasing capabilities and reduces overfitting.

For training, we employed Parameter-Efficient Fine-Tuning (*PEFT*) using *LoRA* (Low-Rank Adaptation), which adapts the model to new data without the need for complete retraining. We quantized the model weights to enhance memory efficiency using *bitsandbytes*. We trained FLAN T5$_{\text{Base}}$ on a NVIDIA A10G Tensor Core GPU for one epoch (35.63 hours) on 1 mio. paraphrase pairs, using an initial learning rate of 1e-3. After one epoch, we achieved a minimum Cross Entropy loss of 1.195. FLAN T5$_{\text{XL}}$ was trained for one epoch (22.38 hours) on 100,000 pairs and achieved 0.88.

For inference, we configure `max_length` to 512 tokens to cap the output at T5's tokenization limit. `do_sample` is set to `True`, allowing for randomized token selection from the model's probability distribution, enhancing the variety of paraphrasing. Additionally, parameters like `temperature`, `no_repeat_ngram_size`, and `length_penalty` are adjustable via the user interface, providing control over randomness, repetition avoidance, and text length.

4.3 User Interface

Our web-based user interface communicates with the sanitization pipeline via *Flask* endpoints. It visualizes token LvCs (gray, yellow, red), allows dynamic adjustments of these levels, and starts the sanitization process. Moreover, a responsive side menu allows users to select the model size and tune hyperparameters for paraphrasing. The main window (Figure 1) shows the original and the sanitized texts, with options for editing and annotating.



During evry evening wn I do my cleaning, after ev1 else has left da building 2, a tall guy enters da building, uses one of da terminals. Most of da times he wears a plaid shirt. On April 5th, I approached him and realized it was John...

Every evening, while I work on my improvement, after everyone else has left the building, a tall man enters the premises, uses one of the terminals, and often wears khaki attire. On one occasion, I approached him and realized that it was he.

Figure 1: The UI's main window showing the input text (left) and the sanitized text (right). We made up the input and converted it to "Internet Slang" (https://www.noslang.com/reverse) to showcase how an extremely obvious writing style is neutralized.

5 EVALUATION

We evaluate our tool quantitatively (Sections 5.1 and 5.2) and demonstrate its workings and usefulness with an example from a real-world whistleblower testimony (Section 5.3). They complement each other in that the first focuses on identification via writing style and the second two on identification via content.

5.1 Re-Identification Through Writing Style: IMDb62 Movie Reviews Dataset

5.1.1 Evaluation metrics. The large unknowns of context knowledge imply that evaluations cannot rely on straightforward measurement methods for A_{id} and E_{dt}. We, therefore, work with the following proxies.

Text-surface similarities To understand the effect of language model size and hyperparameter settings on lexical and syntactic variations from original texts, we utilize two ROUGE scores: ROUGE-L (Longest Common Subsequence) to determine to which extent the overall structure and sequence of information in the text changes. And ROUGE-S (Skip-Bigram) to measure word pair changes and changes in phrasings.

Risk Without further assumptions about the (real-world case-specific) background knowledge, it is impossible to exactly quantify the ultimate risk of re-identification (see Section 3.1). We therefore only measure the part of A_{id} where (a) the context knowledge is more easily circumscribed (texts from the same author) and (b) benchmarks are likely to generalize across case studies: the risk of re-identification based on stylometric features, measured as authorship attribution accuracy (AAA).

Utility It is also to be expected that the rewriting reduces E_{dt}, yet again it is impossible to exactly determine (without real-world case-specific background knowledge and legal assessment) whether the detail supplied is sufficient to allow for legal follow-up of the report or even only to create alarm that could then be followed up. We, therefore, measure E_{dt} utility through two proxies: a semantic similarity measure and a sentiment classifier. To estimate semantic similarity (**SSim**), we calculate the cosine similarity of both texts' sentence embeddings using the SentenceTransformer[10] Python framework. To determine the absolute sentiment score difference (**SSD**), we classify the texts' sentiment using an off-the-shelf BERT-based classifier[11] from Hugging Face Hub.

All measures are normalized to take on values between 0 and 1, and although the absolute values of the scores between these endpoints (except for authorship attribution) cannot be interpreted directly, the comparison of relative orders and changes will give us a first indication of the impacts of different rewriting strategies on A_{id} and E_{dt}.

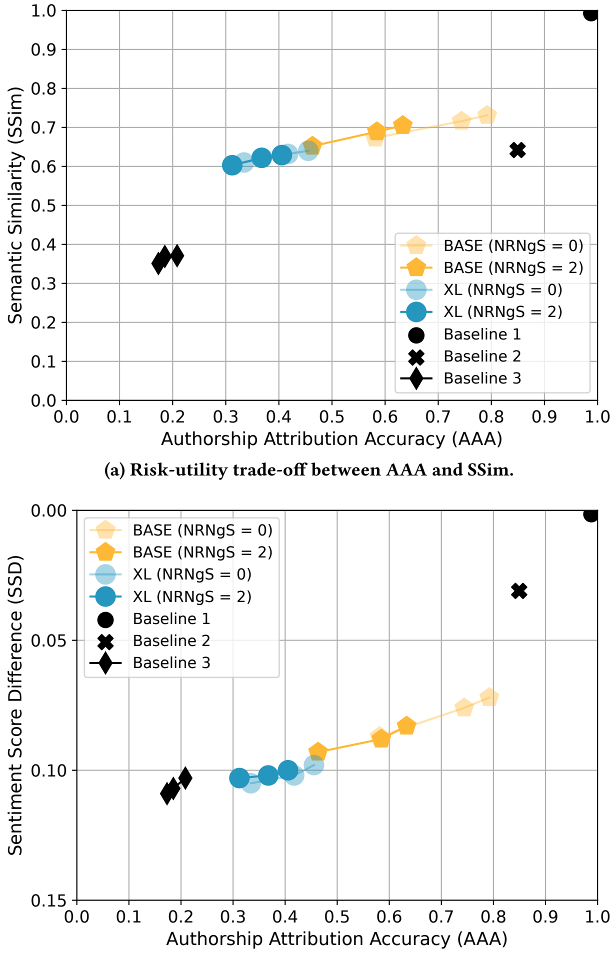
5.1.2 Data, language models, and settings. We investigate protection against authorship attribution attacks with the popular IMDb62 movie reviews dataset [59], which contains 62,000 movie reviews by 62 distinct authors. We assess AAA using the "Valla" software package [64], specifically its two most effective models: one based on character *n*-grams and the other on BERT. This approach covers both ends of the the authorship attribution spectrum [3], from low-level, largely topic-independent character *n*-grams to the context-rich features of the pre-trained BERT model.

The evaluation was conducted on AWS EC2 "g4dn.xlarge" instances with NVIDIA T4 GPUs. We processed 130 movie reviews for each of the 62 authors across twelve FLAN T5 configurations, totaling 96,720 texts with character counts spanning from 184 to 5248. Each review was sanitized with its textual elements assigned their default LvCs (see 3.5).

Both model sizes, "Base" (250M parameters) and "XL" (3B parameters) were tested with temperature values T of 0.2, 0.5, and 0.8, as well as with `no_repeat_ngram_size` (NRNgS) set to 0 or 2. The former, *temperature*, controls the randomness of the next-word predictions by scaling the logits before applying softmax, which makes the predictions more or less deterministic. For our scenario, this causes smaller or greater perturbation of the original text's meaning. The latter, NRNgS, disallows n consecutive tokens to be repeated in the generated text, which for our scenario means deviating more or less from the original writing style.

The Risk-utility trade-offs of all configurations are compared to three baselines: $Baseline_1$ is the original text. In $Baseline_2$, similar to state-of-the-art related work [24, 44], we only redact named entities by replacing them with placeholders, such as "[PERSON]" and do not utilize our language model. Similarly, in $Baseline_3$ we only remove named entities but rephrase the texts using our best-performing model configuration regarding AA protection.

5.1.3 Results. The n-gram-based and BERT-based "Valla" classifiers achieved AAA baselines of `98.81%` and `98.80%`, respectively. As expected, the AAA and text-surface similarities varied significantly depending on the model configuration. The XL-model generated texts with much smaller ROUGE-L and ROUGE-S scores, i.e. more lexical and syntactic deviation from the original texts. Using $NRNgS = 2$ slightly decreased AAA in all configurations while not significantly affecting semantic similarity, which is why we use this for all the following results.



(a) Risk-utility trade-off between AAA and SSim.

(b) Risk-utility trade-off between AAA and SSD.

Figure 2: Risk-utility trade-offs.

Figure 2 (a) shows the risk-utility trade-off between AAA and SSim. "Top-left" (0,1) would be the - fictitious - best result. For each model configuration, increasing T caused AAA to drop but also decreased utility by ~ 8%/4% (BASE/XL) for SSim and ~12%/3% (BASE/XL) for SSD. The figure shows that the investigated settings create a trade-off curve, with XL ($T = 0.8$, $NRNgS = 2$) allowing for a large reduction in AAA (to `31.22%`, as opposed to the original text $baseline_1$ of 98.81%), while BASE ($T = 0.2$, $NRNgS = 0$) retains the most SSim (0.731, as opposed to the original texts, which have $SSim = 1$ to themselves).

Figure 2 (b) shows the risk-utility trade-off between AAA and SSD (the plot shows 1-SSD to retain "top left" as the optimal point). The results mirror those of AAA-SSim, except for $baseline_2$: because only named entities (not considered sentiment-carrying) are removed, the sentiment score changes only minimally.

5.1.4 Discussion. In summary, all our models offer a good compromise between baselines representing state-of-the-art approaches. They have lower risk and higher or comparable utility compared to $baseline_3$, where only named entities are removed. This indicates

[10] all-mpnet-base-v2
[11] bert-base-multilingual-uncased

the effectiveness of LLM-based rephrasing in authorship attribution. $Baseline_2$, which involves suppressing named entities and rephrasing, shows the lowest risk due to limited content left for the LLM to reconstruct, resulting in mostly short, arbitrary sentences, as reflected by low SSim scores.

5.2 Re-Identification Through Content: European Court of Human Rights Cases

Pilán et al.'s [48] Text Anonymization Benchmark (TAB) includes a corpus of 1,268 English-language court cases from the European Court of Human Rights, in which directly- and quasi-identifying nominal and adjectival phrases were manually annotated. It solves several issues that previous datasets have, such as being "pseudo-anonymized", including only few categories of named entities, not differentiating between identifier types, containing only famous individuals, or being small. TAB's annotation is focused on protecting the identity of the plaintiff (also referred to as "applicant").

5.2.1 Evaluation Metrics. TAB introduces two metrics, entity-level recall ($ER_{di/qi}$) to measure privacy protection and token-level-weighted precision (WP_{di+qi}) for utility preservation. Entity-level means that an entity is only considered safely removed if all of its mentions are. WP_{di+qi} uses BERT to determine the information content of a token t by estimating the probability of t being predicted at position i. Thus, precision is low if many t with high information content are removed. Both metrics use micro-averaging over all annotators to account for multiple valid annotations. Because our tool automatically rephrases the anonymized texts, we make two changes. First, since we cannot reliably measure WP_{di+qi}, we fall back to our previously introduced proxies for measuring E_{dt} utility. Secondly, we categorize newly introduced entities from LLM hallucination that may change the meaning of the sanitized text.

The legal texts, which must prefer direct and commonly-known identifiers, are likely to present none or far fewer of the background-knowledge-specific re-identification challenges of our domain. Thus, again the metrics used here should be regarded as proxies.

Risk We measure A_{id} using $ER_{di/qi}$ and count slightly rephrased names of entities as "not removed" using the Levenshtein distance. For example, rephrasing "USA" as "U.S.A" has the same influence on $ER_{di/qi}$ as failing to remove "USA".

Utility We estimate E_{dt} through *SSim*. In addition, we determine all entities in the sanitized text that are not in the original text (again using the Levenshtein distance). We categorize them into (1) *rephrased harmful entities* (semantically identical to at least one entity that should have been masked), (2) *rephrased harmless entities*, and (3) *newly introduced entities.* We measure semantic similarity by calculating the cosine similarity of each named entity phrase's sentence embedding to those in the original text.

5.2.2 Data, language models, and settings. The TAB corpus comprises the first two sections (introduction and statement of facts) of each court case. For our evaluation, we use the *test* split which contains 127 cases of which each has, on average, 2174 characters (356 words) and 13.62 annotated phrases. We perform all experiments using the "XL" (3B parameter) model with temperature values T of 0.2, 0.5, and 0.8, as well as with *NRNgS* set to 2.

5.2.3 Results and Discussion. $ER_{di/qi}$ and SSim vary slightly, but not significantly for different T values. For *T = 0.2*, we get an entity-level recall on quasi-identifiers (ER_{qi}) of 0.93, which is slightly better than Pilán et al.'s [48] best performing model trained directly on the TAB corpus (0.92). However, our result for direct identifiers ER_{di} is 0.53, while theirs achieves 1.0, i.e. does not miss a single high-risk entity. Closer inspection reveals that our low results for direct identifiers come mainly from (i) the SpaCy NER failing to detect the entity type CODE (e.g. "10424/05") and (ii) the LLM re-introducing names of named entities that are spelled slightly differently (e.g. "Mr Abdisamad Adow Sufi" instead of "Mr Abdisamad Adow Sufy").

Regarding utility, all three model configurations achieve similar SSim scores ranging from 0.67 (T = 0.8) to 0.69 (T = 0.2). These results fall into the same range achieved using the IMDb62 movie reviews dataset. However, in addition to re-introducing entities that should have been masked, we found that, on average, the LLM introduces 5.24 new entities (28.49%) per court case. While some of these, depending on the context, can be considered harmless noise (e.g. "European Supreme Tribunal"), manual inspection revealed that many change the meaning and legitimacy of the sanitized texts. For example, 4.7% contain names of people that do not appear in the original text, 43.3% contain new article numbers, 20.5% contain new dates, and 11.8% include names of potentially unrelated countries.

The frequency of such hallucinations could also be a consequence of the specific text genre of court cases, and future work should examine to what extent this also occurs in whistleblower testimonies and how it affects the manual post-processing over the generated text that is previewed in our semi-automated tool.

5.3 Re-Identification Through Content: Whistleblower Testimony Excerpts

We further investigated our tool's rewritings of two excerpts (Tables 3, 4) from a whistleblower's hearing in the Hunter Biden tax evasion case, as released by the United States House Committee on Ways and Means.[12] This qualitative view on our results provides for a detailed understanding of which identifiers were rewritten and how.[13]

5.3.1 Approach. First, we compiled the essential E_{dt} upon which we based our analysis on. Next, we assessed the textual features in both excerpts to enhance our tool's automatic Level of Concern (LvC) estimations, aiming for the lowest author identifiability (A_{id}). Finally, we input these annotations into the user interface to produce the rewritings.

5.3.2 E_{dt} and A_{id}. Based on the information from the original texts in tables 3 and 4 alone, we define E_{dt} as follows, with E_{dt1}, E_{dt2} being a subset of excerpt 1 and E_{dt3} a subset of excerpt 2.

$$E_{dt} := \begin{cases} \text{"The Tax Division approved charges but for no apparent} \\ \text{reason changed their decision to a declination.",} \\ \text{"The declination occurred after significant effort was put} \\ \text{into the investigation by the whistleblower.",} \\ \text{"In their effort in doing what is right, the whistleblower} \\ \text{suffered on a professional and personal level."} \end{cases}$$

[12] https://waysandmeans.house.gov/?p=39854458 [Accessed 29-April-2024], "#2"

[13] To answer these questions, it is immaterial whether the text sample describes a concrete act of wrongdoing (as in our fictitious Ex. 1) or not (as here).

In *exc*1 (Table 3), we classified “joining the case” (first-person indexical) and implications of a nation-wide investigation as highly concerning. Additionally, we marked all “case” mentions as highly concerning to evaluate consistent suppression. “DOJ Tax”, being a stylometric identifier because it is no official abbreviation, received a medium LvC, and “thousands of hours” was similarly categorized, potentially indicating the authors role as lead in the case.

In *exc*2 (Table 4), we classified the lexical identifier “2018”, which could be cross-referenced relatively easily, as well as all descriptive identifiers concerning the author’s sexual orientation and outing as highly concerning. Furthermore, emotional descriptors (“sleep, vacations, gray hairs, et cetera”) are given medium LvC, similar to references of case investment (“thousands of hours” and “95 percent”), mirroring the approach from *exc*1.

5.3.3 Results and Discussion. $Exc1_{sanitized}$ retains E_{dt2}, but not E_{dt1}, as “DOJ Tax” is replaced with “proper noun” due to the non-existence of a corresponding entity in Wikidata. Consequently, it defaults to the token’s POS tag. For A_{id}, all identified risks were addressed (e.g., “considerable time” replaces “thousands of hours.”). However, the generalization of “case” led to inconsistent terms like “matter”, “situation”, and “issue” due to the $NRNgS = 2$ setting. This is beneficial for reducing authorship attribution accuracy but may confuse readers not familiar with the original context.

$Exc2_{sanitized}$ maintains parts of E_{dt3}, though terms like “X amount of time” and “Y amount of the investigation” add little value due to their lack of specificity. Notably, “amount o of” represents a rare LLM-induced spelling error, underscoring the need for human editing for real-world use. The emotional state’s broad generalization to “physical health, leisure, grey body covering” is odd and less suitable than a singular term would be. Despite this, $Exc2_{sanitized}$ effectively minimizes A_{id} by addressing all other identified risks.

Table 3: LvC-annotated whistleblower testimony exc_1 (excerpt 1) with identifiers (top) and $exc1_{sanitized}$ (bottom).

Original: “**Prior** to joining the **case**, **DOJ Tax** had approved tax charges for the **case** and the **case** was in the process of progressing towards indictment [...] After working **thousands of hours** on that captive **case**, poring over evidence, interviewing **witnesses** all over the **U.S.**, the decision was made by **DOJ Tax** to change the approval to a declination and not charge the **case**.”
Lexical IDs: DOJ Tax; U.S.
Indexical IDs: [implicit: me] joining the case (first person)
Descriptive IDs: interviewing witnesses all over the U.S. (nation-wide investigation); thousands of hours (author involvement)

Sanitized: “The proper noun had approved tax charges for the matter and the situation was moving towards indictment, but after spending considerable time on that captive matter, poring over evidence, the decision was made by proper noun to defer the approval and not charge the issue.”

Table 4: LvC-annotated whistleblower testimony $exc2$ (excerpt 2) with identifiers (top) and $exc2_{sanitized}$ (bottom).

Original: “I had opened this investigation in **2018**, have spent **thousands of hours** on the case, worked to complete **95 percent** of the investigation, have sacrificed **sleep**, **vacations**, **gray hairs**, **et cetera**. **My husband** and I, in identifying **me** as the **case agent**, were **both publicly** outed and ridiculed on social media due to **our sexual orientation**.”
Lexical IDs: 2018; thousands of hours; 95 percent
Indexical IDs: me as the case agent (role of author); My husband (author’s marital status)
Descriptive IDs: I had opened this investigation in 2018 (can be cross-referenced); My husband and I + publicly outed and ridiculed [...] due to our sexual orientation (author’s sexual orientation and public event); sacrificed sleep, [...], gray hairs (emotional state)

Sanitized: “I had opened this investigation on a certain date, had spent X amount of time on the case, worked to complete Y amount of the investigation, sacrificing my physical health, leisure, grey body covering, etc.”

6 CONCLUSIONS, LIMITATIONS AND FUTURE WORK

We evaluated our our tool’s effectiveness using ECHR court cases and excerpts from a real-world whistleblower testimony and measured the protection against authorship attribution attacks and information loss statistically using the popular IMDb62 movie reviews dataset. Our method can significantly reduce authorship attribution accuracy from 98.81% to 31.22%, while preserving up to 73.1% of the original content’s semantics, as measured by the established cosine similarity sentence embeddings. Our qualitative analysis revealed that minor wording changes significantly impact A_{id} and E_{dt}, and highlighted our tool’s strengths in reducing A_{id} through generalization, perturbation, and suppression.

Our tool’s usefulness in real-world whistleblowing scenarios remains to be tested, particularly with human users. Challenges arise from the possibility of the tool introducing unrelated entities through model hallucination and its limitations in addressing complex syntactic structures and co-references. Still, our LLM-based approach has proved to be promising in matters of counteracting the limitations of state-of the art approaches. The fine-tuned model effectively reduces authorship attribution and improves text coherence – two of the main shortcomings of previous works. At the same time, it introduces novel challenges, such as limited control over the accuracy and consistency of the rephrased content.

Future work will focus on refining our tool through evaluations involving human participants and domain experts. Given the crucial importance of context knowledge for re-identification risks and the challenges in identifying all textual features that contribute to re-identification, future work will also pay increasing attention to enhancing anonymization awareness. This would not only apply to the whistleblowing use case, but extend to the protection of free speech in other areas too, including journalism, political activism, and social media.

We envision an interactive awareness tool as a more dynamic alternative to conventional static writing guides on whistleblowing platforms. This tool would incorporate insights from our research as well as insights from practitioners, aiming to educate users about subtle textual nuances that could pose re-identification risks, thereby creating a deeper understanding and more effective use of anonymization practices in high-risk disclosures. At the same time, we need to draw on practitioners' and legal experts' knowledge to better understand what textual changes are detrimental (or conducive) to utility and incorporate these insights into the guidance provided by the awareness tool.

7 ETHICAL CONSIDERATIONS, RESEARCHERS POSITIONALITY, AND POSSIBLE ADVERSE IMPACTS

In the following paragraphs, we discuss five key challenges, interweaving a potential adverse impacts statement, an ethical considerations statement (what we have done or can do), and positionalities.

We are computer scientists (some of us with a background also in social and legal sciences) who have programming expertise (instrumental for mitigating challenges C1–C4), understanding of data protection law (C1), research expertise in bias and fairness, including methods for risk mitigation when working with LLMs (C2), and collaborators with human-subjects studies expertise (C3). None of us has been a whistleblower. We outline below how future collaborators and/or deployers with other positionalities can contribute relevant complementary expertise on C1–C5.

C1 – Data Protection: Our tool does not collect or store any user data. Original as well as re-written texts are discarded after each run, and they are not used to train the model further. Our tool does not require an internet connection beyond the initial downloading of pre-trained language models and optional queries to Wikidata servers. While querying Wikidata enhances the efficacy of our tool by enabling the generalization of certain words, users should be aware that these queries might expose confidential information to external servers. To mitigate this risk, our implementation remains functional when offline, albeit with slightly reduced efficacy due to the lack of real-time Wikidata look-ups. In a real-life deployment, technical and organizational measures would need to be implemented in order to safeguard the confidential personal or organizational data that remain in the reports; this will also require security and legal expertise.

C2 – Bias and (Un-)fairness: Our tool may inadvertently introduce or perpetuate biases present in the training data. FLAN T5 was trained on C4, which is generated from the April 2019 Common Crawl dataset. Dodge et al. [13] discovered that C4 has a "negative sentiment bias against Arab identities" and excludes "documents associated with Black and Hispanic authors" as well as documents "mentioning sexual orientations" [p. 8] by its blocklist filter. Therefore, similar to other pre-trained models [32], FLAN T5 is "potentially vulnerable to generating equivalently inappropriate content or replicating inherent biases" [8, p. 52]. This may bias our level of concern measures. For example, certain names, professions, or locations may be classified as "medium concerning" or "highly concerning" more often because they are considered "surprising", which may unfairly impact the narratives involving them. Future work should, therefore, include evaluating and mitigating these biases and possibly experiments with other datasets and pre-trained models.

C3 – Over-Reliance and Retaliation: The results of our quantitative evaluation are promising, but an extensive qualitative evaluation is necessary to determine whether our approach translates to real-world situations. Therefore, users of our tool must remain aware of its potential to alter the original intent of their text significantly and, depending on the context, possibly offer limited protection against retaliation. Over-reliance on our tool may lead to a false sense of security, resulting in increased vulnerability to retaliation. We intend to assess the extent of this form of automation bias [10] in a subsequent user study, discuss with people who are working in the field (e.g., whistleblower protection activists) how to best reduce it, and also evaluate these future mitigation measures.

C4 – Resource consumption: Training LLMs is resource-intensive. By re-using the existing model and enlisting distilled LLM learning, this impact could be reduced in future work.

C5 – Tool Misuse: Even though our tool aims to mitigate the risk of whistleblower re-identification, malicious actors might misuse our tool for obfuscating dangerous information or illegally converting copyrighted material. By providing our source code and fine-tuned models publicly, we open avenues for ethical use and misuse alike. Therefore, we emphasize that our sole aim in developing our tool is to facilitate legal, ethical whistleblowing. Future refinements and real-world evaluations will require collaboration with legal and social experts to better understand the practical implications and potential misuse scenarios.

ACKNOWLEDGMENTS

We acknowledge funding from the German Federal Ministry of Education and Research (BMBF) – Nr 16DII134.

This publication has been partially supported by the EXDIGIT (Excellence in Digital Sciences and Interdisciplinary Technologies) project, funded by Land Salzburg under grant number 20204-WISS/263/6-6022.

REFERENCES

[1] Vigjilenca Abazi. 2020. The European Union whistleblower directive: a 'game changer'for whistleblowing protection? *Industrial Law Journal* 49, 4 (2020), 640–656.

[2] Allison Adams, 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.* Linköping University Electronic Press, Linköping, 1–7.

[3] Malik Altakrori, Jackie Chi Kit Cheung, and Benjamin CM Fung. 2021. The topic confusion task: A novel evaluation scenario for authorship attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2021.* Association for Computational Linguistics, Punta Cana, Dominican Republic, 4242–4256. https://doi.org/10.18653/v1/2021.findings-emnlp.359

[4] Iyadh Ben Cheikh Larbi, Aljoscha Burchardt, and Roland Roller. 2022. Which anonymization technique is best for which NLP task?–It depends. A Systematic Study on Clinical Text Processing. *arXiv e-prints* (2022), arXiv–2209.

[5] Bettina Berendt and Stefan Schiffner. 2022. Whistleblower protection in the digital age-why "anonymous" is not enough.: From technology to a wider view of governance. *The International Review of Information Ethics* 31, 1 (2022).

[6] Rachelle Bosua, Simon Milton, Suelette Dreyfus, and Reeva Lederman. 2014. Going public: Researching external whistleblowing in a new media age. In *International handbook on whistleblowing research.* Edward Elgar Publishing, 250–272.

[7] Can Eyupoglu Can, Muhammed Ali Aydin, Abdul Halim Zaim, and Ahmet Sertbas. 2018. An Efficient Big Data Anonymization Algorithm Based on Chaos and Perturbation Techniques. *Entropy* 20, 5 (2018), 373. article no.: 373; https://www.mdpi.com/1099-4300/20/5/373.

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).

[9] Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. 2019. AI Now 2019 Report. https://ainowinstitute.org/publication/ai-now-2019-report-2

[10] Mary Cummings. 2004. Automation Bias in Intelligent Time Critical Decision Support Systems. In *Proc. of the AIAA 1st Intelligent Systems Technical Conference*. doi:10.2514/6.2004-6313.

[11] Rita de Sousa Costa and Inês de Castro Ruivo. 2020. Preliminary Remarks and Practical Insights on How the Whistleblower Protection Directive Adopts the GDPR Principles. In *Privacy Technologies and Policy - 8th Annual Privacy Forum, APF 2020, Lisbon, Portugal, October 22-23, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12121)*, Luís Antunes, Maurizio Naldi, Giuseppe F. Italiano, Kai Rannenberg, and Prokopios Drogkaris (Eds.). Springer, 95–109. https://doi.org/10.1007/978-3-030-55196-4_6

[12] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association* 24, 3 (2017), 596–606.

[13] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758* (2021).

[14] Josep Domingo-Ferrer. 2007. A three-dimensional conceptual framework for database privacy. In *Secure Data Management: 4th VLDB Workshop, SDM 2007, Vienna, Austria, September 23-24, 2007. Proceedings 4*. Springer, 193–202.

[15] Josep Domingo-Ferrer, David Sánchez, and Jordi Soria-Comas. 2016. Database anonymization: privacy models, data utility, and microaggregation-based inter-model connections. *Synthesis Lectures on Information Security, Privacy, & Trust* 8, 1 (2016), 1–136.

[16] Cynthia Dwork. 2006. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*. Springer, 1–12.

[17] European Data Protection Supervisor (EDPS) and Spanish Data Protection Agency (AEPD). 2021. 10 Misunderstandings Related to Anonymisation. https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf. Joint statement on anonymisation of personal data according to EU GDPR.

[18] Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA: BERT fine-tuning for Authorship Attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*. 127–137.

[19] Ronen Feldman and James Sanger. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

[20] Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *Principles of Security and Trust: 8th International Conference, POST 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6–11, 2019, Proceedings 8*. Springer International Publishing, 123–148.

[21] Christian Hauser, Nadine Hergovits, and Helene Blumer. 2019. Whistleblowing Report 2019. http://whistleblowingreport.org/

[22] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2019. *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*. Technical Report. IEEE. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

[23] Alistair EW Johnson, Lucas Bulgarelli, and Tom J Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*. 214–221.

[24] Bennett Kleinberg, Toby Davies, and Maximilian Mozes. 2022. Textwash–automated open-source text anonymisation. *arXiv preprint arXiv:2208.13081* (2022).

[25] Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and PG Demidov. 2019. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*. IEEE, 184–195.

[26] Alaor Leite. 2021. Whistleblowing und das System der Rechtfertigungsgründe Das erlaubte Whistleblowing nach dem Geschaftsgeheimnisgesetz als, fürdernder Rechtfertigungsgrund". *Goltdammer's Archiv für Strafrecht* 168, 3 (2021), 129–146.

[27] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2006. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*. IEEE, 106–115.

[28] Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4188–4203.

[29] Yinhan Liu, 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* (2019).

[30] Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics* 75 (2017), S34–S42.

[31] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3–es.

[32] Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2022. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing* (2022).

[33] Tanya M Marcum and Jacob Young. 2019. Blowing the whistle in the digital age: are you really anonymous? The perils and pitfalls of anonymity in whistleblowing law. *DePaul Bus. & Comm. LJ* 17 (2019), 1.

[34] Brian Martin. 2003. Illusions of whistleblower protection. *UTS L. Rev.* 5 (2003), 119.

[35] Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The Limits of Word Level Differential Privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 867–881.

[36] Andrew WE McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman, and Rachel Greenstadt. 2012. Use fewer instances of the letter "i": Toward writing style anonymization. In *Privacy Enhancing Technologies: 12th International Symposium, PETS 2012, Vigo, Spain, July 11-13, 2012. Proceedings 12*. Springer, 299–318.

[37] Joseph McGlynn III and Brian K Richardson. 2014. Private support, public alienation: Whistle-blowers and the paradox of social support. *Western Journal of Communication* 78, 2 (2014), 213–237.

[38] Brijesh Mehta, Udai Pratap Rao, Ruchika Gupta, and Mauro Conti. 2019. Towards privacy preserving unstructured big data publishing. *Journal of Intelligent & Fuzzy Systems* 36, 4 (2019), 3471–3482.

[39] Gayani Nanayakkara, Nirmalie Wiratunga, David Corsar, Kyle Martin, and Anjana Wijekoon. 2022. Clinical dialogue transcription error correction using Seq2Seq models. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*. Springer, 41–57.

[40] Arvind Narayanan and Vitaly Shmatikov. 2010. Myths and fallacies of" personally identifiable information". *Commun. ACM* 53, 6 (2010), 24–26.

[41] Janet P Near and Marcia P Miceli. 1985. Organizational dissidence: The case of whistle-blowing. *Journal of business ethics* 4, 1 (1985), 1–16.

[42] Hoang Nguyen and Sandro Cavallari. 2020. Neural multi-task text normalization and sanitization with pointer-generator. In *Proceedings of the First Workshop on Natural Language Interfaces*. 37–47.

[43] Annika Willoch Olstad, Anthi Papadopoulou, and Pierre Lison. 2023. Generation of Replacement Options in Text Sanitization. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. 292–300.

[44] Anthi Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022. Neural Text Sanitization with Explicit Measures of Privacy Risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*. 217–229.

[45] Constantinos Patsakis and Nikolaos Lykousas. 2023. Man vs the machine: The Struggle for Effective Text Anonymisation in the Age of Large Language Models. *arXiv preprint arXiv:2303.12429* (2023).

[46] Andreas Pfitzmann and Marit Hansen. 2005. *Anonymity, unlinkability, unobservability, pseudonymity, and identity management-a consolidated proposal for terminology – v. 0.28*. Technical Report. https://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.28.pdf

[47] David J. Phillips. 2004. Privacy policy and PETs: the influence of policy regimes on the development and social implications of privacy enhancing technologies. *New Media and Society* 6, 6 (2004), 691—-706.

[48] Ildikó Pilán, Pierre Lison, Lilja Ovrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics* 48, 4 (2022), 1053–1101.

[49] Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*. 17–26.

[50] Joyce Rothschild and Terance D Miethe. 1999. Whistle-blower disclosures and management retaliation: The battle to control information about organization corruption. *Work and occupations* 26, 1 (1999), 107–128.

[51] Mary Saade. 2023. Women & Whistleblowing. *Hastings Journal on Gender and the Law* 34, 1 (2023), 43.

[52] Shikha Sachdeva and Narendra Singh Chaudhary. 2022. Exploring whistleblowing intentions of Indian nurses: a qualitative study. *International Journal of Organizational Analysis* ahead-of-print (2022).

[53] Pierangela Samarati and Latanya Sweeney. 1998. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* Technical Report. Harvard Data Privacy Lab. https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf

[54] David Sánchez and Montserrat Batet. 2016. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology* 67, 1 (2016), 148–163.

[55] David Sánchez, Montserrat Batet, and Alexandre Viejo. 2012. Detecting sensitive information from textual documents: an information-theoretic approach. In *Modeling Decisions for Artificial Intelligence: 9th International Conference, MDAI 2012, Girona, Catalonia, Spain, November 21-23, 2012. Proceedings 9*. Springer, 173–184.

[56] David Sánchez, Montserrat Batet, and Alexandre Viejo. 2013. Automatic general-purpose sanitization of textual documents. *IEEE Transactions on Information Forensics and Security* 8, 6 (2013), 853–862.

[57] David Sánchez, Montserrat Batet, and Alexandre Viejo. 2014. Utility-preserving privacy protection of textual healthcare documents. *Journal of biomedical informatics* 52 (2014), 189–198.

[58] Kim R Sawyer, Jackie Johnson, and Mark Holub. 2010. The necessary illegitimacy of the whistleblower. *Business & Professional Ethics Journal* (2010), 85–107.

[59] Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics* 40, 2 (2014), 269–310.

[60] Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. A4nt: author attribute anonymity by adversarial training of neural machine translation. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*. 1633–1650.

[61] Lauren M Stuart, Saltanat Tazhibayeva, Amy R Wagoner, and Julia M Taylor. 2013. On identifying authors with style. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 3048–3053.

[62] Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of biomedical informatics* 58 (2015), S11–S19.

[63] Marie Terracol. 2019. Building on the EU directive for whistleblower protection: analysis and recommendations. (2019).

[64] Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. 2022. On the state of the art in authorship attribution and authorship verification. *arXiv preprint arXiv:2209.06869* (2022).

[65] Özlem Uzuner, Tawanda C Sibanda, Yuan Luo, and Peter Szolovits. 2008. A de-identifier for medical discharge summaries. *Artificial intelligence in medicine* 42, 1 (2008), 13–35.

[66] Meghan Van Portfliet and Kate Kenny. 2022. Whistleblowing advocacy: Solidarity and fascinance. *Organization* 29, 2 (2022), 345–366.

[67] Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2021. Progress in machine translation. *Engineering* (2021).

[68] Benjamin Weggenmann and Florian Kerschbaum. 2018. Syntf: Synthetic and differentially private term frequency vectors for privacy-preserving text mining. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 305–314.

[69] Martin Weingardt. 2004. *Fehler zeichnen uns aus: Transdisziplinäre Grundlagen zur Theorie und Produktivität des Fehlers in Schule und Arbeitswelt*. Julius Klinkhardt.

[70] Simone White. 2018. A matter of life & death: whistleblowing legislation in the EU. In *Eucrim: The European Criminal Law Associations' Forum*, Vol. 3. 170–177.

[71] Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*. 247–257.

[72] Shubham Yadav, Santosh Singh Rathore, and Satyendra Singh Chouhan. 2020. Authorship Identification Using Stylometry and Document Fingerprinting. In *Big Data Analytics: 8th International Conference, BDA 2020, Sonepat, India, December 15–18, 2020, Proceedings 8*. Springer, 278–288.

[73] Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman SM Chow. 2021. Differential privacy for text analytics via natural text sanitization. *arXiv preprint arXiv:2106.01221* (2021).

[74] Ying Zhao and Jinjun Chen. 2022. A survey on differential privacy for unstructured data content. *ACM Computing Surveys (CSUR)* 54, 10s (2022), 1–28.