Unsupervised utility evaluation of text anonymization methods via neural language models

Benet Manzanares-Salor[a], David Sánchez[a], Pierre Lison[b]

[a]*Department of Computer Engineering and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, ComSCIAM-Center for Computational Science and Applied Mathematics, Universitat Rovira i Virgili, Av. Paisos Catalans 26, Tarragona, 43007, Spain*
[b]*Norwegian Computing Center, Postboks 114 Blindern, Oslo, NO-0314, Norway*

Abstract

Text anonymization methods strive to find a balance between privacy protection and utility preservation, where the latter refers to the fact that the anonymized documents are still analytically useful for research or business tasks. The performance of these methods is evaluated empirically, by comparing their outputs with human-based anonymizations through the standard precision and recall metrics. Whereas recall is used as a proxy for the level of privacy attained, precision is interpreted as the amount of utility preserved. Nonetheless, these metrics were not designed for evaluating privacy-oriented tasks and present several drawbacks. First, they assume a unique ground truth whereas, in text anonymization, several masking choices can be equally valid to prevent re-identification. Second, the human annotations used as ground truth are inherently subjective and prone to errors. Third, both metrics weight terms uniformly, thereby ignoring the varied impact that the terms' semantics have on utility and re-identification risk. To overcome these limitations, in this paper we present the first unsupervised utility metric for anonymized texts. Our metric relies on neural language models to quantify the utility loss incurred by anonymization. Empirical experiments on document clustering show that our metric captures the actual utility of the anonymized outcomes more accurately than precision, while being more sensitive to varying anonymization intensities. Together with a previously

Email addresses: benet.manzanares@urv.cat (Benet Manzanares-Salor), david.sanchez@urv.cat (David Sánchez), plison@nr.no (Pierre Lison)

proposed privacy metric, our proposal defines a complete evaluation framework for text anonymization that does not require costly human annotations. Using this framework, we also report a comprehensive evaluation of a variety of text anonymization methods.

Keywords: Text anonymization, Utility evaluation, Neural language models, Information content

---

1. Introduction

Text constitutes the standard encoding for most unstructured or semi-structured knowledge sources, which makes it essential for many research tasks [1, 2, 3, 4, 5]. Nevertheless, text often contains (sensitive) personal information that, in accordance with privacy regulations such as the European General Data Protection Regulation (GDPR), should be protected prior to releasing it or making it available for secondary use. Privacy protection can be achieved via anonymization, a process by which data should no longer be attributed to specific subjects. The latter renders the data no longer personal and, therefore, outside the scope of the GDPR.

Data anonymization has been widely employed to protect structured (tabular) datasets [6, 7]. However, unstructured data such as plain text present significant challenges [8, 9, 10, 11]. The main difficulty is that the potentially re-identifying attributes that may appear in a document are unbounded and, often, not obviously associated with the corresponding subject. As a consequence, automatic text anonymization methods are significantly less mature than those meant for structured datasets [8]. Most of these methods rely on natural language processing (NLP) techniques to detect and mask terms that could lead to re-identification [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30], such as personal identifiers, addresses or names.

Due to the heuristic nature of text anonymization methods, their performance should be empirically evaluated *ex post*. The quality of the anonymization is measured as the balance between the level of privacy attained and the degree of utility preserved [6]. On the one hand, privacy is inversely proportional to disclosure risk. On the other hand, data utility can be understood as the amount of meaningful information preserved in the anonymized outcomes, and it is the inverse of the information loss incurred by anonymization. For example in case of a retail surveillance camera, privacy protection means that the identity of monitored people cannot be disclosed, and utility refers

to the ability of algorithms to still being capable to recognize the behavior of monitored people to, *e.g.*, detecting shoplifting. Whereas privacy protection is a legal requirement that should be fulfilled and its enforcement necessarily entails a loss of information, keeping data analytically useful for secondary use is the main goal behind data releases. Therefore, anonymization methods strive to find a balance between privacy and utility.

In the statistical disclosure control field (SDC) [6], the evaluation of anonymization methods is performed by empirically measuring the residual disclosure risk and the statistical utility of the anonymized data. The privacy assessment is conducted by exposing the anonymized data to re-identification attacks, such as record linkage [31], while preserved utility is quantified by comparing the original data with their anonymized counterparts [7]. Utility metrics include measuring the difference (or error) between the original and anonymized versions of individual records or between relevant statistical features of the dataset, such as attribute means, variances, or covariances.

In contrast, in the literature on text anonymization, privacy and utility evaluation rely on manual annotations that are used as the anonymization ground truth [12, 13, 15, 17, 19, 20, 21, 22, 24, 26, 32]. In particular, the information retrieval (IR) metrics *precision* and *recall* are employed to measure the performance of text anonymization. Precision corresponds to the proportion of terms masked by the system that were also masked in the ground truth, and it is interpreted as a measure of utility preservation: the less unnecessarily masked terms, the more information/utility is preserved. Conversely, recall assesses the proportion of re-identifying terms (again, defined by the ground truth) that have been detected, and it is understood as the inverse to re-identification risk or, in order words, as a measure of privacy protection. Nevertheless, these metrics have severe limitations as evaluation means for text anonymization. First, the fact that these metrics are meant for IR tasks in which a single best solution is assumed, contrasts with the fact that, in data anonymization, multiple masking choices –each one involving a different combination of terms– could be equally valid in preventing re-identification [8]. Secondly, the manual annotations that are used as ground truth are prone to errors, omissions, and inconsistencies [33]. Finally, while precision and recall uniformly account for the contribution of terms, not all missed/unnecessarily masked terms contribute equally to the disclosure risk/utility loss [34].

1.1. Contributions and plan

To overcome the limitations of precision-based utility evaluation, in this paper we present the first unsupervised utility metric for text anonymization. Our utility assessment is based on an information theoretic measure we also propose here, which estimates the semantics encompassed by (original or masked) texts. On this basis, our utility metric quantifies the difference between the semantics encompassed by the original documents and their anonymized versions, similarly to how utility is accurately assessed on structured data in the SDC literature.

We demonstrate the effectiveness of our proposal by assessing the level of utility preserved by several text anonymization methods on a common dataset, and by comparing it with the practical utility of their anonymized outcomes when applied to an fundamental analytical task: document clustering. The results show that our metric has a significantly stronger correlation with the practical utility of anonymized texts than the standard precision metric. This makes it a more accurate utility metric, with the additional benefit of not requiring manual annotations.

This work complements the privacy metric we presented in [35], thereby offering a complete framework for evaluating text anonymization. By leveraging this framework, we also report a comprehensive evaluation of a variety of text anonymization methods. Compared to precision and recall, our metrics provide a more accurate assessment of the attained privacy/utility balance, and highlight the intrinsic limitations not only of NLP-based methods, but also of manual anonymizations.

The remainder of the paper is organized as follows. Section 2 provides background on (text) anonymization methods, discusses related works on privacy and utility evaluation, and introduces the neural language models on which our approach relies. In Section 3, we propose and formalize our metric. Section 4 reports and discusses the empirical results obtained by applying our metric on a variety of anonymization methods, and compares it with the standard precision metric. Section 5 offers a comprehensive evaluation of all the anonymization methods considered, reporting their empirical privacy, and comparing their privacy/utility balance. Finally, Section 6 gathers the conclusions and depicts several lines of future work.

2. Background

2.1. Anonymization

Anonymization consists of the complete and irreversible removal of all information that could directly or indirectly be used to re-identify the individuals referred in a dataset [36].

The attributes that can potentially lead to re-identification can be categorized into *identifiers*, which are unique values of an individual such as her complete name or passport number, and *quasi-identifiers*, which are publicly known information (typically demographic and geospatial), such as gender, zip code, or birth date. While identifiers allow re-identification in isolation, quasi-identifiers should be combined with other quasi-identifiers present in the same dataset in order to define a unique (*i.e.*, potentially re-identifying) sequences of values. The identified public knowledge or data sources containing (quasi-)identifiers of an individual are known as *background knowledge*. Whereas identifiers are limited, any identified publicly available information on an individual can act as a quasi-identifier, thereby making quasi-identifying attributes virtually unlimited.

Proper anonymization therefore requires the complete removal of all identifiers and the masking or selective removal of quasi-identifiers. Simply removing identifiers and pre-defined sets of quasi-identifiers does not qualify as anonymization –because it does not effectively prevent re-identification–, and is therefore considered de-identification [8, 37].

Because structured data are organized in records with predefined attributes, this simplifies the task of categorizing (and masking) identifying and quasi-identifying values. In contrast, spotting (quasi-)identifiers in unstructured text is challenging because i) (quasi-)identifying terms appearing in a document are unbounded, and ii) quite often, they are not obviously associated to the individuals they refer to. To tackle this challenge, text anonymization methods rely on (suboptimal) heuristics to detect and mask (quasi-)identifiers and therefore can hardly offer formal privacy guarantees [8].

2.2. Privacy evaluation

Due to the lack of formal privacy guarantees, the level of privacy protection attained by text anonymization methods should be evaluated *ex post*, that is, through empirical evaluations of the residual disclosure risk of the protected outcomes. Because most text anonymization methods rely

on named-entity recognition (NER) models and are evaluated by means of IR-based metrics, the latter have been adopted as the *de facto* standard to assess anonymized outputs. In particular, privacy evaluation is performed as is done for completeness in NER tasks, that is, by quantifying the *recall* [12, 13, 15, 17, 19, 20, 21, 22, 24, 26, 32]. Recall is defined as the proportion of elements of interest that were masked by the system to be evaluated with respect to a ground truth:

$$Recall = \frac{\#TruePositives}{\#TruePositives + \#FalseNegatives} \tag{1}$$

where #TruePositives is the number of elements detected by the system and #FalseNegatives is the number of missed ones. In text anonymization, these elements correspond to (quasi-)identifying words or text spans that should be masked according to human experts' opinions, which are considered the ground truth.

Nevertheless, recall-based privacy evaluation presents several drawbacks [34, 38]. On one hand, it assumes that all identified/missed elements contribute equally to mitigate/increase re-identification risk, which is rarely the case [34]. Indeed, failing to mask an identifier (such as the family name of an individual) is much more disclosive than just missing the job or whereabouts. On the other hand, recall is not privacy metric *per se* and just quantifies the agreement between the system's output and the (often subjective and error-prone) human annotations.

A privacy-oriented alternative to recall-based evaluation was presented in [35]. This approach, named *Text Re-Identification Risk* (TRIR), relies on a re-identification attack designed for anonymized text, which follows a similar principle as the record linkage attack used to evaluate privacy risks in structured data [31]. Specifically, a classifier model tailored for re-identification is trained on the background knowledge available on the individuals referred in the documents to be protected. Then, the re-identification accuracy of the attack on the anonymized texts is used to quantify their residual disclosure risk, as follows:

$$TRIR = \frac{\#ReIdentifications}{|A_I|} \tag{2}$$

where #ReIdentifications is the number correct re-identification achieved by the attack and A_I is the set of individuals/documents in the protected

dataset. This method avoids the shortcomings of recall-based evaluation and does not resort to human annotations [35].

2.3. Utility evaluation

Being the complementary of recall, the *precision* metric is used in the literature to quantify the utility preserved in the anonymized outcomes [12, 13, 15, 17, 19, 20, 21, 22, 24, 26, 32]. Precision calculates the ratio of correctly masked terms by the anonymization method:

$$Precision = \frac{\#TruePositives}{\#TruePositives + \#FalsePositives} \tag{3}$$

where #TruePositives is the number of text spans that were correctly masked by the anonymization method with respect to the ground truth, and #FalsePositives is the number of unnecessarily masked ones.

Most of the criticisms raised against recall-based privacy assessment also apply to precision-based utility evaluation. First, precision does not provide a comprehensive figure of the utility retained in the anonymized document, because the semantics of correctly masked terms, even though being masked in the anonymized outcome, are not considered in the utility assessment. Moreover, precision does not explicitly measure text's analytical utility, but just counts the number of (unnecessarily) masked terms. This assumes that every unnecessarily masked term contributes equally to utility loss when, in fact, the utility and readability of an anonymized document can be more or less impaired depending on the semantics conveyed by each masked term and the context in which it appeared. Finally, IR-based metrics assume the existence of a single ground truth. Although this assumption may be reasonable for NER or other NLP-oriented tasks, it does not hold for text anonymization, where several masking choices (each encompassing a different combination of quasi-identifying terms) could be equally valid to prevent re-identification [8]. An extension of the standard precision metric, which weights the contribution of terms' semantics, was proposed in [34]. Nevertheless, it still suffers from the issues derived from the use of manual annotations as ground truth.

In contrast, in the SDC area [6], the assessment of data utility accounts for the actual analytical utility that one can expect from the anonymized output, and it is properly approached by measuring the distributional and statistical differences between the original and anonymized datasets. These differences can be computed at the record level or by using abstract characterizations of the datasets. In the first case, the original attribute values

of each record are individually contrasted with their anonymized versions, thereby measuring the information loss incurred by each modification made by the anonymization method (*e.g.*, generalization, suppression, noise addition, etc.). On the other hand, considering abstract characterizations includes measuring how data distributions and their properties (*e.g.*, attribute means or variances, covariances, etc.) have been modified as a result of the anonymization process. To measure the divergence between the original and anonymized attribute values, a variety of dissimilarity metrics can be employed according to the attributes' data types. Numeric values can be compared with the well-known mean square error, the mean absolute error, or mean variation metrics, whereas categorical values require more complex methods such as contingency tables, Shannon's entropy [39, 40] or ontology-based metrics [41, 42, 43].

2.4. Neural language modeling

Our method extensively leverages neural language models, which have recently revolutionized the NLP field. In the following we provide background on these models that is needed to understand the choices made in our proposal.

A language model aims to learn the intrinsic structure of languages, modeling the meaning of words and the structure of sentences and documents. To this end, language models are pre-trained in an unsupervised fashion on a large corpus of documents (from *e.g.*, Wikipedia, book collections and other online sources), so that they capture a broad range of text types and linguistic constructions. The resulting language models are assumed to encapsulate general knowledge, being suitable for transfer learning.

Although building these large pre-training models is costly, they are usually freely available for research, industrial, or personal use [44]. Language models are typically based on the transformer architecture [45, 46, 47, 48]. Compared to previous neural architectures based on recurrent layers [49, 50, 51], transformer models can handle long-range dependencies between tokens while simultaneously allowing tokens to be processed in parallel (which is crucial when handling longer texts). This is achieved in practice through a self-attention mechanism by which the vector representation of each token is expressed as a weighted sum of a linear transformation of the vectors from the preceding layer. The weights of this sum correspond to attention scores expressing the relevance of each token in the surrounding context for the representation of the currently processed token. This makes it possible

to efficiently derive contextualized vector representations (contextual word embeddings) for each token of a given sequence. Those contextual word embeddings are computed through multiple neural layers, each one being represented by its own set of parameters. The model parameters are optimized in an self-supervised fashion (what is called pre-training) from tasks such as masking a randomly selected word and subsequently predicting its value (*masked language modelling*), predicting the next word in a sequence (*causal language modelling*) or determining whether two sentences follow each other in a document (*next sentence prediction*).

One of the most well-established transformer-based language models is BERT (Bidirectional Encoder Representations from Transformers) [45]. After fine-tuning, BERT has been shown to reach or surpass human-level performance in multiple language-related tasks. Many variations of BERT have been proposed that improve its performance [52] or reduce the size of the model and, therefore, its computational requirements [53]. All these language models are freely available at Hugging Face's transformers library [44] with pre-trained weights and an easy-to-use implementation.

3. An information content-based utility metric for anonymized texts

In this section, we present our utility metric, named *Text Preserved Information* (TPI). It aims to provide a practical evaluation means to assess the ability of anonymization methods to preserve text's utility/semantics, and do so without resorting to manual annotations.

Given that semantics play the most critical role in determining text utility [54], our metric assesses utility as the difference between the semantics encompassed by the original document and its anonymized version. In the field of computational linguistics, the semantics encompassed by textual terms are often measured according to the information content (IC) they convey [55]. The IC of a term is defined as the inverse of its probability of occurrence. Therefore, terms with the lowest probability (consequently, the highest IC) are the least predictable and thus provide the most significant information/semantics/utility in the context in which they appear. In the NLP literature, several methods have been proposed to calculate the IC of textual terms [55, 32, 56]. In this paper, we propose a novel and efficient method, named *Text Information Content* (TIC), which leverages neural language models to estimate the term's probabilities required for IC calculation and that is free from the limitations and dependencies of existing IC calculation methods.

3.1. Formalization

Let O be the original and unmodified document to be protected. This document contains an ordered list of N terms, $O = [t_1, \ldots, t_N]$ (including term repetitions). Let P be the protected version of O. P also consists of an ordered list of M terms, $P = [t_1, \ldots, t_M]$. Note that the number of terms in P might differ from that in O due to term suppressions performed by the anonymization method.

As introduced above, to measure the semantics/utility of a term t, we leverage its information content (IC) [55]:

$$IC(t) = -log(prob(t)) \tag{4}$$

Details on the calculation of terms' probability ($prob(t)$) will be provided in Section 3.2.

To measure the utility of a document D, either original (O) or anonymized (P), we define its *Text Information Content* (TIC) as the sum of the IC of all the terms it contains:

$$TIC(D) = \sum_{t \in D} IC(t) \tag{5}$$

Then, to measure the proportion of utility from O that has been preserved in P after the anonymization process, we propose the *Text Preserved Information* (TPI) metric, which is computed as the ratio of their respective TICs:

$$TPI(P, O) = \frac{TIC(P)}{TIC(O)} \tag{6}$$

TPI provides an intuitive metric of the amount of O's utility preserved in P, which ranges from 0 (*i.e.*, P has no utility at all) to 1 (*i.e.*, the utility of O is perfectly preserved in P). Consistently with the masking operations performed by anonymization methods, TPI will never be greater than 1 because P's utility is necessarily a subset of O's. We propose the division operation for TPI rather than subtraction (which is also commonly employed to measure information loss), because the latter would produce non-normalized absolute values that would make comparability difficult. In fact, TPI's [0..1] range allows a direct comparison with the standard precision metric (Eq. 3) and the calculation of F-scores when combined with any normalized privacy metric. The latter will enable to assess the global performance of an anonymization method and the balance between privacy and utility.

Notice, however, that the meaning of TPI values are different to those of precision: whereas $precision = 1$ means that the anonymized output perfectly matches the manual ground truth, $TPI = 1$ means that O has not been modified at all. Consequently, precision just gives an idea of the agreement between an automated anonymization and a particular manual anonymization (among the many other valid anonymizations that can be conceived). In contrast, TPI provides an assessment of the actual utility preserved in the anonymized outcomes, which is consistent with the (statistical) utility metrics employed in the SDC area for structured databases [6]. Our metric can be seen as an adaptation of this methodology where documents correspond to records, TIC serves as an abstract characterization of the utility the data convey, and TPI computes the dissimilarity between the TIC of the original and anonymized documents, giving practitioners a clearer idea of the actual utility they can expect. The practical differences between TPI and precision will be empirically evaluated in Section 4.1.

3.2. Probability calculus

Gathering accurate term probabilities is crucial for TIC calculation and, subsequently, for the TPI (Eq. 6) metric. In the following, we discuss a variety of approaches that can be employed to calculate these probabilities, highlight their limitations, and propose our own calculation method based on neural language modeling.

Traditionally, the probability of a term has been empirically measured as its relative frequency in a corpus [55]. This approach requires a well-designed strategy for consistently counting the occurrences of terms at a conceptual level, which implies disambiguating polysemous words and aggregating occurrences of synonyms. Moreover, the corpus should be large and representative enough so that the distribution of terms captures the way they are used in society. This is challenging given the number of rare words (*e.g.*, names, addresses) that should be properly represented and that, due to their rareness, are the usual target of anonymization. In order to address this data sparseness issue, some authors have used the Web as corpus [57, 32]. To gather probabilities efficiently, terms are queried into a web search engine, and the number of results returned relative to the number of webs indexed by the search engine are interpreted as probabilities. The most significant drawbacks of this approach are scalability and cost, because the number of queries to be performed to a search engine and the time required to obtain the results could be very large for long documents.

To avoid the cost of corpus-based approaches, other authors rely on the so-called intrinsic IC calculation models. These models derive term probabilities from their specificity, which is estimated from their number of hyponyms and/or hypernyms in a structured knowledge base (*e.g.*, ontology) [58, 56]. Intrinsic IC models assume that the less hyponyms and the more hypernyms a term has, the more specific (and less probable) the term is. The shortcoming of this approach is its dependence on the availability of a large enough knowledge base that covers (ideally all) the terms present in the documents to assess. This is hardly achievable in practice because static knowledge bases usually model concepts (*i.e.*, abstractions) rather than instances (*i.e.*, real-world entities), the latter being, due to their specificity, the most privacy-sensitive terms.

Moreover, both corpus-based and intrinsic IC models ignore the fact that the informativeness of terms depends on their context. In particular, words acquire meaning through their combination with other words. This is known as *compositionality*, and states that the meaning of a complex expression is determined by the meanings of its constituent terms and the rules used to combine them. Therefore, understanding the semantics of a term appearing in a context (and estimating its probability of appearance within that context) requires analyzing the relationships of the term with its context. For example, the probability of the noun "penguin" is higher in the sentence "I saw a penguin in Antarctica" than in the sentence "There is a penguin in my room". In contrast, the aforementioned ontology-based and corpus-based IC calculations assess terms in isolation.

With the goal of computing IC efficiently and more accurately than existing approaches, we leverage advances in neural language models to obtain contextualized term probabilities. As mentioned in Section 2.4, language models pre-trained on huge corpora contain general linguistic knowledge, which include the distribution of terms within sentences. Models pre-trained on the *masked language modeling* (MLM) task learn to predict masked terms based on their context.

Based on this principle, we propose replacing each target term with a "[MASK]" token, feeding the masked sentence into the MLM model, and using the model's predicted probabilities for these masked terms as their occurrence likelihoods. For example, to determine the probability of the word "juice" in the sentence "I drink apple juice" the model is provided with the input "I drink apple [MASK]". The model then generates a probability distribution across its entire vocabulary for that masked slot, and the specific

value assigned to the "juice" token is selected as the final likelihood. Suppose the model assigns a high probability value of 85% to the token "juice". This would indicate that the token is predictable from the context and, therefore, that it has a low information content (*i.e.*, Eq. 4). Conversely, consider the sentence "I eat an apple *sandwich*" where the model assigns a low probability value of 1% to the token "sandwich", suggesting that "sandwich" is unexpected and carries high information content in this context.

In brief, this method can provide context-aware probability estimations based on the general knowledge extracted from pre-trained models. Moreover, the approach is more practical and feasible than the aforementioned methods, because it does not require manually-engineered resources such as ontologies or costly web queries. In this respect, we can employ the wide variety of language models pre-trained on the MLM task, such as BERT [45], which are freely available at [44].

However, using this methodology to obtain the IC of all the terms in a document (as it is required for TIC) presents some challenges and requires an adequate definition of what a term is and a context masking strategy. In the following, we describe the challenges to be tackled and our solution in detail.

First, we need to consider that the straightforward approach of directly using the language model's tokens as terms would lead to incorrect results. This is because tokens do not usually correspond to the units of text that convey a complete meaning, such as words or chunks, but to smaller subparts. For instance, if the WordPiece tokenization [59] in which BERT relies on is employed, terms such as "New York" or "looking" would be split into two tokens: "New"+"York" and "look"+"#ing", respectively. Therefore, when predicting a token-defined term, parts of the word or chunk may be available in the context, leading to overestimated probabilities. For example, for the chunk "New York", it would be easy for the model to predict the token "New" having "York" in the context, and vice versa. This results in the probability of the text span "New York" being higher than it should. On this basis, we propose to perform a chunk preprocessing, where we use words as terms except for those belonging to a chunk, which will be interpreted as a single term. For instance, in the sentence "I was in New York", the corresponding terms will be: "I", "was", "in" and "New York". From the point of view of the language model, this approach may result in terms that will be tokenized into more than one token. For these multi-token terms, we take the minimum probability of the term's tokens as the term probability.

The intuition is that the probability of the term (chunk) as a whole cannot be higher than the probability of the least probable component token. An alternative to taking the minimum probability would be to multiply the token probabilities. However, this significantly overestimates the informativeness of longer chunks, because it assumes that their component tokens are independent (whereas words rarely are [60]).

Secondly, replacing all the terms to be assessed by the "[MASK]" tag at once is not suitable because the model will have no context to perform the prediction of each term. Instead, we propose masking context terms iteratively, that is, we mask a subset of the terms at each iteration (leaving the rest untouched as context), get the probabilities of those terms, and repeat the process on other non-previously masked terms until all of them have been assessed. We propose two strategies to this end: *Sentence-Term Alternating* (S-TA) and *N-Term Alternating* (N-TA). In S-TA, only one term per sentence is masked at each iteration, keeping the remainder of the sentence as context. In this way, the model receives as input one sentence with only a single term masked at a time. In Algorithm 1, a detailed description of the procedure is depicted. On the other hand, in N-TA, one term out of each N is masked, keeping the remainder (including surrounding terms and terms from other sentences) as context. Algorithm 2 provides a step-by-step description of this strategy. In this case, the model receives the maximum number of tokens possible (*e.g.*, 512 for BERT), with approximately $1/N$ of those terms uniformly masked. Due to the common input length limitation of language models, the document is split according to the maximum number of supported tokens. Notice that this splitting is not usually required for S-TA, because sentences rarely reach the maximum length.

Intuitively, S-TA is expected to provide more reliable probabilities than N-TA with a small N, because it keeps more immediate context around the masked terms, that is, the sentence in which they appear. In the best case, N-TA can be equivalent to S-TA if N corresponds to the number of terms in the sentence. However, nta has its advantages because it can estimate the probabilities of all terms with fewer iterations, and this results in fewer model predictions, which are the costlier operation of the process. More specifically, whereas S-TA predictions scale with the number of terms per sentence and the number of sentences, the number of predictions required for N-TA depends on N and the number of document splits needed. For instance, in the best case for S-TA, where there is a single sentence as long as a split (*i.e.*, 512 tokens for BERT), S-TA will require as many predictions

Algorithm 1 Sentence-Term Alternating (S-TA)

Require: $terms$
$termsProbs \leftarrow Vector[terms.Length]$
for $(sentence,\ startIndex)$ **in** $sentenize(terms)$ **do**
    for i **in** $[1, \ldots, sentence.Length]$ **do**
        $maskedSentence \leftarrow copy(sentence)$
        $maskedSentence[i] \leftarrow$ "$[MASK]$"
        $probs \leftarrow MLM(maskedSentence)$
        $termsProbs[startIndex + i] \leftarrow probs[i].forTerm(sentence[i])$
    end for
end for
return $termsProbs$

Algorithm 2 N-Term Alternating (N-TA)

Require: $terms,\ N$
$termsProbs \leftarrow Vector[terms.Length]$
for i **in** $[1, \ldots, N]$ **do**
    $maskedTerms \leftarrow copy(terms)$
    $maskedIndices \leftarrow \emptyset$
    for j **in** $[i, \ldots, terms.Length]$ **do**
        if $(j - i) \bmod N = 0$ **then**
            $maskedTerms[j] \leftarrow$ "$[MASK]$"
            $maskedIndices.insert(j)$
        end if
    end for
    $probs \leftarrow MLM(maskedTerms)$
    for k **in** $maskedIndices$ **do**
        $termsProbs[k] \leftarrow probs[k].forTerm(terms[k])$
    end for
end for
return $termsProbs$

as sentence terms, while N-TA will need N predictions. Subsequently, N-TA with an N lower than the average number of sentences' terms will be faster than S-TA.

The behavior of both strategies is depicted in Figure 1 in a sample text.

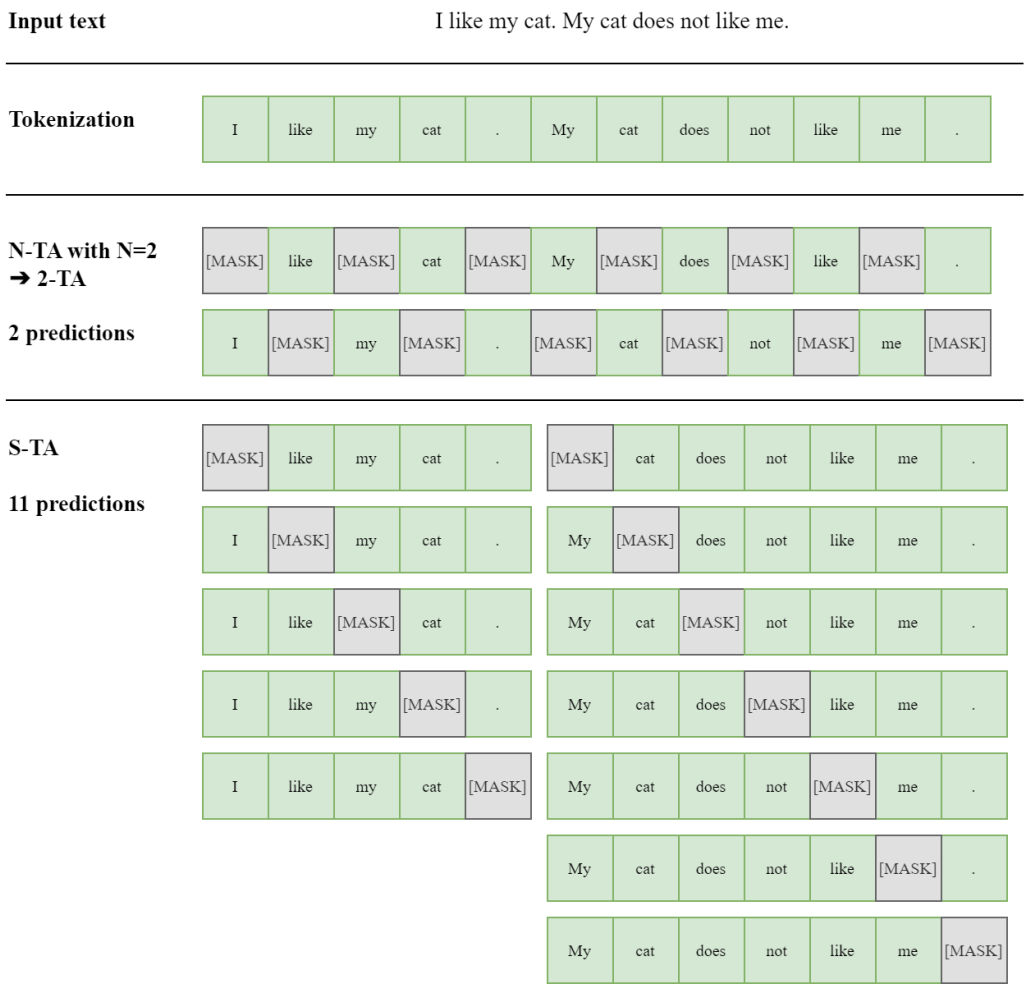


Figure 1: Comparison between the two proposed context masking strategies, N-TA and S-TA, applied to a two-sentence sample text. N-TA uses $N = 2$ (2-TA), while S-TA operates at sentence level. For each strategy, the strings to be processed by the language model –partitioned per token and including the masks– are represented. Notice the lower computational cost of 2-TA relative to S-TA due to the lower number of required model predictions.

Employing either of these two strategies, the probability of all terms in the

original document O can be obtained. However, if the same approach is applied to the protected document P, where some terms may have been replaced by privacy-preserving masks, the probabilities of these masks will probably be overestimated, resulting in a TIC value that can be even higher than the TIC of O. This is because privacy-preserving masks are typically non-meaningful suppression tags (*e.g.*, "***") or abstract generalizations (*e.g.*, NER classes names such as "PERSON" or "LOCATION"). From the point of view of the language model, these terms are outliers and their assessment would result in a too low probability (and, subsequent, too high IC). For instance, because a language model trained on clear text does not expect the word "PERSON" in the privacy-protected sentence "PERSON is an actor", it will likely consider "PERSON" less probable than the original actor's name, thereby resulting in an inconsistent IC assessment.

To circumvent this issue, we consider privacy-preserving masks as non-informative terms with a fixed probability of 1 (hence, zero IC). This assumes that the masking operation applied to sensitive terms was suppression, which is the most common strategy in the literature due to the difficulty of obtaining suitable and utility-preserving generalizations to replace sensitive terms [61]. Due to this assumption, the *useful* terms in the protected document P can be defined as a strict subset of the terms in the original document O, that is, P contains the terms in O minus the protected ones. Consequently, we can compute the TIC of P according to the probabilities of the term calculated for O simply by subtracting the IC of the protected terms from $TIC(O)$ to get $TIC(P)$. Equation 7 formalizes this computation, where $IsProtected(t)$ is a binary function that returns 1 if the term t is protected and 0 otherwise.

$$TIC(P) = TIC(O) - \sum_{t \in P} IsProtected(t) \cdot IC(t) \tag{7}$$

By incorporating this calculus into Eq. 6, we obtain the following detailed expression for TPI:

$$TPI(P, O) = 1 - \frac{\sum_{t \in P} IsProtected(t) \cdot IC(t)}{TIC(O)} \tag{8}$$

4. Experiments on utility evaluation

In the following, we report empirical results on the application of TPI to several text anonymization methods on a common corpus of documents.

To offer a broad coverage of methods, we considered a variety of approaches as well as a reference manual anonymization. Finally, we also compare TPI to the standard precision metric, and measure the correlation between the utility assessment of both metrics and the practical utility of the anonymized outcomes observed in a document clustering task.

As experimental data, we used the privacy-oriented dataset introduced in [62]. It consists of 553 abstracts of short Wikipedia biographies, which have been randomly chosen from the set of biographies presented in [63]. Wikipedia biographies are commonly used in the text anonymization literature as evaluation data, because their high density of (quasi-)identifying terms makes them a challenging anonymization scenario [8, 32, 57, 64]. The random sampling resulted in a wide variety of biography domains, ranging from football players to World War II generals. This configures a more general and larger corpus than most privacy evaluation datasets employed in the literature, which are usually domain-specific (*e.g.*, medical texts [19, 12, 26, 22]) and/or comparatively small [61]. The 553 abstracts have been manually anonymized by several human annotators by following the guidelines described in [62].

4.1. Anonymization methods

Most automated approaches to text anonymization use sequence labeling to identify text spans belonging to predefined categories (named entities such as names, locations or dates), which may lead to re-identification [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27]. These categories are detected by handcrafted rules or machine learning, which define a named entity recognition (NER) model. To anonymize text, detected entities are removed or masked by replacing them with their corresponding category label. We considered the following NER-based tools that have been employed in the literature to anonymize texts [61, 8, 34]:

- **Stanford NER** [65]: offers three pre-trained NER models: NER3, which is able to detect LOCATION, ORGANIZATION and PERSON types; NER4, which detects LOCATION, ORGANIZATION, PERSON and MISC (miscellaneous) types; and NER7, which detects LOCATION, ORGANIZATION, PERSON, MONEY, PERCENT, DATE and TIME.

- **spaCy NER**[1]: we employed the en_core_web_lg neural model trained on the Ontonotes v5 corpus [66], which is capable of detecting named entities of GPE (*e.g.*, countries, cities), LOC (non-GPE locations such as mountain ranges), ORG, PERSON, MONEY, DATE, PERCENT, CARDINAL, TIME (times smaller than a day), EVENT (*e.g.*, wars), FAC (*e.g.*, buildings), LANGUAGE, LAW (named documents made into laws), NORP (nationalities or religious or political group), ORDINAL, PRODUCT (*e.g.*, vehicles), QUANTITY and WORK_OF_ART (*e.g.*, books titles, songs) types.

- **Microsoft Presidio**[2]: an anonymization-specific tool based on NER. Among the diversity of entity types supported by Presidio, we enabled those corresponding to quasi-identifying information: LOCATION, PERSON, NRP –person's nationality, religious or political group– and DATE_TIME.

Once the tool and category types are fixed, NER-based approaches produce a *fixed* protected output, which cannot be tailored to specific privacy/utility requirements due to the lack of privacy-oriented parameters.

In addition to NER-based approaches, a few methods have been proposed in the privacy-preserving data publishing (PPDP) area. In contrast to the former methods, they explicitly consider the notion of disclosure risk, thereby not limiting the type of information that can act as quasi-identifier. Moreover, as is the case for anonymization methods meant for structured databases, they also allow to configure the desired balance between privacy and utility via parameters. Although several methods exist, most of them can hardly be applied in practice, because they either only focus on masking (and assume that sensitive terms have been manually detected) [67, 68, 69], or present scalability issues arising from their dependency on external resources, such as web search engines [32, 70, 64]. As an exception to the former, we picked up two recent and practical PPDP-based approaches that can be directly applied to our experimental setting.

- **Disclosure protection based on word embeddings** [61]: This approach takes advantage of a word embedding model, such as Word2Vec [71], to detect text spans that are semantically related to the individual

[1] `https://spacy.io/api/entityrecognizer`
[2] `https://github.com/microsoft/presidio`

to be protected. The embedding model is unsupervisedly trained from scratch by using the documents to be protected and, if that corpus is small, by complementing it with an external related corpus. Because word embeddings capture the semantic relationships between terms, it is expected that the words closely related to the individual to be protected would have embeddings similar to the embedding of the individual's name because both have a tendency to co-occur. Therefore, words whose embeddings have a cosine similarity with the individual's name higher than a user-defined threshold are considered (quasi-)identifiers and are subsequently masked. The similarity threshold t allows balancing privacy and utility according to the data holder's priorities and risk tolerance. In Section 4.5, three distinct values for this hyperparameter are assessed: $t = 0.5$ (the median), along with $t = 0.25$ and $t = 0.75$, which capture equidistant points on the similarity spectrum. For the rest of the experiments, $t = 0.5$ is adopted as a neutral, intermediate default value, enabling an evaluation of the method's performance without specifically favoring privacy or utility. Although a meaningful generalization-based approach relying on ontologies is proposed, in order to ensure a fair comparison with the other considered methods (all of them based on suppressions or abstract generalizations), here we evaluate it as if sensitive terms were suppressed. In our experiments, we trained the Word2Vec model by using the 553 biographies to be protected and an external corpus of 19,000 Wikipedia biographies of movie actors that was employed in the experiments of the original paper. The public implementation of the method[3] has been employed to run the experiments. In the following, this method will be referred to as *Word2Vec*.

- **K-anonymity-based on knowledge graphs** [62]: This method exploits a knowledge graph capturing the distribution of terms in a large corpus (*i.e.*, Wikidata[4]) to detect terms or text spans that appear in less than k documents. According to this, the method enforces k-anonymity [72] on the set of documents to be protected by incrementally suppressing terms until no combination of terms is present in less than K documents. The masking can be carried out in a *greedy* manner, where

[3] `https://github.com/fadiabdulf/automatic_text_anonymization`
[4] `https://www.wikidata.org/`

the least frequent term is masked first, or through a *random* approach, where the next sensitive term to be masked is arbitrarily selected. The greedy strategy is expected to result in a lower –but not necessarily minimum– number of masked terms, which would contribute to better retain utility. In the following, the two versions of this method will be referred to as *k-anonymity greedy* and *k-anonymity random*. Since the authors of this paper also proposed the dataset we employ in our experiments, we used their results (obtained with a privacy parameter of $k = 5$) 'as is'.

In addition to the automatic methods listed above, we also consider the manual anonymization conducted in [62]. These annotations are available in the paper's repository[5]. The annotation process was performed as follows. First, the annotators found terms disclosing personal information, and secondly, they decided which of those terms ought to be masked to conceal the identity of the individual in the biography. This process followed carefully designed guidelines, similar to those used in [34]. For the first step, NER-like pre-annotations obtained with spaCy were provided, but annotators were instructed to use them only as a starting point and make corrections as needed. They were also responsible for assigning a semantic category to the spotted terms (*e.g.*, person, location, organization). In the second step, annotators had to classify the detected text spans into direct identifiers, quasi-identifiers or non-identifiers, the latter being left in clear. Finally, the detected (quasi-)identifiers were replaced by their semantic category, as usually done in NER-based anonymization. The documents were annotated by 5 annotators (four undergraduate law students and one NLP researcher) using the TagTog tool[6]. In addition to assessing the utility resulting from these manual annotations, we also used them as ground truth for computing the precision and recall metrics of the methods described above.

4.2. Results

We next report on the results of the application of TPI to the above-described methods. The calculation of TIC of both original and anonymized documents is implemented as described in Section 3 (Eq. 8), including the chunking preprocessing (using spaCy's noun chunks) and the S-TA and N-TA

[5] `https://github.com/anthipapa/bootstrapping-anonymization`
[6] `https://docs.tagtog.com/`

Table 1: Summary of anonymization methods categorized into NER-based, PPDP-oriented, and manual approaches. The right column specifies the target entity categories for NER models, configuration parameters for PPDP techniques, and the manual annotation setup.

Method	**Categories/Parameters**
Standord NER3	LOCATION, ORGANIZATION, PERSON
Standord NER4	LOCATION, ORGANIZATION, PERSON, MISC
Standord NER7	LOCATION, ORGANIZATION, PERSON, MONEY, PERCENT, DATE, TIME
spaCy NER	GPE, LOC, ORG, PERSON, MONEY, DATE, PERCENT, CARDINAL, TIME, EVENT, FAC, LANGUAGE, LAW, NORP, ORDINAL, PRODUCT, QUANTITY, WORK_OF_ART LOC, ORG, PERSON, DATE, CARDINAL, EVENT, GPE, FAC, LANGUAGE, LAW, MONEY, NORP, ORDINAL, PERCENT, PRODUCT, TIME, QUANTITY, WORK_OF_ART
Microsoft Presidio	LOCATION, PERSON, NRP, DATE_TIME
Word2Vec	$t = 0.5$ as cosine similarity threshold
k-anonymity Greedy	$k = 5$ with a greedy suppression strategy
k-anonymity Random	$k = 5$ with a random suppression strategy
Manual annotation	5 annotators; guidelines from [34]; spaCy-based pre-annotations

strategies. Unless otherwise stated, TPI is based on S-TA strategy because, as we will discuss later, it is able to provide the most accurate TIC assessment. Term probabilities were obtained using the BERT language model [45] with its standard pre-training for MLM on Wikipedia and the BookCorpus [73]. TPI calculation was performed on the cloud-powered IDE Google Colaboratory[7]. All experiments were executed in a stable environment consisting of an Intel Xeon CPU, 12GB of RAM and an Nvidia Tesla T4 GPU

[7] `https://colab.research.google.com/`

with 16GB of VRAM. This hardware is more than enough for the utility assessment of the 553 documents because no (costly) training process is involved.

Figure 2 reports the TPI values for the different anonymization methods and compares them with the standard precision metric. The error bars correspond to the standard deviation for each metric and method in all the documents. The *average* column represents the mean of all automatic methods (*i.e.*, it does not consider manual anonymization).

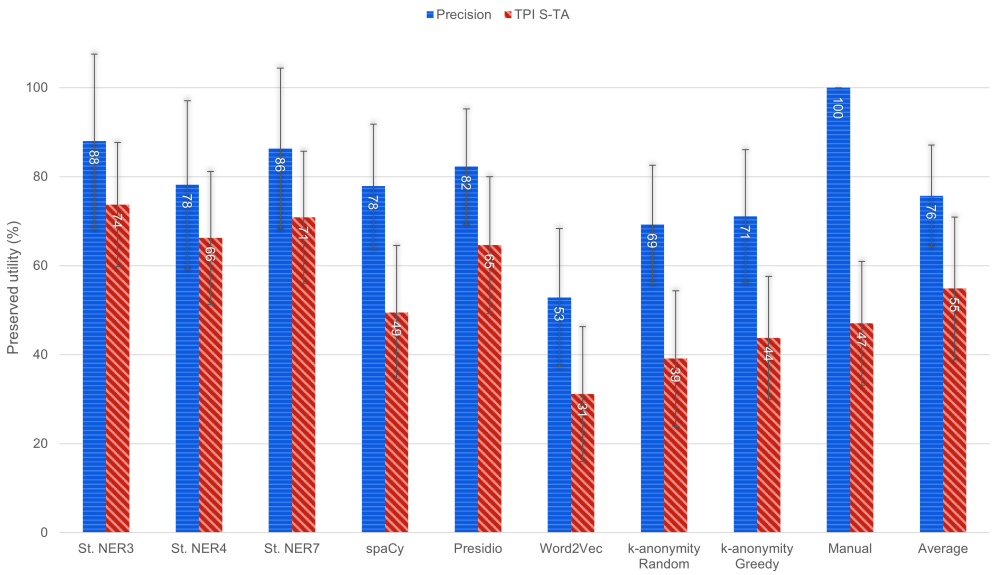


Figure 2: TPI and precision across anonymization methods. Blue bars indicate precision, while red bars represent TPI using the S-TA context masking strategy. Variability bars show the standard deviation among documents. The "Average" bar aggregates performance across all methods (excluding manual annotations, which are the ground truth for precision). Higher values are better.

The results show that TPI is highly correlated with precision. Concretely, for automatic anonymization methods (it does not make sense to consider the manual anonymization since its precision will always be 100%), their Pearson coefficient is $r = 0.923$. This correlation establishes TPI as a robust method for assessing preserved utility without requiring costly human annotations. The fact that precision values are deterministically higher than those of TPI is consistent with the claim made in Section 3.1: while the upper bound for precision represents a perfect agreement between automatic

and manual anonymizations, for TPI it means that the utility of the original document has been perfectly preserved or, in other words, that the document has not been modified. Therefore, TPI provides a clearer view on the (loss of) utility caused by anonymization, which does not depend on the criteria of human annotators. Moreover, per method standard deviations (indicated by the error bars in the figure) are generally lower for TPI than for precision. This suggests that TPI provides a more consistent assessment that is less document-dependent. On the other hand, the deviation of the average is significantly larger for TPI, which means that TPI better differentiates the behavior of the different methods. This is especially noticeable when comparing NER-based methods against PPDP ones: whereas precision tends to produce less differentiated results among them, TPI shows larger differences, which are expected given the distinct principles in which they rely.

Another advantage of TPI is the possibility of measuring the utility preserved by manual anonymizations. In particular, the TPI value for a human anonymization can be considered as a reference for automatic methods. Consequently, any method with a TPI higher than that of the manual anonymization, as was the case for all the NER-based methods, can be regarded as privacy-lacking. This is because this excess of TPI would come from non-masked (quasi-)identifying terms.

Regarding PPDP methods, the lowest TPI corresponded to the Word2Vec approach, which can be at least partially argued by the fact that it evaluates and masks (quasi-)identifiers individually. This means that any term above the similarity threshold will be masked. In contrast, the k-anonymity-based approach tries to minimize the amount of masked terms by breaking, via term suppression, non-k-anonymous combinations of terms. This means that not all potential quasi-identifiers will be masked but just those required to enforce k-anonymity. Moreover, as intuitively introduced above, the greedy approach has proven to be more successful in preserving utility than the random technique.

4.3. Task-oriented utility evaluation

The following experiment tries to answer the question: which metric captures best the practical utility of the anonymized outputs?

In order to have an empirical reference for the practical utility of the anonymized documents, we applied them on a fundamental NLP task: document clustering. We selected this task because it is unsupervised (therefore, it is not dependent on task-tailored models or human criteria), and because

all the documents' contents are leveraged during clustering. Specifically, we considered the clusters obtained from the original documents as a reference. Then, we repeated the same process on the anonymized documents produced by the different methods and compared the obtained clusters with the reference. The *similarity* between both clusterizations, *i.e.*, original vs. anonymized with a concrete method, will give us an indication of the practical utility that the method was able to preserve. We used the well-known *normalized mutual information* (NMI) [74] as a similarity metric, which produces values in the [0..1] range.

The clustering method that we used was inspired by the approach presented in [75]. First, a BERT-based feature extraction was performed, which produces embeddings for all the document's tokens. Averaging them, we can obtain the document embedding. In the case of anonymized texts, the embeddings of masked text spans (being many of them very general or even meaningless deletion masks) were excluded from the averaging. Subsequently, the document embeddings were clustered with the K-means++ algorithm [76] with $K = 4$. This parameter was determined by evaluating a range of cluster counts (*i.e.*, $K \in \{2, 4, 8, 16, 32\}$) on the original documents, selecting the value that yielded the highest silhouette score [77]. Additionally, to minimize the effect of the random seeds, the clustering of each set of documents was repeated 50 times and the results that minimized inertia were selected.

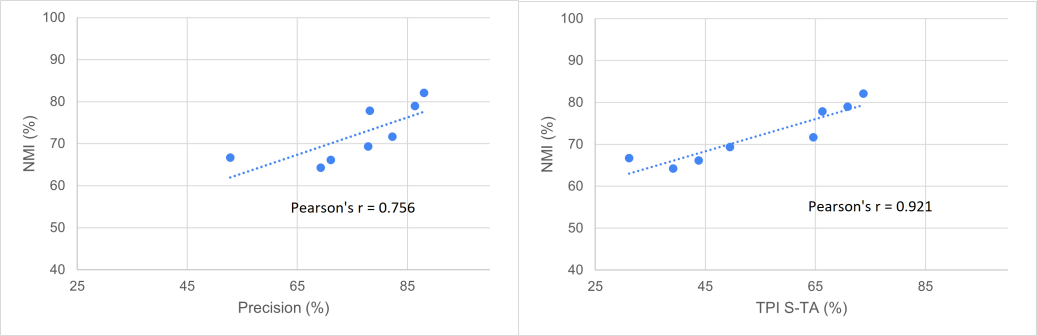
To calculate the NMI for each method, the clustering and evaluation processes were repeated 5 times, and the average NMI was reported. As a result, each clustering was performed 250 ($= 50 \cdot 5$) times, so that the impact of the random seeds is minimized even more.

Figure 3 plots the correlation between the resulting cluster similarity (NMI) for the different anonymization methods and precision and TPI.

The empirical utility (NMI) presents a very strong positive correlation with the respective TPI of each method ($r = 0.921$), which is significantly higher than the correlation observed for precision ($r = 0.756$). The p-value for both correlations is below 0.05, which means that the difference observed is statistically significant. These results provide evidence to support the assertion that TPI provides a more accurate characterization of documents' utility than precision.

4.4. Comparing context masking strategies

The next experiments are designed to evaluate the influence of the context masking strategies discussed in Section 3.2. This is equivalent to measuring



(a) (b)

Figure 3: Correlation between practical utility based on NMI cluster similarity and (a) Precision and (b) TPI S-TA, for automatic anonymization methods. Each scatter plot displays Pearsons r alongside a dashed trend line illustrating the linear relationship between metrics.

the effect of changing the context used by the language model to predict terms, which directly influences the calculation of TIC and TPI.

Figure 4 compares TPI values with the S-TA masking strategy (sentence-level context) with those obtained using N-TA with $N = [2, 4, 6]$, where N defines the length of the context; for example, 2-TA masks one term of every two, while 4-TA masks one term of every four. We employ N up to 6 since it roughly corresponds to $1/6 = 16,\widehat{6}\%$ term masking, which is similar to the 15% token masking used for MLM pre-training of BERT [45]. In fact, we did not observe substantial differences using $N > 6$, which will result in a lower masking ratio and in a larger (but unexploited) context. On the other hand, S-TA is assumed to provide the most consistent results, due to working at a sentence-level (that is, the usually most useful context) and masking sentence's terms one by one.

We can observe that increasing N for the N-TA strategy approximates TPI's results to those obtained with S-TA. In particular, as can be observed in the average column, values with 2-TA overestimate the S-TA values by an average of 8%, while 6-TA reduces this overestimation to an average of 2%. The fact that TPI values decrease inversely proportional to the context length can be argued by the fact that larger contexts enable a more accurate characterization of the TIC of terms, being the most informative ones those that have been subjected to protection.

Finally, in Table 2, we report the runtime of the TIC calculation (the dom-

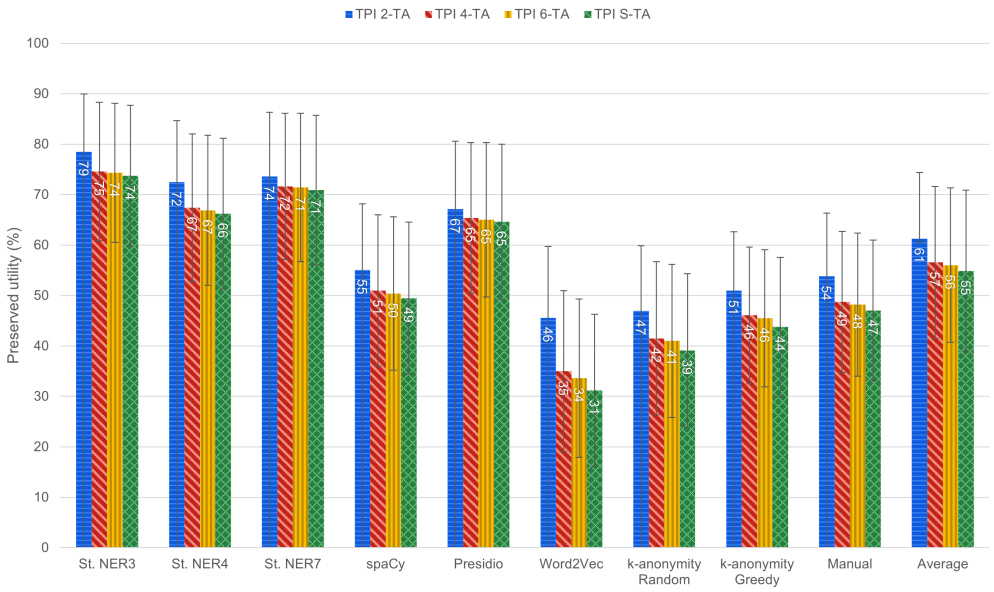


Figure 4: Impact of context masking on TPI measurements for the every anonymization method. This chart contrasts the S-TA strategy with multiple N-TA strategies ($N \in \{2, 4, 6\}$). The "Average" bar aggregates values across anonymization approaches. Higher values are better.

inant computational cost of TPI) for the different context masking strategies. Notice that the runtime with S-TA is around 10 times greater than with 6-TA. Nevertheless, Figure 4 shows that 6-TA provides very similar results to S-TA. This is in line with the reasoning made in Section 3.2, which postulates N-TA as a lower-cost technique that approximates the results of S-TA.

Notably, the runtime for TIC is on the order of minutes, which is negligible compared to the cost of generating the manual annotations needed for the standard precision metric. The latter generally requires several hours of human labor, as evidenced by the 800 hours devoted to the annotation of the TAB dataset [34]. The economic cost is also significantly reduced because the salaries of human annotators can be replaced by the substantially lower cost of computational resources. In summary, the unsupervised nature of TPI brings significant savings compared to the standard human-based evaluation and renders it far more practical.

Table 2: TIC calculation runtime (the dominant computational cost of TPI assessment) using S-TA and multiple N-TA ($N \in \{2, 4, 6\}$) context masking strategies for the original (non-protected) documents. Lower values are better.

Context masking strategy	Runtime
2-TA	39”
4-TA	61”
6-TA	65”
S-TA	10’ 21”

4.5. Sensitivity analysis

In this section, we analyze the sensitivity of our metric and of the standard precision to different anonymization intensities. For the NER-based methods, we simulated varying protection levels by randomly ablating the generated maskings, as these methods lack configurable parameters. Specifically, we applied two ablation levels retaining 75% and 50% of the maskings produced by St. NER4. A lower retention percentage corresponds to higher expected utility, with the 50% setting preserving half of the maskings of the baseline approach. The ablation process ensures that the 75% configuration is a superset of the 50% configuration.

For the PPDP-oriented method, we experimented with the similarity threshold t of the Word2Vec-based approach. Note that decreasing this parameter results in more terms being masked (*i.e.*, lower utility), because the similarity-based filter becomes less restrictive. The set of terms masked at a given threshold always encompasses those masked at a lower, more restrictive threshold. In this experiment, we evaluate three threshold values ($t = 0.25$, $t = 0.5$, and $t = 0.75$) to assess performance in the [0..1] similarity range.

Figure 5 shows that TPI exhibits a clear sensitivity advantage over precision: while TPI effectively captures the expected changes in utility with respect to different anonymization intensities, precision remains largely unaffected. This behavior can be attributed to the formulation of precision in Eq. 3, which normalizes the count of correct maskings (*i.e.*, $\#TruePositives$) by the total number of maskings performed by the evaluated method (*i.e.*, $\#TruePositives + \#FalsePositives$). Consequently, precision remains invariant to the masking count, focusing exclusively on the alignment with ground truth: it assesses the *qualitative* consistency of the masking set as a (any-size) subset of manual annotations. Because the underlying selection

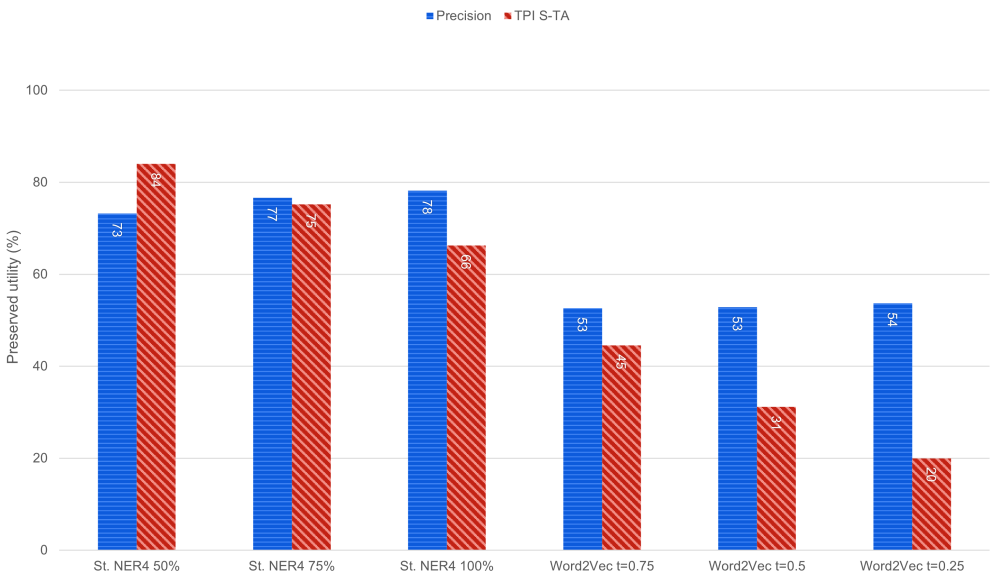


Figure 5: Comparison between TPI S-TA (red) and precision (blue) across varying protection levels for both St. NER4 and Word2Vec with varying parameters. For St. NER4, we applied ablation at 50% and 75% masking preservation, with lower percentages resulting in fewer maskings. For PPDP, we varied the Word2Vec similarity threshold ($t \in \{0.25, 0.5, 0.75\}$), where a lower threshold increases the number of masked terms.

logic is consistent across the different protection levels considered, the ratio of correct to total maskings stays constant in front of the absolute volume of suppressed terms. This inherent scalar invariance makes precision insensitive to the *quantitative* changes in utility that correspond to adjustments in anonymization intensity.

5. Evaluation of text anonymization methods

In this section, we leverage the results obtained through our utility metric and the TRIR privacy metric we presented in [35], in order to offer a comprehensive evaluation of the anonymization methods introduced in the previous section. We utilize the implementation available at `https://github.com/BenetManzanaresSalor/TAE`. As discussed in Section 2.2, both metrics share similar working principles: they are able to assess anonymized documents without relying on manual annotations, they try to measure, respectively, the *actual* utility and privacy retained/achieved in the anonymized outcomes,

and they provide results within the [0..1] range, so that their magnitudes are comparable and integrable. In particular, by aggregating both metrics, we are able to quantify the privacy/utility balance achieved by anonymization methods and, therefore, measure and compare their global performance. Finally, we also compare the figures produced by our metrics with precision and recall.

5.1. Privacy assessment

As introduced in Section 2.2, the TRIR privacy metric estimates the residual re-identification risk of a set of anonymized documents according to the success of a re-identification attack applied to those documents. The attack employs a classifier trained on (raw/non-annotated) publicly available identified sources describing features of the individuals referred in the protected documents. Those public sources constitute the background knowledge that potential attackers may leverage to perform re-identifications.

Following the experimental setting described in [35], we employed the (identified) bodies of the Wikipedia biographies considered so far as background knowledge to train the classifier, and applied the latter to re-identify the abstracts of the same articles protected with the different anonymization methods. In addition, we also measured the baseline re-identifiability of the articles by applying the same classifier on the original unprotected abstracts. To obtain the article bodies, which were not available in the original evaluation dataset [62], we scraped the Wikipedia. We observed that around 8% of the biographies lacked a body section, consisting only of an abstract with no references. This was expected since the dataset included several short articles corresponding to less well-known individuals. Therefore, the re-identification accuracy is capped at 92% for all methods.

In Figure 6 we report, in percentage, the TRIR-based privacy protection assessment as the inverse of the TRIR-based re-identification risk, that is, $100-TRIR$. For reference, we also report the recall –which is standardly used to measure privacy– computed with respect to the manual anonymization. Consequently, the recall for the original unmasked documents is 0%, while that for the manually anonymized ones is 100%. Note also that, whereas recall represents the agreement between automatic methods and manual annotations, TRIR estimates the residual disclosure risk. This is in line with the argument stated in Section 3.1 when comparing precision and TPI.

Focusing on TRIR, as expected, the worse protection (*i.e.*, highest TRIR) corresponds to the original documents. However, it is quite above the mini-

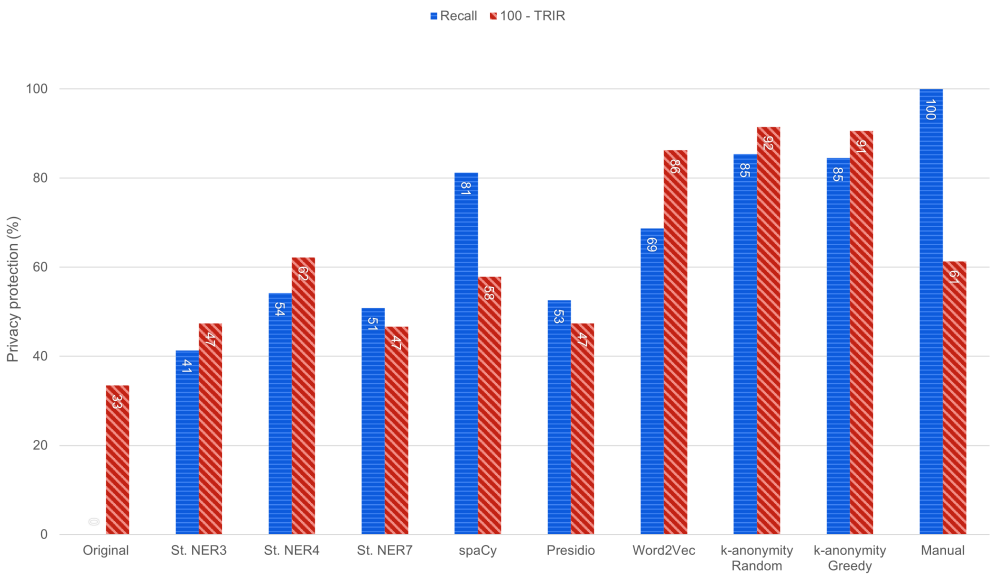


Figure 6: Privacy protection across anonymization methods according to recall (blue) and the complement of TRIR, expressed as 100-TRIR (red). Higher values indicate stronger protection.

mum theoretical privacy level that, according to the discussion above, would be $100 - 92 = 8\%$. This could be attributed to the large number of individuals to be re-identified, which poses a significant challenge for the (highly dimensional multi-class) classifier. Moreover, many of the bodies used as background knowledge were brief, therefore limiting the training data available for re-identifying the corresponding individuals.

Examining NER-based methods, we see that most of them barely improve the baseline TRIR level, with the best results obtained by Stanford's NER4 model. This aligns with the criticisms raised in [8, 78], which argue that, because the use of NER models restricts the types of entities masked, the kind of protection offered is more akin to de-identification than to anonymization. In this sense, St. NER4 model provides the best results among NER-based methods because it incorporates a MISC category that encompasses a larger variety of (quasi-)identifiers than the other models' categories. Moreover, NER-based methods offer a *fixed* protection, that is, all detected entities are systematically masked, whereas undetected ones are left in clear. As a result, it is not possible to configure the trade-off between utility and privacy, as it is common in sound anonymization methods [6]. This implies that the

level of privacy attained is the best they can get, and no parameters can be configured to, *e.g.*, prioritize privacy over utility.

Furthermore, according to TRIR, the manual annotations attained a privacy level as low as the best NER-based method. This is far from optimal protection, with re-identification probability significantly exceeding the random guess baseline (*i.e.*, 39% vs 0.18%). Given that these annotations were produced by multiple annotators following meticulously designed guidelines, and considering that risks exceeding random guess levels were observed in other datasets [35, 79], this suggests inherent limitations in human-based approaches. Contrary to typical natural language processing tasks, where manual annotations are often regarded as the ground truth, text anonymization may require additional skills beyond linguistic comprehension. These skills include recalling combinations of quasi-identifiers and assessing their uniqueness, areas where humans may lack proficiency. This raises critical concerns about the reliability of manual annotations as ground truth. A suboptimal reference not only undermines the validity of precision and recall as evaluation metrics, but also leads to an unfair evaluation of more privacy-resilient methods (Word2Vec and k-anonymity-based ones, in this case). In this specific case, a potential issue is the over-reliance of these annotations on privacy-deficient NER-based pre-annotations. This NER-bias can be seen in the significant divergence between TRIR and recall for the spaCy tool, which does not occur for the other NER models; note that a similar phenomenon, although not as noticeable, can be seen for TPI vs. precision in Figure 2. Because pre-annotations were actually done with spaCy, it makes sense that the results of that tool better match with the manual annotations that relied (apparently too much) on them, therefore resulting in a high recall. On the other hand, the recall of other NER models with similar or even better levels of TRIR-based privacy was much lower. Consequently, by using this manual anonymization as ground truth, we are giving an unfair advantage to spaCy and, by extension, to similar (NER-based) methods.

On the other hand, PPDP methods yielded the best privacy figures (that is, the lowest TRIR, albeit still higher than the random guess, which is 0.18% for this dataset). This can be attributed, at least in part, to the fact that they characterize and anonymize the corpus as a whole, instead of treating documents (and terms) in isolation as NER-based methods do. This facilitates the detection of the rarest (combinations of) of terms, which are usually the ones that entail more risk. Moreover, on the contrary to NER-based approaches, PPDP methods provide parameters (*i.e.*, k for the

k-anonymity-based approaches and the similarity threshold for Word2Vec), which allow configuring the level of privacy protection –and, inversely, utility preservation– according to the user needs.

With the parameters used, the best TRIR was obtained by the k-anonymity random method, although it was less than 1% better than its greedy counterpart. However, this comes at the expense of a lower utility, as reported in Figure 2, which therefore results in the greedy method having a better privacy/utility trade-off. The Word2Vec-based approach is the closest, but with significantly worse utility (see Figure 2). The better performance of the k-anonymity-based methods can be explained by the high quality statistics provided by the knowledge graph on which they are based. From this graph, created by counting the number of appearances of Wiki-tagged entities, it is possible to accurately compute the number of appearances (*i.e.*, the k-anonymity level) of terms and term combinations at the Wikipedia level. However, extending this knowledge graph to other domains can be costly. On the other hand, the Word2Vec model relies on lower quality data (*i.e.*, a comparatively small collection of non-annotated documents), which are easier to compile and require neither preprocessing nor tagging.

5.2. F-scores

Precision and recall can be aggregated to compute F-scores that, in the context of text anonymization, account for the balance between privacy and utility. Being our metrics viable alternatives to precision and recall, and providing both of them values normalized within the [0..1] range, we can also compute F-scores based on TPI and TRIR. In this section, we report the standard F_1-score, and the more privacy-oriented F_2-score, for both precision and recall, and our metrics. The F_2-score, which doubles the weight of the privacy metric, is motivated by the fact that achieving a strong privacy protection is usually a priority when releasing sensitive data, and it is the preferred score of several authors [80, 81].

The F_1 and F_2-scores are calculated as instantiations of F_β (Eq. 9). Notice that, as we have done in Section 5.1, privacy has been computed as $(1 - TRIR)$ to be properly aggregated with utility (TPI). Figures 7 and 8 report, respectively, the F_1 and F_2-scores resulting from our metrics and from precision and recall.

$$F_\beta = (1 + \beta^2) \cdot \frac{Utility \cdot Privacy}{(\beta^2 \cdot Utility) + Privacy} \tag{9}$$

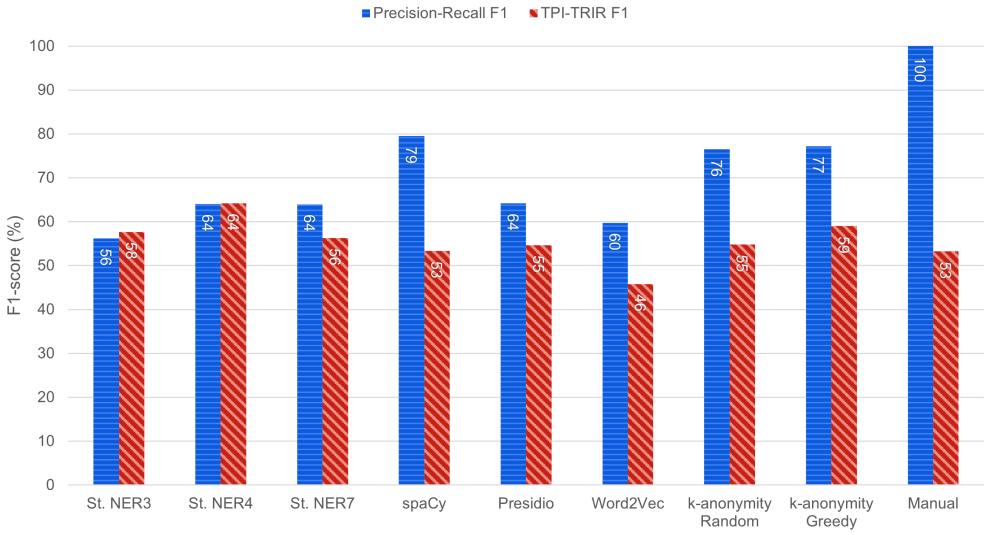


Figure 7: Comparison of F_1-scores across anonymization methods, aggregating utility and privacy metrics. The chart contrasts traditional precision-recall-based F_1 (blue) with TPI-TRIR-based F_1 (red). Higher values are better.

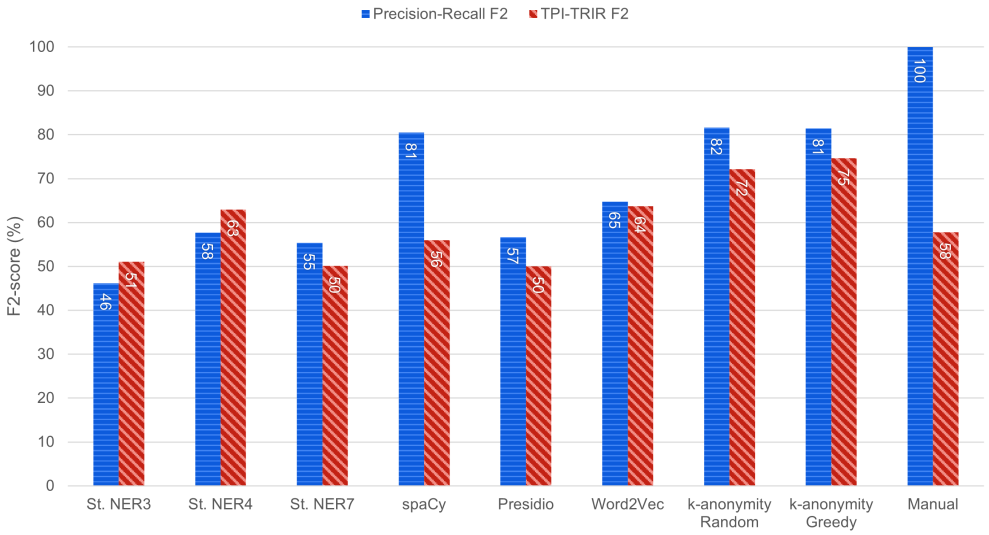


Figure 8: Comparison of F_2-scores across anonymization methods, aggregating utility and privacy metrics while prioritizing the latter. The chart contrasts traditional precision-recall-based F_2 (blue) with TPI-TRIR-based F_2 (red). Higher values are better.

We can see that the relative ranking of automatic methods obtained by sorting them via their F_1 and F_2-scores differ for each pair of metrics. Considering that our utility assessment is highly correlated with precision ($r = 0.923$ for the automatic methods, as discussed in Section 4.2), the divergences mainly come from the privacy evaluation ($r = 0.8$ for the automatic methods' results reported in Figure 6). In this respect, in Section 5.1 we observed that the manual annotations were privacy lacking, thereby rendering metrics relying on them as ground truth unreliable. Moreover, they were also biased by the spaCy-based pre-annotations, which resulted in an unfair advantage for that tool. This can be clearly seen by the (almost) maximum (precision,recall)-based F-scores obtained by spaCy, with figures that are significantly higher than for other NER-based methods, and which significantly diverge from the (TPI,TRIR)-based F-scores.

Focusing on our metrics, we can see that, whereas the F_1-score gives the edge to the NER4 model, the F_2-score ranks PPDP methods first. Given that TPI produces significantly smaller values than precision (as discussed in Section 4.2), whereas TRIR's distribution is in line with recall's (as shown in Section 5.1), the F_2-score provides a more even aggregation of our metrics' scales. This is consistent with the individual privacy and utility figures reported in the previous sections. In fact, the correlation between the (TPI,TRIR) and (precision,recall) F_2 scores is $r = 0.73$, which corresponds to the result of multiplying the correlations between precision and TPI ($r = 0.923$), and recall and TRIR ($r = 0.8$). Therefore, the F_2-score, which should be preferred in privacy-oriented applications [80, 81], should also be preferred for our metrics.

6. Conclusions

We have proposed the first unsupervised utility evaluation metric for text anonymization. Compared to the widely employed precision, TPI provides an alternative that does not require costly and error-prone manual annotations. Experiments depict a strong correlation between TPI and precision, making TPI a viable unsupervised alternative to the latter. Furthermore, our task-oriented evaluation suggests that TPI better captures the practical utility retained by the anonymized outcomes.

Together with our previous privacy metric [35], TPI defines a complete evaluation framework that is free from the issues that arise when employing precision, recall and manual annotations to evaluate anonymization meth-

ods. The suitability of our metrics is illustrated through the comprehensive evaluation of a variety of text anonymization methods reported in Section 5. The reported results provide empirical evidence for the criticisms raised in [8, 61] on the drawbacks of using NER methods for text anonymization. Moreover, they also show that privacy-oriented methods [71, 62] produce more robust anonymizations, whereas manual anonymization may be hampered by inaccuracies and bias and may introduce compromises when used as ground truth.

As future work, we plan to use text similarity metrics to accurately estimate utility when generalizations (rather than suppressions) are used as privacy-preserving masks. This would provide a more accurate utility assessment for the (very few) text anonymization methods that mask terms through meaningful generalizations [61]. The idea is to employ the similarity between each generalization and the original text span to weight the loss of information resulting from each masking.

Another future research direction is to investigate the impact of utilizing other language models for probability (and subsequently IC) estimation. TPI is fundamentally model-agnostic, and can accommodate any model that assigns probabilities to terms as long as those probabilities are derived from the same model across the entire corpus, in order to ensure consistency in numerical scale and distribution. Nevertheless, a compelling research question emerges regarding the use of *large language models* (LLMs) [82, 83]: these models are generally trained for *causal* language modeling, predicting the next token based solely on preceding context. This contrasts with models trained for *masked* language modeling, such as BERT [45], which consider both preceding and succeeding context. Consequently, the proposed context masking strategies (*i.e.*, S-TA and N-TA) are not applicable, and thus a new solution should be envisaged. An intuitive solution could involve incremental probability estimation, that is, processing terms sequentially. However, this introduces a potential positional bias, where a term's predictability may be more reflective of its index in the text (and thus the available context) than its semantic value. Addressing these challenges is essential to fully harness the potential of LLMs for TPI calculation.

Finally, our metric can also be leveraged to guide the masking of the potential quasi-identifiers detected during the anonymization process. The idea is to use the estimated IC of the potential quasi-identifiers to select the masking combination that minimizes the information loss at a document level.

Acknowledgments

We acknowledge support from the Norwegian Research Council (CLEANUP project (http://cleanup.nr.no/), grant nr. 308904), the Government of Catalonia (Acadèmia d'Excel.lència Prize to D. Sánchez) and MCIN/AEI under grant PID2024-157271NB-I00 "CLEARING-IT". B. Manzanares-Salor is also supported by the Spanish Government under an FPU grant (ref. FPU23/01785).

References

[1] S. Marcos-Pablos, F. J. García-Peñalvo, Information retrieval methodology for aiding scientific database search, Soft Computing 24 (8) (2020) 5551–5560.

[2] T. Hao, X. Chen, G. Li, J. Yan, A bibliometric analysis of text mining in medical research, Soft Computing 22 (2018) 7875–7892.

[3] W. Duch, P. Matykiewicz, J. Pestian, Neurolinguistic approach to natural language processing with applications to medical text analysis, Neural Networks 21 (10) (2008) 1500–1510, iCONIP 2007. doi:https://doi.org/10.1016/j.neunet.2008.05.008.
URL https://www.sciencedirect.com/science/article/pii/S0893608008000956

[4] Z. Lan, J. Yu, S. Liu, J. Yao, D. Huang, J. Su, Towards better text image machine translation with multimodal codebook and multi-stage training, Neural Networks 189 (2025) 107599. doi:https://doi.org/10.1016/j.neunet.2025.107599.
URL https://www.sciencedirect.com/science/article/pii/S0893608025004794

[5] S. Adhikari, S. Thapa, U. Naseem, H. Y. Lu, G. Bharathy, M. Prasad, Explainable hybrid word representations for sentiment analysis of financial news, Neural Networks 164 (2023) 115–123. doi:https://doi.org/10.1016/j.neunet.2023.04.011.
URL https://www.sciencedirect.com/science/article/pii/S0893608023001880

[6] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, P.-P. De Wolf, Statistical Disclosure Control, Vol. 2, Wiley New York, New York, 2012.

[7] J. Domingo-Ferrer, D. Sánchez, J. Soria-Comas, Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Intermodel Connections, Vol. 8, Synthesis Lectures on Information Security, Privacy, and Trust, Online, 2016.

[8] P. Lison, I. Pilán, D. Sánchez, M. Batet, L. Øvrelid, 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), Association for Computational Linguistics, Online, 2021, pp. 4188–4203.

[9] G. M. Csányi, D. Nagy, R. Vági, J. P. Vadász, T. Orosz, Challenges and open problems of legal document anonymization, Symmetry 13 (8) (2021) 1490.

[10] S. Shahriar, R. Dara, R. Akalu, A comprehensive review of current trends, challenges, and opportunities in text data privacy, Computers & Security 151 (2025) 104358. doi:https://doi.org/10.1016/j.cose.2025.104358.
URL https://www.sciencedirect.com/science/article/pii/S0167404825000471

[11] Y. Ren, K. Ramesh, Y. Yao, A. Field, How do we measure privacy in text? a survey of text anonymization metrics, in: K. Inui, S. Sakti, H. Wang, D. F. Wong, P. Bhattacharyya, B. Banerjee, A. Ekbal, T. Chakraborty, D. P. Singh (Eds.), Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, Mumbai, India, 2025, pp. 1532–1544.
URL https://aclanthology.org/2025.findings-ijcnlp.94/

[12] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, M. H. Samore, Automatic de-identification of textual documents in the electronic health

record: a review of recent research, BMC Medical Research Methodology 10 (1) (2010) 1–16.

[13] J. Aberdeen, S. Bayer, R. Yeniterzi, B. Wellner, C. Clark, D. Hanauer, B. Malin, L. Hirschman, The mitre identification scrubber toolkit: Design, training, and assessment, International Journal of Medical Informatics 79 (12) (2010) 849–859.

[14] A. Chen, J. Jonnagaddala, C. Nekkantti, S.-T. Liaw, Generation of Surrogates for De-Identification of Electronic Health Records, IOS Press, Lyon, France, 2019, book section 1, pp. 70–73.

[15] F. Dernoncourt, J. Y. Lee, O. Uzuner, P. Szolovits, De-identification of patient notes with recurrent neural networks, Journal of the American Medical Informatics Association 24 (3) (2017) 596–606.

[16] Y. Elazar, Y. Goldberg, Adversarial removal of demographic attributes from text data, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 11–21.

[17] F. Hassan, J. Domingo-Ferrer, J. Soria-Comas, Anonymization of unstructured data via named-entity recognition, in: International Conference on Modeling Decisions for Artificial Intelligence, Springer, Mallorca, Spain, 2018, pp. 296–305.

[18] Y. Huang, Z. Song, D. Chen, K. Li, S. Arora, Texthide: Tackling data privacy in language understanding tasks, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2020, pp. 1368–1382.

[19] A. E. Johnson, L. Bulgarelli, T. J. Pollard, Deidentification of free-text medical records using pre-trained bidirectional transformers, in: Proceedings of the ACM Conference on Health, Inference, and Learning, Association for Computing Machinery, Toronto, Ontario, Canada, 2020, pp. 214–221.

[20] Z. Liu, B. Tang, X. Wang, Q. Chen, De-identification of clinical notes via recurrent neural network and conditional random field, Journal of Biomedical Informatics 75 (2017) S34–S42.

[21] N. Mamede, J. Baptista, F. Dias, Automated anonymization of text documents, in: IEEE Congress on Evolutionary Computation, IEEE, Vancouver, BC, Canada, 2016, pp. 1287–1294.

[22] I. Neamatullah, M. M. Douglass, L.-W. H. Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark, G. D. Clifford, Automated de-identification of free-text medical records, BMC Medical Informatics and Decision Making 8 (1) (2008) 1–17.

[23] S. Reddy, K. Knight, Obfuscating gender in social media writing, in: Proceedings of the First Workshop on NLP and Computational Social Science, Association for Computational Linguistics, Austin, Texas, USA, 2016, pp. 17–26.

[24] G. Szarvas, R. Farkas, R. Busa-Fekete, State-of-the-art anonymization of medical records using an iterative machine learning framework, Journal of the American Medical Informatics Association 14 (5) (2007) 574–580.

[25] Q. Xu, L. Qu, C. Xu, R. Cui, Privacy-aware text rewriting, in: Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics, Tokyo, Japan, 2019, pp. 247–257.

[26] H. Yang, J. M. Garibaldi, Automatic detection of protected health information from clinic narratives, Journal of Biomedical Informatics 58 (2015) S30–S38.

[27] V. Yogarajan, M. Mayo, B. Pfahringer, A survey of automatic de-identification of longitudinal clinical narratives, arXiv:1810.06765 (2018).

[28] R. Juez-Hernandez, L. Quijano-Snchez, F. Liberatore, J. Gmez, Agora: An intelligent system for the anonymization, information extraction and automatic mapping of sensitive documents, Applied Soft Computing 145 (2023) 110540. doi:https://doi.org/10.1016/j.asoc.2023.110540. URL https://www.sciencedirect.com/science/article/pii/S1568494623005586

[29] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. M. Lundberg, H. Nori, H. Palangi, M. T.

Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with GPT-4, CoRR abs/2303.12712 (2023). arXiv:2303.12712, doi:10.48550/ARXIV.2303.12712.
URL https://doi.org/10.48550/arXiv.2303.12712

[30] S. Singhal, A. F. Zambrano, M. Pankiewicz, X. Liu, C. Porter, R. S. Baker, De-identifying student personally identifying information with gpt-4, in: Proceedings of the 17th International Conference on Educational Data Mining, 2024, pp. 559–565.

[31] V. Torra, K. Stokes, A formalization of record linkage and its application to data protection, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 20 (06) (2012) 907–919.

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

[33] E. Bier, R. Chow, P. Gollé, T. H. King, J. Staddon, The rules of redaction: Identify, protect, review (and repeat), IEEE Security & Privacy 7 (6) (2009) 46–53.

[34] I. Pilán, P. Lison, L. Ørelid, A. Papadopoulou, D. Sánchez, M. Batet, The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization, Computational Linguistics (2022) 1053–1101.

[35] B. Manzanares-Salor, D. Sánchez, P. Lison, Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack, Data Mining and Knowledge Discovery 38 (6) (2024) 4040–4075.

[36] E. Mackey, M. Elliot, K. O'Hara, The Anonymisation Decision-making Framework, UKAN Publications, Manchester, 2016.

[37] R. Chevrier, V. Foufi, C. Gaudet-Blavignac, A. Robert, C. Lovis, Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review, Journal of Medical Internet Research 21 (5) (2019) e13484.

[38] M. Mozes, B. Kleinberg, No intruder, no validity: Evaluation criteria for privacy-preserving text anonymization, arXiv:2103.09263 (2021).

[39] C. E. Shannon, A mathematical theory of communication, The Bell system technical journal 27 (3) (1948) 379–423.

[40] V. Domingo-Ferrer, Josep; Torra, A quantitative comparison of disclosure control methods for microdata, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies (2001) 111–133.

[41] J. Domingo-Ferrer, D. Sánchez, G. Rufian-Torrell, Anonymization of nominal data based on semantic marginality, Information Sciences 242 (2013) 35–48.

[42] S. Martínez, D. Sánchez, A. Valls, A semantic framework to protect the privacy of electronic health records with non-numerical attributes, Journal of biomedical informatics 46 (2) (2013) 294–303.

[43] S. Martínez, D. Sánchez, A. Valls, Semantic adaptive microaggregation of categorical microdata, Computers & Security 31 (5) (2012) 653–672.

[44] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45.

[45] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, l. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, Neural Information Processing Systems Foundation, Long Beach, CA, USA, 2017, pp. 5998–6008.

[47] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. J. J. M. L. R. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (140) (2020) 54855551.

[48] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, On the opportunities and risks of foundation models, Radiology: Artificial Intelligence 4 (4) (2021) e220119.

[49] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[50] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, in: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 103–111.

[51] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp. 2227–2237.

[52] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv:1907.11692 (2019).

[53] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv:1910.01108 (2019).

[54] M. Batet, D. Sánchez, Semantic disclosure control: semantics meets data privacy, Online Information Review 42 (3) (2018) 290–303.

[55] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, pp. 448–453.

[56] M. Batet, D. Sánchez, Leveraging synonymy and polysemy to improve semantic similarity assessments based on intrinsic information content, Artificial Intelligence Review 53 (3) (2020) 2023–2041.

[57] D. Sánchez, M. Batet, A. Viejo, Automatic general-purpose sanitization of textual documents, IEEE Transactions on Information Forensics and Security 8 (6) (2013) 853–862.

[58] N. Seco, T. Veale, J. Hayes, An intrinsic information content metric for semantic similarity in wordnet, in: Proceedings of the 16th European Conference on Artificial Intelligence, ECAI'04, IOS Press, NLD, 2004, pp. 1089–1090.

[59] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation, arXiv:1609.08144 (2016).

[60] B. Anandan, C. Clifton, Significance of term relationships on anonymization, in: IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, 2011, p. 253256.

[61] F. Hassan, D. Sánchez, J. Domingo-Ferrer, Utility-preserving privacy protection of textual documents via word embeddings, IEEE Transactions on Knowledge and Data Engineering 35 (2023) 1058–1071.

[62] A. Papadopoulou, P. Lison, L. Øvrelid, I. Pilán, Bootstrapping text anonymization models with distant supervision, in: Thirteenth Language Resources and Evaluation Conference, 2022, p. 44774487.

[63] R. Lebret, D. Grangier, M. Auli, Neural text generation from structured data with application to the biography domain, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 1203–1213.

[64] J. Staddon, P. Golle, B. Zimny, Web-based inference detection, in: USENIX Security Symposium, Association for Computing Machinery, Boston, MA, USA, 2007, pp. 1–16.

[65] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, D. McClosky, The stanford corenlp natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 55–60.

[66] R. Weischedel, E. Hovy, M. Marcus, M. Palmer, R. Belvin, S. Pradhan, L. Ramshaw, N. Xue, OntoNotes: A large training corpus for enhanced processing, Springer, New York, 2011, book section 1, pp. 54–63.

[67] C. Cumby, R. Ghani, A machine learning based system for semi-automatically redacting documents, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 25, AAAI Press, San Francisco, California, USA, 2011, pp. 1628–1635.

[68] V. T. Chakaravarthy, H. Gupta, P. Roy, M. K. Mohania, Efficient techniques for document sanitization, in: Proceedings of the 17th ACM conference on Information and knowledge management, Association for Computing Machinery, Napa Valley, California, USA, 2008, pp. 843–852.

[69] B. Anandan, C. Clifton, W. Jiang, M. Murugesan, P. Pastrana-Camacho, L. Si, t-plausibility: Generalizing words to desensitize text, Transactions on Data Privacy 5 (3) (2012) 505–534.

[70] D. Sánchez, M. Batet, Toward sensitive document release with privacy guarantees, Engineering Applications of Artificial Intelligence 59 (2017) 23–34.

[71] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv:1301.3781 (2013).

[72] L. Sweeney, k-anonymity: A model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (05) (2002) 557–570.

[73] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, Santiago, Chile, 2015, pp. 19–27.

[74] A. L. N. Fred, A. K. Jain, Robust data clustering, in: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA, IEEE Computer Society, 2003, pp. 128–136. doi:10.1109/CVPR.2003.1211462.
URL https://doi.org/10.1109/CVPR.2003.1211462

[75] A. Subakti, H. Murfi, N. Hariadi, The performance of bert as data representation of text clustering, Journal of Big Data 9 (1) (2022) 1–21.

[76] D. Arthur, S. Vassilvitskii, K-means++: The advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07, Society for Industrial and Applied Mathematics, USA, 2007, pp. 1027–1035.

[77] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65. doi:https://doi.org/10.1016/0377-0427(87)90125-7.
URL https://www.sciencedirect.com/science/article/pii/0377042787901257

[78] B. Manzanares-Salor, D. Snchez, Enhancing text anonymization via re-identification risk-based explainability, Knowledge-Based Systems 310 (2025) 112945. doi:https://doi.org/10.1016/j.knosys.2024.112945.
URL https://www.sciencedirect.com/science/article/pii/S095070512401579X

[79] B. Manzanares-Salor, D. Sánchez, P. Lison, Automatic evaluation of disclosure risks of text anonymization methods, in: Privacy in Statistical Databases, Vol. 13463 of Privacy in Statistical Databases, Springer, Paris, France, 2022, pp. 157–171.

[80] O. Ferrández, B. R. South, S. Shen, F. J. Friedlin, M. H. Samore, S. M. Meystre, Evaluating current automatic de-identification methods with veteran's health administration clinical documents, BMC medical research methodology 12 (2012) 1–16.

[81] O. Mendels, Nlp approaches to data anonymization (2020).
URL https://towardsdatascience.com/nlp-approaches-to-data-anonymization-1fb5bde6b929

[82] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al., Mixtral of experts, arXiv preprint arXiv:2401.04088 (2024).

[83] D. P. Woodruff, V. Cohen-Addad, L. Jain, J. Mao, S. Zuo, M. Bateni, S. Branzei, M. P. Brenner, L. Chen, Y. Feng, L. Fortnow, G. Fu, Z. Guan, Z. Hadizadeh, M. T. Hajiaghayi, M. JafariRaviz, A. Javanmard, K. C. S., K. ichi Kawarabayashi, R. Kumar, S. Lattanzi, E. Lee, Y. Li, I. Panageas, D. Paparas, B. Przybocki, B. Subercaseaux, O. Svensson, S. Taherijam, X. Wu, E. Yogev, M. Zadimoghaddam, S. Zhou, Y. Matias, J. Manyika, V. Mirrokni, Accelerating scientific research with gemini: Case studies and common techniques (2026). arXiv:2602.03837.
URL https://arxiv.org/abs/2602.03837