

UNIVERSITAT
ROVIRA i VIRGILI

Evaluation and Enhancement of Text Anonymization with Language Models

Benet
Manzanares-
Salor

DOCTORAL THESIS
2025

Benet MANZANARES-SALOR

Evaluation and Enhancement of Text Anonymization with Language Models

DOCTORAL THESIS

Supervisor:
Dr. David SÁNCHEZ

Computer Science and Mathematics of Security doctoral program

Department of Computer Engineering and Mathematics



UNIVERSITAT ROVIRA i VIRGILI

Oct 2025

Abstract

Text anonymization is required to share or release personal textual data for secondary use while complying with privacy regulations. Nevertheless, compared to the well-studied structured data, unstructured text presents additional challenges in both anonymization and its evaluation.

Manual anonymization is still a standard practice, and anonymization evaluation mostly uses (subjective and error-prone) human annotations as the ground truth. This evaluation also employs information retrieval metrics that do not directly assess privacy protection and utility preservation, often leading to unreliable assessments of anonymization quality.

At the methodological level, most automatic approaches to text anonymization rely on named entity recognition (NER), narrowly targeting predefined sets of entity types while overlooking other identifying terms. Furthermore, the vast majority of methods favor suppression over replacement, limiting the utility of the outcomes.

This thesis leverages the capabilities of modern language models to address these limitations by introducing novel approaches for the automatic evaluation and anonymization of text.

We first propose a new evaluation framework grounded in the principles of structured data anonymization, comprising: TRIR, a privacy metric that quantifies re-identification risk through a language model (LM)-based linkage attack; TPI, which measures utility preservation of the anonymized outcomes according to the loss of information content (IC) incurred by masking; and TPS, which extends TPI with semantic similarity to more accurately evaluate utility preservation in replacement-based masking.

To enhance the protection of NER-based approaches, we also present PETRE, which applies explainability techniques on our linkage attack to

discover remaining identifying terms and enforce configurable protection guarantees based on k-anonymity.

We further survey the applications of large language models (LLMs) for both the detection and masking of identifying terms, and introduce a novel approach that combines the strengths of current approaches. Finally, we propose INTACT, an automatic masking method that leverages LLMs to generate and select truthful generalizations that can be used as replacement for sensitive terms while preserving their general meaning.

Experimental results on multiple real-world datasets demonstrate the effectiveness of proposals, laying the foundation for a new paradigm in fully automatic, accurate, and reliable text anonymization.

Resum

La anonimització de text és necessària per compartir o divulgar dades personals en format textual per a usos secundaris complint amb les normatives de privacitat. No obstant això, en comparació amb les dades estructurades –ja àmpliament estudiades–, el text no estructurat presenta desafiaments addicionals tant en la seva anonimització com en la seva avaluació.

L'anonimització manual continua sent una pràctica comuna, i la seva avaluació es basa principalment en anotacions humanes (subjectives i propenses a errors) com a referència. A més, aquesta avaluació empra mètriques de recuperació d'informació que no avaluen directament la protecció de la privacitat ni la preservació de la utilitat, la qual cosa sovint condueix a valoracions poc fiables de la qualitat de l'anonimització.

A nivell metodològic, la majoria dels enfocaments automàtics per a l'anonimització de text es basen en la *named entity recognition* (NER), centrant la seva atenció únicament en conjunts predefinits de tipus d'entitat i passant per alt altres termes identificadors. Així mateix, la gran majoria dels mètodes afavoreixen la supressió per sobre del reemplaçament, cosa que limita la utilitat dels resultats.

Aquesta tesi aprofita les capacitats dels *language models* (LMs) moderns per abordar aquestes limitacions mitjançant la introducció d'enfocaments novedosos per a l'avaluació i anonimització automàtica de text.

Primer, proposem un nou marc d'avaluació basat en els principis de l'anonimització de dades estructurades, que comprèn: TRIR, una mètrica de privacitat que quantifica el risc de reidentificació a través d'un atac de vinculació basat en un LM; TPI, que mesura la preservació de la utilitat dels resultats anonimitzats segons la pèrdua de *information content* (IC) provocada per l'emmascarament; i TPS, que amplia TPI amb similitud semàntica per avaluar amb més precisió la preservació de la utilitat en l'emmascarament basat en reemplaçament.

Per millorar la protecció dels enfocaments basats en NER, també presentem PETRE, que aplica tècniques d'explicabilitat al nostre atac de vinculació per descobrir els termes identificadors restants i obtenir garanties de privacitat configurables basades en el *k*-anonimat.

A més, explorem les aplicacions dels *large language models* (LLM) tant per a la detecció com per a l'emmascarament de termes identificadors, i introduïm un nou enfocament que combina els punts forts dels mètodes actuals. Finalment, proposem INTACT, un mètode automàtic d'emmascarament que aprofita els LLM per generar i seleccionar generalitzacions verídiques que puguin ser utilitzades com a reemplaçament de termes sensibles, preservant el seu significat general.

Els resultats experimentals sobre múltiples conjunts de dades del món real demostren l'efectivitat de les propostes, assentant les bases per a un nou paradigma d'anonimització de text totalment automàtic, precís i fiable.

Resumen

La anonimización de texto es necesaria para compartir o divulgar datos personales en formato textual para usos secundarios cumpliendo con las normativas de privacidad. No obstante, en comparación con los datos estructurados –ya ampliamente estudiados–, el texto no estructurado presenta desafíos adicionales tanto en su anonimización como en su evaluación.

La anonimización manual sigue siendo una práctica común, y su evaluación se basa principalmente en anotaciones humanas (subjetivas y propensas a errores) como referencia. Además, dicha evaluación emplea métricas de recuperación de información que no evalúan directamente la protección de la privacidad ni la preservación de la utilidad, lo que a menudo conduce a valoraciones poco fiables de la calidad de la anonimización.

A nivel metodológico, la mayoría de los enfoques automáticos para la anonimización de texto se basan en *named entity recognition* (NER), centrando su atención únicamente en conjuntos predefinidos de tipos de entidad y pasando por alto otros términos identificadores. Asimismo, la gran mayoría de los métodos favorecen la supresión por encima del reemplazo, lo que limita la utilidad de los resultados.

Esta tesis aprovecha las capacidades de los *language models* (LMs) modernos para abordar estas limitaciones mediante la introducción de enfoques novedosos para la evaluación y anonimización automática de texto.

Primero, proponemos un nuevo marco de evaluación basado en los principios de la anonimización de datos estructurados, que comprende: TRIR, una métrica de privacidad que cuantifica el riesgo de reidentificación a través de un ataque de vinculación basado en un LM; TPI, que mide la preservación de la utilidad de los resultados anonimizados según la pérdida de *information content* (IC) provocada por el enmascaramiento; y TPS, que extiende TPI con similitud semántica para

evaluar con mayor precisión la preservación de la utilidad en el enmascaramiento basado en reemplazo.

Para mejorar la protección de los enfoques basados en NER, también presentamos PETRE, que aplica técnicas de explicabilidad a nuestro ataque de vinculación para descubrir los términos identificadores restantes y obtener garantias de privacidad configurables basadas en *k*-anonimato.

Además, exploramos aplicaciones de los *large language models* (LLM) tanto para la detección como para el enmascaramiento de términos identificadores, e introducimos un nuevo enfoque que combina las fortalezas de los métodos actuales. Finalmente, proponemos INTACT, un método automático de enmascaramiento que aprovecha los LLM para generar y seleccionar generalizaciones verídicas que puedan ser utilizadas como reemplazo de términos sensibles, preservando su significado general.

Los resultados experimentales sobre múltiples conjuntos de datos del mundo real demuestran la efectividad de las propuestas, sentando las bases para un nuevo paradigma de anonimización de texto totalmente automático, preciso y fiable.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor David Sánchez, for their unwavering support and guidance throughout the development of this thesis.Their expertise and constructive feedback have not only significantly enhanced my research skills but have also enabled me to expand the boundaries of my academic pursuits. I wish to extend my heartfelt appreciation to Dr. Jesús Manjón Paniagua for their invaluable support, which proved instrumental during my doctoral studies. My sincere thanks also go to all members of the CRISES group, with special mention to Dr. Najeeb Jebreel for their advice and assistance. Finally, I am profoundly grateful to my family and friends for their constant presence and encouragement throughout this challenging journey.

This research was funded by the Spanish Government under an FPU grant (ref. FPU23/01785). We also acknowledge support from the Norwegian Research Council (CLEANUP project (`http://cleanup.nr.no/`), grant nr. 308904), the Government of Catalonia (ICREA Acadèmia Prize to D. Sánchez, and grant 2021SGR-00115), MCIN/AEI/ 10.13039/ 5011-00011033 and "ERDF A way of making Europe" under grant PID2021-123637NB-I00 "CURLING", and the EU's NextGenerationEU/PRTR via INCIBE (project "HERMES" and INCIBE-URV cybersecurity chair).

Contents

Abstract **iii**

Resum **v**

Resumen **vii**

Acknowledgements **ix**

1 Introduction **1**
1.1 Motivation 1
1.2 Contributions 6
1.3 Outline 7

2 Background **9**
2.1 Anonymization 9
2.1.1 Structured data anonymization 10
2.1.2 Text anonymization 12
2.1.3 Redaction and sanitization 15
2.1.4 Perturbative methods 16
2.2 Evaluation 17
2.2.1 Structured data 18
2.2.2 Textual data 19
2.3 Language Models 23
2.4 Datasets 26

3 Evaluating disclosure risk **31**
3.1 Re-identification attack and risk metric for anonymized text 33
3.1.1 Formalization 34
3.1.2 Attacker's resources 37
3.2 Choosing and implementing the classifier model 39
3.2.1 Selecting the language model 39
3.2.2 From language model to classifier model 41
3.2.3 From documents to model input 41
3.2.4 Building the classifier 43
3.2.5 Prediction 46
3.2.6 Selection of hyperparameters 48
3.3 Evaluation results 49
3.3.1 Evaluated methods 49
3.3.2 Evaluation corpus and background documents . . . 52
3.3.3 Testing environment and hyperparameters 55
3.3.4 Influence of attacker's resources 56
3.3.5 Comparison with recall and human annotations . . 63
3.4 Conclusions 68

4 Unsupervised utility evaluation **71**
4.1 Information content-based utility metric for anonymized texts 72
4.1.1 Formalization 73
4.1.2 Probability calculus 75
4.2 Experiments on utility evaluation 83
4.2.1 Anonymization methods 83
4.2.2 Results 85
4.2.3 Task-oriented utility evaluation 87
4.2.4 Comparing utility calculation strategies 89
4.3 Evaluation of text anonymization methods 91
4.3.1 Privacy assessment 91
4.3.2 F-scores 96

4.4 Conclusions 98

5 Enhancing privacy protection **101**
5.1 Privacy Enhancement for Text via Risk-oriented Explainability 103
5.1.1 Architecture 104
5.1.2 Formalization 109
5.2 PETRE implementation 112
5.3 Experiments 114
5.3.1 Evaluation datasets 115
5.3.2 Anonymization methods 116
5.3.3 Background knowledge 117
5.3.4 Evaluation metrics 118
5.3.5 Setting the privacy parameter k 119
5.3.6 Results 120
5.4 Conclusions 132

6 LLM-based text anonymization **133**
6.1 Analysis and enhancement of LLM-based anonymization . 135
6.1.1 Analysis 135
6.1.2 Our prompt 139
6.2 Evaluation framework 142
6.2.1 Privacy metric 143
6.2.2 Utility metric 143
6.3 Evaluation results 144
6.3.1 Dataset 144
6.3.2 Implementation details 145
6.3.3 Non-LLM-based baselines 147
6.3.4 Results and discussion 147
6.3.5 Ablation study of our prompt 153
6.4 Conclusions 155

7 Truthful text sanitization **159**
7.1 Text sanitization method 161
7.1.1 Replacement generation 162
7.1.2 Replacement selection 164
7.1.3 Full sanitization process 166
7.2 Evaluation framework 169
7.2.1 Utility assessment 169
7.2.2 Privacy assessment 172
7.3 Experimental results 173
7.3.1 Dataset 173
7.3.2 Implementation details 173
7.3.3 Baselines 179
7.3.4 General statistics 180
7.3.5 Utility evaluation 182
7.3.6 Privacy evaluation 185
7.3.7 Truthfulness and abstraction level 188
7.3.8 Generalizability to other LLMs 192
7.4 Discussion 196
7.5 Conclusions 198

8 Conclusions and future work **201**
8.1 Publications 205
8.2 Future work 207

Bibliography **211**

List of Figures

3.1 TRIA scenario, including individuals and documents sets from the point of view of the attacker. 35
3.2 Workflow for the build_classifier method from Algorithm 1. 44
3.3 Our proposed workflow for the predict method. 47
3.4 TRIR using different background knowledge breadths for several anonymization approaches. Horizontal lines depict the upper bounds for *500_filtered* and *2000_filtered*. . . 57
3.5 TRIR using the first 25% of the content from the background knowledge texts. 59
3.6 Comparison between different split sizes at fine-tuning for the *50_eval* scenario. 61
3.7 Ablation study on the inclusion of anonymized bodies in training for the *50_eval* scenario. 62
3.8 Effect of different language models on TRIR. 64
3.9 Comparison between the standard recall metric and the most and least correlated TRIR's attack configurations. . . 66
3.10 Precision of automatic anonymization methods. 67

4.1 Comparison between N-TA and S-TA on an sample sentence 81
4.2 TPI and precision for different anonymization methods . . 86
4.3 Correlation between the NMI cluster similarity for the different automatic anonymization methods and (A) Precision and (B) TPI S-TA 89

4.4 TPI values calculated with different context masking strategies 90
4.5 Privacy protection attained for the different anonymization methods according to recall and TRIR 93
4.6 TPI and TRIR-based F_1-score vs. precision and recall-based F_1-score for the different anonymization methods 97
4.7 TPI and TRIR-based F_2-score vs. precision and recall-based F_2-score for the different anonymization methods 97

5.1 PETRE architecture and workflow 105
5.2 Evaluation metrics of starting anonymizations and their PETRE-anonymized counterparts 121
5.3 Correlation between the recall of starting anonymizations and the improvement rate brought by PETRE 123
5.4 Correlation between PETRE's runtime and its recall improvement rate 125
5.5 PETRE mean runtime per document for different k values and starting anonymizations 126
5.6 Mean ranks for each PETRE-enhanced starting anonymization with different values of k 127
5.7 Evaluation metrics of starting anonymizations and their PETRE-anonymized counterparts for different background knowledge sets 129

6.1 Privacy and utility evaluation of the considered anonymization methods 148
6.2 Privacy and utility evaluation across ablations of our proposed prompt. Starting from the rightmost point of each line, the models used are GPT-3.5-Turbo, GPT-4o-mini and GPT-4o, respectively. 154

7.1 Overview of the text sanitization process with INTACT, together with a sample system output. 162

7.2 Frequency of INTACT's replacement choice for TAB. *Entity Type* is selected when all five candidates are guessed by the inference attacks, leading to a default strategy of replacing the term by its entity type label (*e.g.*, *PERSON*). . 181
7.3 TPS for the different replacement strategies 182
7.4 NMI for the clustering resulting from the evaluated replacement strategies 184
7.5 TRIR for different replacement strategies 187
7.6 Frequency of INTACT's replacement choice for the test set of TAB with the Mistral 7B Instruct model. 194
7.7 Frequency of INTACT's replacement choice for the test set of TAB with the Llama 3.1 8B Instruct model. 195

List of Tables

2.1 Datasets 27

3.1 Comparison of widely-used BERT-based language models 40
3.2 Correlation between recall and TRIR with different attacker's resources for the different anonymization methods. 65

4.1 TIC calculation runtime with different context masking strategies 91

5.1 PETRE-enhancing examples. Dashed underline corresponds to terms masked by the starting anonymization method, whereas solid underline shows additional terms masked by PETRE. 122
5.2 Mean PETRE runtime per document for each starting anonymization 125
5.3 Results for St.NER7 and Presidio with PETRE in the TAB dataset 131

6.1 Example of anonymized outcomes of LLM-based methods. Each PII detected by each method is replaced by the word *SENSITIVE*. 151
6.2 Average tokens (`prompt + generation = total`) for anonymizing a document across all method-model pairs. 153

7.1 One-shot prompt for replacement generation. 175

7.2 Example sentences and replacements used in the one-shot in-context learning setup for replacement generation. . . . 176
7.3 Prompt for guessing a PII term based on a replacement candidate and the rest of the text with replacements instead of PIIs. 178
7.4 Statistics of using entity type as replacement in TAB in cases where the inference attack successfully guesses all LLM-generated (or rule-based) replacements proposed. . . 181
7.5 Human evaluation of replacements on the *truthfulness* and *abstraction* level for three selected sanitization methods: Presidio, Dou et al and INTACT (our approach). 191
7.6 Utility and privacy evaluation for LLM generalizability on the TAB test subset with Llama 3.1 8B. (↑ for higher is better, ↓ for lower is better) 193

List of Abbreviations

SDC	Statistical Disclosure Control
PPDP	Privacy-Preserving Data Publishing
NLP	Natural Language Processing
NER	Named Entity Recognition
LM	Language Model
LLM	Large Language Model
MLM	Masked Language Modeling
IR	Information Retrieval
IC	Information Content

A mi Yayo y Amama

Chapter 1

Introduction

1.1 Motivation

Text is a ubiquitous mean of sharing information among humans, and defines the standard encoding for most unstructured or semi-structured knowledge sources. The availability of text data is therefore crucial for many data mining and machine learning tasks. Examples of textual sources employed for research include medical reports used in pharmacological studies (Jin et al., 2022; Meystre et al., 2010), publications in social networks employed for classification (Guo and Sarker, 2023; Rivas and Hristidis, 2021) or contextual analysis (Dolatabadi et al., 2023; Gutiérrez-Batista et al., 2018), and reviews or comments leveraged for customer satisfaction analysis (Alantari et al., 2022; Zhao, Xu, and Wang, 2019).

However, text data often contain sensitive information about individuals, as in the case of medical reports (U.S. Department of Labor, 2004) or court cases (Cornell Law School, 2007). Sharing or releasing such textual datasets therefore poses a direct conflict with privacy regulations, including, but not limited to, the European General Data Protection Regulation (GDPR) (European Commission, 2016). For instance, in the aforementioned medical reports, their direct release could expose highly confidential patient information (e.g., sensitive diagnoses), which

would adversely affect individuals' standing in society (e.g., personal relationships or employment opportunities).

To address privacy issues, data holders can obtain the informed consent of the individuals the data refer to (which may be infeasible), or anonymize the data, a process by which it should not be possible to attribute the data to concrete subjects. The latter makes the data no longer personal and, therefore, outside the scope of regulations such as the GDPR. Anonymization of structured data (i.e., tabular datasets) is a mature field, with a broad range of methods for statistical disclosure control (SDC) (Hundepool et al., 2012; Samarati, 2001; Li, Li, and Venkatasubramanian, 2006; Machanavajjhala et al., 2007; Dwork, 2006; Domingo-Ferrer, Sánchez, and Blanco-Justicia, 2021), including multiple privacy models (such as k-anonymity (Samarati, 2001) or ϵ-differential privacy (Dwork, 2006)) that offer formal privacy guarantees. The information to protect are attributes that allow to identify the corresponding individual. These identifiers should either be suppressed or masked using techniques such as generalization, aggregation or noise addition. In structured databases, identifying these attributes is relatively straightforward since the data is organized in attribute-basis (*i.e.*, one column per attribute).

However, the anonymization of unstructured data such as plain text is a far more intricate task (Lison et al., 2021; Csányi et al., 2021). Difficulties arise because identifying attributes appearing in textual resources are unbounded and, usually, not obviously associated to the corresponding subject. As a consequence, automatic text anonymization methods are significantly less mature than those meant for structured datasets (Lison et al., 2021; Csányi et al., 2021), and rarely providing any privacy guarantee.

In fact, the detection of identifying attributes that was easy for structured databases, becomes an intricate and demanding challenge for textual data. Real-world implementations of this task often rely on manual annotations made by one or multiple human experts (Maduell et al.,

2023; Lima-López et al., 2023; Bier et al., 2009). Some mechanisms have been proposed to mitigate the arduous work of manual anonymization. They consist mainly in preliminary annotations of potentially disclosive information grounded on Named Entity Recognition (NER) tools, with the aim of decreasing the volume of manual annotations (Pilán et al., 2022; Papadopoulou et al., 2022). Originally developed in the field of Natural Language Processing (NLP) Jehangir, Radhakrishnan, and Agarwal, 2023; Nadeau and Sekine, 2007, NER detects and classifies named entities -—such as person names, organizations, dates, and drug names—- into predefined categories. Over time, NER has gained popularity as a text anonymization approach, focusing on detecting the most disclosure-prone entity types.

However, NER-based text anonymization presents multiple drawbacks. First, as many authors have argued (Sánchez and Batet, 2016; Hassan, Sánchez, and Domingo-Ferrer, 2023; Lison et al., 2021; Csányi et al., 2021; Pilán et al., 2022), masking only predefined named entities fails to adequately mitigate re-identification risk, since any term can be disclosive. Moreover, NER-based methods lack flexibility -—they mask all detected entities uniformly, preventing data holders from defining specific privacy requirements.

In contrast, text anonymization methods proposed in the area of privacy-preserving data publishing (PPDP) (Hassan, Sánchez, and Domingo-Ferrer, 2023; Papadopoulou et al., 2022; Sánchez and Batet, 2017; Sánchez and Batet, 2016; Chakaravarthy et al., 2008; Cumby and Ghani, 2011; Anandan et al., 2012; Mosallanezhad, Beigi, and Liu, 2019; Staddon, Golle, and Zimny, 2007; Fernandes, Dras, and McIver, 2019) consider any information that jeopardizes subject's anonymity and often allow to configure the desired protection level. However, the distortion added by some of these methods and several scalability issues make them unfeasible in many real-world scenarios (Lison et al., 2021).

Furthermore, most approaches in the text anonymization literature focus only on the detection of disclosive terms, using straightforward

techniques for their masking. That is, suppress them or replace them by generic semantic categories indicating named entity types such as *PERSON* or *LOCATION* (Johnson, Bulgarelli, and Pollard, 2020; Yang and Garibaldi, 2015; Yogarajan, Mayo, and Pfahringer, 2018; Huang et al., 2020; Mamede, Baptista, and Dias, 2016; Bubeck et al., 2023; Papadopoulou et al., 2022; Pilán et al., 2022). This relatively crude masking can undermine the utility of the protected data for research and analysis.

Evaluating the performance of automatic text anonymization methods presents significant challenges, mirroring the inherent difficulties of the anonymization task itself. Like in the SDC literature (Hundepool et al., 2012), effective text anonymization requires balancing the trade-off between the attained level of privacy and the preservation of data utility. On the one hand, privacy is inversely proportional to re-identification risk, which corresponds to the probability of inferring who the data belongs to. 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 the 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 goal behind data releases.

For tabular databases, evaluation of anonymization methods is performed by measuring the re-identification risk and the statistical utility of the anonymized data. Privacy assessment can be conducted empirically by subjecting the anonymized dataset to re-identification attacks, such as record linkage (Domingo-Ferrer and Torra, 2003; Nin, Herranz, and Torra, 2007; Torra, Abowd, and Domingo-Ferrer, 2006; Torra and Stokes, 2012), where the risk is measured as the proportion of records

that can be successfully matched to their original identities. However, many approaches provide an *ex ante* formal privacy guarantee, not requiring an explicit privacy evaluation. For instance, k-anonymity (Samarati, 2001) ensures that the probability of re-identification does not exceed $\frac{1}{k}$. On the other hand, utility is quantified by comparing the original data with their anonymized counterparts (Domingo-Ferrer, Sánchez, and Soria-Comas, 2016). Metrics for utility 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.

For textual data, the vast majority of anonymization techniques (and all NER-based, particularly) do not provide privacy guarantees, so *ex post* evaluation is required to know the level of protection achieved. The *de facto* way to evaluate automatic text anonymization methods is to compare them with manual anonymization (Meystre et al., 2010; Aberdeen et al., 2010; Dernoncourt et al., 2017; Hassan, Domingo-Ferrer, and Soria-Comas, 2018; Johnson, Bulgarelli, and Pollard, 2020; Liu et al., 2017; Mamede, Baptista, and Dias, 2016; Neamatullah et al., 2008; Szarvas, Farkas, and Busa-Fekete, 2007; Yang and Garibaldi, 2015; Sánchez and Batet, 2016), which is assumed to define the ground truth. Particularly, the information retrieval (IR) metrics *recall* and *precision* are employed to measure the performance of text anonymization. Recall assesses the proportion of re-identifying terms (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. Conversely, 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. Nevertheless, these metrics have severe limitations as evaluation means for text anonymization. First, they assume there is a single best solution, while in data anonymization multiple masking choices could be equally valid

in preventing re-identification (Lison et al., 2021). Secondly, the manual annotations that are used as ground truth are prone to errors, omissions, and inconsistencies (Bier et al., 2009). Thirdly, whereas both metrics uniformly account the contribution of terms, not all missed/unnecessarily masked terms contribute equally to the re-identification risk/utility loss (Pilán et al., 2022). Additionally, these metrics are limited to suppression-based masking, which does not allow the assessment of replacement-based masking.

On the other hand, Language Models (LMs) (Devlin et al., 2019; Sanh et al., 2019; Liu et al., 2019; Wang et al., n.d.; Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023; Reid et al., 2024) have recently emerged as powerful tools for text processing and generation. They exhibit advanced proficiency in natural language understanding, positioning them as promising candidates for addressing key challenges in text anonymization. For instance, their ability to discern nuanced linguistic patterns can enable precise detection of (identifying) information tied to the individual to protect, what is crucial for ensuring that privacy requirements are met and avoiding unnecessarily hampering utility. By leveraging these capabilities, LMs can enhance both document protection and evaluation of privacy and utility while offering practical solutions.

1.2 Contributions

In this thesis, we propose automatic and privacy-grounded methods for both text evaluation and anonymization, leveraging the capabilities of language models to process and characterize text from a risk perspective. Specifically, we propose:

- *Text Re-Identification Attack* (TRIA), a method analogous to record linkage but specifically designed for textual data. Based on this attack, we define *Text Re-Identification Risk* (TRIR), a privacy metric that empirically quantifies the re-identification risk of a set of

documents. Unlike recall, TRIR directly measures re-identification risk, is fully automatic (requires no human annotations), and can be configured to reflect the knowledge and resources expected for an attacker.

- *Text Preserved Information* (TPI), a utility metric that measures the proportion of information from the original document preserved by the anonymized version, mirroring the techniques used for structured data. In contrast to precision, TPI focuses on the comparison between anonymized and original documents and operates entirely automatically.

- *Privacy Enhancement for Text via Risk-oriented Explainability* (PETRE), a method designed to strengthen the privacy protection of anonymized documents by enforcing an empirical privacy guarantee. PETRE operates in an automatic way and offers an intuitive configuration of the desired privacy level based on probabilistic k-anonymity Soria-Comas and Domingo-Ferrer, 2012.

- We conduct a comprehensive comparative study of state-of-the-art methods for LLM-based anonymization. In addition, we propose a novel method building on the strengths and limitations identified in our analysis.

- *INference-guided Truthful sAnitization for Clear Text* (INTACT), a technique for the generation of truthful replacements for disclosive terms. Alongside this, we propose *Text Preserved Similarity* (TPS) a utility metric specifically designed for anonymizations using replacement-based masking.

1.3 Outline

This thesis is divided into 8 chapters, organized as follows:

- **Chapter 2**: provides background on (text) anonymization, its evaluation, and language models.
- **Chapter 3**: introduces our machine learning-based re-identification attack, TRIA, and the associated privacy metric, TRIR.
- **Chapter 4**: presents our utility metric for text anonymization, TPI.
- **Chapter 5**: describes our privacy-enhancing method based on explainability and re-identification risk, PETRE.
- **Chapter 6**: reports our analysis of current LLM-based text anonymization approaches and our novel method based on the strengths and limitations detected.
- **Chapter 7**: introduces our approach to truthful replacement generation and selection, INTACT, along with our replacement-compatible utility metric, TPS.
- **Chapter 8**: provides a summary of the key contributions of this thesis, list the associated publications and defines the potential directions for future research.

Chapter 2

Background

2.1 Anonymization

Anonymization consists in irreversibly masking a dataset to preclude the identification of the referred entities, such as individuals or organizations. Because adequately anonymized data cannot be unequivocally attributed to specific individuals, they are no longer considered personal and, therefore, outside the scope of privacy regulations such as the GDPR (European Commission, 2016). The data distortion (or removal) introduced by the masking process necessarily impairs the informativeness/utility of the protected outcomes proportionally to the level of privacy attained. This defines a trade-off between privacy protection and data utility preservation that anonymization methods strive to optimize.

In the context of anonymization, data values can be classified into the following types:

- *Identifiers*: They allow to unequivocally identify the individual they refer to. Examples of identifiers include the complete name of an individual or IDs.

- *Quasi-identifiers*: Even though they do not enable unequivocal re-identification in isolation, they can when evaluated in combination with other quasi-identifiers, because those combinations are unequivocally associated to specific individuals in public identified sources (*e.g.*, electoral rolls). The public sources that can be leveraged for re-identification are known as *background knowledge*. Examples of quasi-identifiers are zip code, age or job name.

- *Confidential values*: Sensitive data about the individual to be protected that should not be attributed to it. Under the scope of the GDPR (European Commission, 2016), these include religious or political opinions, sexual orientation or healthcare information.

- *Non-confidential data*: They neither contribute towards re-identification nor disclose sensitive information. They can be either general terms (*e.g.*, stop words) or information not related to the individual of interest.

On these basis, utility-preserving anonymization consists in removing all identifiers, masking the minimum amount of quasi-identifiers needed to prevent re-identification, and preserving confidential and non-confidential values. Masking typically consists of suppressing attribute values or replacing them with generalizations. Simply removing identifiers does not qualify as anonymization –as it does not effectively prevent re-identification–, and is therefore considered de-identification (Lison et al., 2021; Chevrier et al., 2019).

2.1.1 Structured data anonymization

The anonymization task varies significantly depending on the type of data. In structured databases, individuals are described as records, each one containing a fixed (and usually reduced) set of attributes. This structure makes it relatively straightforward for a human expert to classify such attributes into identifiers and quasi-identifiers. Anonymization

methods are then applied on the basis of this attribute classification, suppressing identifiers and seeking to produce the quasi-identifiers masking that best preserves the utility of the data while enforcing a certain level of privacy protection. These methods commonly enforce a privacy model, which provides clear formal privacy guarantees on the protected outcomes. In particular, the well-known k-anonymity model requires each individual in the anonymized dataset to be indistinguishable from at least $k-1$ other individuals (Samarati, 2001) with respect to their quasi-identifying values. As a result k-anonymity upperbounds the re-identification probability to that of Eq. 2.1. By setting the value of k, users can tune the trade-off between privacy protection and utility preservation.

$$\textit{Re-identification probability} \leq \frac{1}{k} \tag{2.1}$$

The same guarantee against re-identification offered by k-anonymity can also be attained through probabilistic k-anonymity (Soria-Comas and Domingo-Ferrer, 2012), a relaxation that removes the indistinguishability requirement but keeps limiting the re-identification risk to that shown in Eq. 2.1. In practice, this means that records do not need to be indistinguishable with respect to their quasi-identifiers as long as the desired probability of re-identification is met. This opens the door to use masking operations other than suppression or common-ground generalizations. Other popular privacy model is ϵ-differential privacy (Dwork, 2006), which guarantees that the presence or absence of any single individual in the dataset has a negligible impact on the outcome of the analysis. Formally, a randomized mechanism $\mathcal{M}$ satisfies ϵ-differential privacy if, for all pairs of neighboring datasets D and D' differing by at most one record, and for all subsets of outputs $S \subseteq \text{Range}(\mathcal{M})$, the following inequality holds:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D') \in S] \tag{2.2}$$

Here, ϵ is the *privacy budget*, which quantifies the level of privacy guarantee –smaller values of ϵ enforce stronger privacy.

2.1.2 Text anonymization

Text anonymization is a far more intricate task than the anonymization of structured data. Nature of language, characterized by its inherent ambiguity, nuances, complex semantics and reliance on context, renders the accurate detection of the personally identifiable information (PII) –which encompasses both identifiers and quasi-identifiers– to be masked an unresolved challenge (Lison et al., 2021; Csányi et al., 2021).

The most well-established approaches for (semi-)automatic text anonymization are those based on NER (Johnson, Bulgarelli, and Pollard, 2020; Meystre et al., 2010; Yang and Garibaldi, 2015; Hassan, Domingo-Ferrer, and Soria-Comas, 2018; Neamatullah et al., 2008; Yogarajan, Mayo, and Pfahringer, 2018; Huang et al., 2020; Mamede, Baptista, and Dias, 2016; Bubeck et al., 2023; Patsakis and Lykousas, 2023). NER initially emerged as an NLP task Jehangir, Radhakrishnan, and Agarwal, 2023; Nadeau and Sekine, 2007, focusing on the detection and classification of named entities. These entities typically include specific categories such as person names, organizations, locations, dates, numerical expressions, and domain-specific terms like drug names or medical codes. It is commonly employed as a pre-processing step in NLP pipelines, serving as the foundation for more complex tasks such as relation extraction, entity linking, question answering, and information retrieval.

For text anonymization, NER is used under the assumption that (quasi-)identifiers correspond to certain types of named entities (*e.g.*, names of people, locations and dates). Then, NER-based anonymization consists in leveraging a machine learning model trained to detect those named entities and replacing the detected ones by their types (*e.g.*, PERSON,

LOCATION or DATE). In very recent works, researchers started leveraging the exceptional NER capabilities of large language models (LLMs) (Bubeck et al., 2023; Singhal et al., 2024) for anonymization purposes.

Although applying NER-based anonymization to any corpus (or even a single document) is straightforward, it also presents multiple limitations (Sánchez and Batet, 2016; Hassan, Sánchez, and Domingo-Ferrer, 2023; Lison et al., 2021; Csányi et al., 2021; Pilán et al., 2022; Papadopoulou et al., 2022). First, the assumption that PIIs exclusively correspond to specific types of named entities is fundamentally flawed. It is well-known in the privacy literature that *any* term or combination of terms –without being restricted by type– could potentially serve as a (quasi-)identifier, as long as it is related to the individual to be protected and appears in identified public sources (Hundepool et al., 2012). Consequently, NER-based anonymization usually leaves many (quasi-)identifying terms unmasked (*i.e.*, those not being named entities), thereby resulting in significant disclosure risks (Lison et al., 2021; Patsakis and Lykousas, 2023; Nyffenegger, Stürmer, and Niklaus, 2023). Secondly, NER-based protection masks the pre-defined types of named entities independently of the data holder's privacy requirements and the subject to be protected. On the contrary to all privacy models and methods in the literature, on privacy in structured databases (Hundepool et al., 2012), NER-based methods do not define parameters to set the desired level of privacy (or to configure the trade-off between privacy and utility).

Alternatively to NER-based approaches, there exist some text anonymization methods that are better aligned with the principles of the privacy-preserving data publishing (PPDP) field (Staddon, Golle, and Zimny, 2007; Fernandes, Dras, and McIver, 2019; Sánchez and Batet, 2016; Sánchez and Batet, 2017; Mosallanezhad, Beigi, and Liu, 2019; Hassan, Sánchez, and Domingo-Ferrer, 2023; Papadopoulou et al., 2022). On the contrary to NER techniques, these approaches focus masking on the subjects to be protected and are not restricted to predefined entity types.

As a result, they offer more robust protection and they usually allow to configure the trade-off between utility and privacy (Lison et al., 2021).

Most of these methods rely on analyzing the correlation or mutual information between the terms appearing in the documents and the subjects to be protected. The smaller the set of correlated individuals, the more disclosive/identifying the terms will be. Approaches differ in how these correlations are obtained. Several authors estimate them from the co-occurrence of terms and identities in the Web (Staddon, Golle, and Zimny, 2007; Sánchez and Batet, 2016; Sánchez and Batet, 2017). Others gather statistics from structured databases, such as Wikidata (Papadopoulou et al., 2022). Finally, in Hassan, Sánchez, and Domingo-Ferrer, 2023 the authors train a word embedding model on the documents to be protected, and use the similarity between the obtained embeddings as a measure of correlation. Nonetheless, all these methods face complexity and scalability issues, caused by either the cost of performing Web searches, creating an in-domain structured database or training a model from scratch. Therefore, they are much more difficult to apply in practice than NER-based approaches. It fact, the only practical PPDP-oriented approach we found is (Hassan, Sánchez, and Domingo-Ferrer, 2023), since training the embedding model from scratch is a costly but feasible task.

The advent of LLMs has recently sparked interest in their potential for text anonymization (Bubeck et al., 2023; Patsakis and Lykousas, 2023; Singhal et al., 2024). In these approaches, LLMs are prompted for PII detection. Depending on the prompt design, a more NER-based or PPDP-oriented anonymization is obtained. Preliminary results suggest that the extensive knowledge of pre-trained LLMs effectively provides more nuanced and context-aware PII detection than traditional approaches. Nevertheless, the experimental evaluation reported so far (Bubeck et al.,

2023; Patsakis and Lykousas, 2023; Singhal et al., 2024) lacks standardization due to being done ad hoc and with significantly different evaluation data, metrics, and models. Moreover, no comprehensive comparison of LLM-based methods has been reported so far. Therefore, there is no solid evidence on how much better LLM-based anonymization is, and how different prompts affect the protected outcome.

2.1.3 Redaction and sanitization

There are two masking strategies for text anonymization: *redaction* and *sanitization*. Whereas document redaction directly suppresses or blacks out terms, thereby maximizing privacy at the cost of the utility and readability of the protected documents, text sanitization replaces terms with broader, more abstract generalizations (Bier et al., 2009). The latter achieves a better privacy-utility trade-off while preserving the truthfulness of the information, as generalizations, by definition, should retain the core semantic content of the original terms.

The majority of current text anonymization methods use either strict redaction or replace the disclosive terms by their corresponding named entity category (such as *PERSON* or *LOCATION*), which is a basic form of sanitization (Lison et al., 2021). Both strategies retain little to no meaningful information from the original terms. More sophisticated text sanitization methods have traditionally leveraged structured knowledge bases (Cumby and Ghani, 2011; Anandan et al., 2012; Anandan and Clifton, 2011; Sánchez, Batet, and Viejo, 2013a; Sánchez, Batet, and Viejo, 2013b; Sánchez and Batet, 2017; Hassan, Sánchez, and Domingo-Ferrer, 2023; Olstad, Papadopoulou, and Lison, 2023), such as WordNet (Fellbaum, 1998), the now defunct Open Directory Project (ODP)[1] or YAGO (Fabian, Gjergji, and Gerhard, 2007), to retrieve term generalizations to be used as replacements. Specifically, generalizations are obtained from the superclasses (*i.e.*, hypernyms) in the ontology of the terms to

[1] `http://odp.org/`

be masked. Nonetheless, sanitization methods based on "static" knowledge bases have limited applicability due to their relatively narrow coverage of disclosive terms. This limited coverage is particularly salient for named entities, which are rarely present in such knowledge bases, while they are often one of the most crucial type of terms to sanitize. Moreover, the largest knowledge bases, such as YAGO or ODP may not provide consistent generalizations, as they originate from arbitrary categories, labels and descriptions (such as Wikipedia categories) that do not fulfill the subsumption property; and this may hamper the truthfulness of the sanitized outcomes.

More recently, a number of studies have also explored language models for protecting sensitive terms (Arranz et al., 2022; Dou et al., 2024). Such models successfully overcome the aforementioned coverage issues of knowledge bases as they can (1) generate replacements for any text span and (2) can ensure those replacements can be inserted in the document in a grammatically meaningful manner. However, there is a lack of development in techniques that seek to ensure that the replacements are truthful generalizations of the terms to be masked. This oversight raises concerns about potential semantic distortions, where replacements may perturb rather than abstract the document's intended meaning.

2.1.4 Perturbative methods

Although text anonymization aims to produce a (readable) version of the document but with masked terms, there are also related methods that perturb the structure and content of the document (Fernandes, Dras, and McIver, 2019; Mosallanezhad, Beigi, and Liu, 2019; Abril, Navarro-Arribas, and Torra, 2011; Volodina et al., 2020). These approaches often prioritize privacy guarantees over strict readability and truthfulness. Notable examples include PPDP-oriented methods grounded on differential privacy (Fernandes, Dras, and McIver, 2019) or reinforcement

learning (Mosallanezhad, Beigi, and Liu, 2019), that transform the documents into embeddings or bags-of-words, and perturb them by using a neural model or noise. Additionally, approaches that replace sensitive terms by other than generalizations –thus falling outside the scope of the text sanitization. These include entity swapping (Abril, Navarro-Arribas, and Torra, 2011), random modification (Volodina et al., 2020) and Presidio's synthetic replacement[2].

Although these techniques preserve statistical properties across a corpus of documents, they compromise the semantic integrity of individual documents. Consequently, they produce non-truthful outcomes that might be unacceptable in some secondary uses (*e.g.*, clinical research), because they invalidate conclusions extracted on document-basis. Furthermore, the non-readable output formats of certain methods (*e.g.*, bags-of-words or embeddings) further limit their usability.

2.2 Evaluation

The quality of data anonymization is characterized by the trade-off between privacy protection and utility preservation, with anonymization methods often seeking to maximize utility for a given privacy level. Thus, accurate privacy and utility evaluation is required for data holders to make an informed decision on the method to be used for their data release. On the one hand, anonymization should prevent re-identification, meaning that the protected data cannot be unequivocally associated with a subject's identity. Thus, *privacy evaluation metrics* should measure the probability of re-identification. On the other hand, *data utility preservation* is, in theory, determined by how closely the results obtained from anonymized data align with those derived from the original data for possible downstream tasks. However, in practice, since datasets may

[2] `https://github.com/microsoft/presidio/blob/main/docs/samples/python/synth_data_with_openai.ipynb`

serve multiple, often unpredictable, downstream purposes, utility is typically evaluated by measuring the similarity between the anonymized and original data, which is assumed to correlate with the similarity of outcomes. This similarity is primarily determined by the extent of information loss introduced by anonymization.

2.2.1 Structured data

In tabular data, privacy models such as k-anonymity (Samarati, 2001) guarantee an *ex ante* upper bound for re-identification probability of $\frac{1}{k}$. For methods without formal privacy guarantees, an empirical evaluation of the residual re-identification risk is needed. To this end, the record linkage re-identification attack is the most widespread measure of risk (Domingo-Ferrer and Torra, 2003; Nin, Herranz, and Torra, 2007; Torra, Abowd, and Domingo-Ferrer, 2006; Torra and Stokes, 2012; Abril, Navarro-Arribas, and Torra, 2012). Record linkage seeks to re-identify protected records by linking the masked quasi-identifying attribute values in the anonymized database with those in the background knowledge, which contains the original quasi-identifying values of known individuals. If the linkage is successful, the protected record would be unequivocally associated with the identity of the corresponding individual. Therefore, the re-identification risk of an anonymized database is defined as the proportion of correctly linked records:

$$\text{Re-identification risk} = \frac{\#CorrectlyLinkedRecords}{\#Records} \tag{2.3}$$

If quasi-identifying attributes in the anonymized and background databases unequivocally overlap, record linkage can be as straightforward as linking each anonymized record to the record in the background knowledge with the most similar quasi-identifying attribute values (Hundepool et al., 2012).

Data utility assessment measures the distributional and statistical differences between the original and anonymized datasets. These differences can be computed at 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 include 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 original and anonymized attribute values, a variety of dissimilarity metrics can be employed according to the attribute 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 (Shannon, 1948; Domingo-Ferrer, 2001) or ontology-based metrics (Domingo-Ferrer, Sánchez, and Rufian-Torrell, 2013; Martínez, Sánchez, and Valls, 2013; Martínez, Sánchez, and Valls, 2012).

2.2.2 Textual data

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 those 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 it is done for completeness in NER tasks, that is, by quantifying the *recall* (Meystre et al., 2010; Aberdeen et al., 2010; Dernoncourt et al., 2017; Hassan, Domingo-Ferrer, and Soria-Comas,

2018; Johnson, Bulgarelli, and Pollard, 2020; Liu et al., 2017; Mamede, Baptista, and Dias, 2016; Neamatullah et al., 2008; Szarvas, Farkas, and Busa-Fekete, 2007; Yang and Garibaldi, 2015; Sánchez and Batet, 2016). Recall is defined as the proportion of elements of interest that were masked by the system to be evaluated w.r.t. a ground truth:

$$Recall = \frac{\#TruePositives}{\#TruePositives + \#FalseNegatives} \tag{2.4}$$

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 terms that should be masked according to human experts' opinions, which are considered the ground truth. Specifically, recall does not assess the residual re-identification risk of anonymized texts, but only compares the anonymized outcomes against manual annotations. These annotations are subjective, and may be prone to errors and omissions (Lison et al., 2021; Pilán et al., 2022). Moreover, manual annotation is time consuming, and necessarily involves several human experts, whose annotations should be integrated through a non-trivial and non-univocal process (Pilán et al., 2022). Another drawback of recall is that it assumes the existence of a single gold standard annotation. Even though this assumption may be reasonable for NER or other NLP-oriented tasks, it does not hold for text anonymization, where several masking choices (each one encompassing a different combination of quasi-identifying terms) could be equally valid to prevent re-identification (Lison et al., 2021). Finally, recall-based evaluation assumes that all identified/missed elements contribute equally to mitigate/increase re-identification risk, which is rarely the case (Pilán et al., 2022). Indeed, failing to mask an identifier (such as the full name of an individual) is much more disclosive than just missing a quasi-identifier such as the birthdate or whereabouts.

Being the complementary of recall, the *precision* metric is used in the literature to quantify the utility preserved in the anonymized outcomes

(Meystre et al., 2010; Aberdeen et al., 2010; Dernoncourt et al., 2017; Hassan, Domingo-Ferrer, and Soria-Comas, 2018; Johnson, Bulgarelli, and Pollard, 2020; Liu et al., 2017; Mamede, Baptista, and Dias, 2016; Neamatullah et al., 2008; Szarvas, Farkas, and Busa-Fekete, 2007; Yang and Garibaldi, 2015; Sánchez and Batet, 2016). Precision calculates the ratio of correctly masked terms by the anonymization method:

$$Precision = \frac{\#TruePositives}{\#TruePositives + \#FalsePositives} \tag{2.5}$$

where #TruePositives is the number of text spans that were correctly masked by the anonymization method w.r.t. 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, precision relies on the premise of a single ground truth, even though multiple masking choices can be equally effective at preventing re-identification while preserving different levels of utility (Lison et al., 2021).

To evaluate different approaches using a single metric, precision and recall are combined into the *F-score*, a measure also adopted from the

NER literature. F-score (Eq. 2.6), is the weighted harmonic mean of precision and recall, offering a comprehensive assessment of anonymization effectiveness from both utility and privacy perspectives.. The β factor determines the weight assigned to recall relative to precision, thereby reflecting the prioritization of privacy over utility.

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} \tag{2.6}$$

Even though the F_1-score –which gives equal weight to both precision and recall– have been traditionally employed, several authors have recently advocated for the use of the F_2-score over the F_1-score (Ferrández et al., 2012; Mendels, 2020). The reason is that recall should take precedence over precision in text anonymization because a low recall (regardless of the level of attained precision) could render anonymization ineffective.

Although precision, recall, and F-score are widely used to evaluate text anonymization, these metrics are fundamentally limited to assessing the quality of the PII detection step, and are unable to assess the suitability of the replacements used in sanitization. Indeed, evaluating sanitized texts against a manually constructed ground truth is challenging because i) there can be multiple valid (*i.e.*, non-disclosive) combinations of replacements for a given text and, ii) replacement candidates are, quite often, unbounded. Establishing a single ground truth for the evaluation is thus impossible. Moreover, requesting human annotators to provide privacy-preserving replacements (or, at the very least, chose appropriate replacements for predefined lists of candidates) requires even more effort than the –already quite costly– detection of disclosive text spans Olstad, Papadopoulou, and Lison, 2023. As a result, the few sanitization strategies proposed in the literature are largely untested Cumby and Ghani, 2011; Anandan et al., 2012; Anandan and Clifton, 2011; Sánchez, Batet, and Viejo, 2013a; Sánchez, Batet, and Viejo, 2013b; Sánchez and Batet, 2017; Hassan, Sánchez, and Domingo-Ferrer, 2023;

Olstad, Papadopoulou, and Lison, 2023.

2.3 Language Models

Language models are statistical models grounded on neural networks that assign probabilities to sequences of words and constitute the state-of-the-art in many NLP tasks, such as translation, text generation, sentiment analysis, summarization, question answering, or text classification (Devlin et al., 2019; Raffel et al., 2020; Bommasani et al., 2021; Brown et al., 2020; OpenAI, 2023; Reid et al., 2024; Touvron et al., 2023; DeepSeek-AI et al., 2025).

Language models are typically based on the Transformer architecture (Vaswani et al., 2017; Bommasani et al., 2021; Raffel et al., 2020; Devlin et al., 2019), which was introduced as an encoder-decoder architecture for language translation, but has since become the main building block for generic language models. Transformers make it possible to efficiently derive contextualized vector representations (contextual embeddings) for each token (*i.e.*, word or word fragment) of a given sequence. Those embeddings are computed through multiple neural layers, each being represented by its own set of parameters. Compared to previous neural architectures based on recurrent layers (Hochreiter and Schmidhuber, 1997; Cho et al., 2014; Peters et al., 2018), Transformer models can handle long-range dependencies between tokens while simultaneously allowing tokens to be processed in parallel. Consequently, transformers are better at handling longer and more complex texts.

A language model is initially pre-trained with a corpus of documents (*e.g.*, Wikipedia and the BookCorpus (Zhu et al., 2015)) (Devlin et al., 2019), thereby capturing a broad range of text types and linguistic constructions. During pre-training, the model parameters are optimized in an unsupervised fashion 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). Pre-trained language models serve as foundational components in modern NLP systems, eliminating the need to build models from scratch. These models can be adapted to a particular task (such as document classification) through a fine-tuning process consisting of a supervised learning step on a task-specific corpus. This learning process is a standard neural network training, where the entire training dataset is iterated over multiple times (epochs), and multiple training samples (batches) are processed in each iteration. Thanks to the general linguistic and factual knowledge acquired during pre-training, satisfactory results can be achieved even if the amount of labeled data available for fine-tuning is small. In addition, results can be improved if, previously to fine-tuning, the language model is additionally pre-trained using the task data. This process, which is conducted with the same training method employed to general pre-training, adapts the model to the specific task domain.

In recent years, the HuggingFace's Transformers library (Wolf et al., 2020) has emerged as the standard free platform for the implementation, pre-training, fine-tuning, usage and sharing of (pre-trained) language models. One of the most popular and well-established transformer-based language models available are BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) and its variations, which either improve its performance (Liu et al., 2019) or reduce the model's size and, therefore, its computational requirements (Sanh et al., 2019). After fine-tuning, BERT and its variations have demonstrated human-level performance and competitiveness with larger language models (Sun et al., 2023) in multiple language-related tasks, including text classification. All these language models are freely available at Hugging Face's transformers library (Wolf et al., 2020) with pre-trained weights and an easy-to-use implementation.

Large language models (LLMs) are language models with billions to trillions of parameters (Brown et al., 2020; OpenAI, 2023; Touvron et al.,

2023; Reid et al., 2024). These models generate sequences by predicting the next token –representing a word or part of a word– one step at a time, in an autoregressive manner. They are pre-trained on vast and diverse corpus in the causal language modeling task, enabling them to learn complex linguistic patterns, acquire broad general knowledge and understand how to follow instructions. Once trained, LLMs can be guided by natural language prompts to perform specific tasks, and can even perform tasks not included in the pre-training data without fine-tuning. This approach has proven highly effective, often matching or even surpassing human performance across a wide range of applications (Bubeck et al., 2023; Tan et al., 2024; Gilardi, Alizadeh, and Kubli, 2023).

Interacting with a LLM typically involves a conversational format, alternating between the roles of *User* (the human participant) and *Assistant* (the model). Optionally, an initial predefined *System* prompt can be included to set the overall behavior or tone of the model. For automatic task solving, the standard approach is to start with a prompt describing the task and providing the task data (with User and/or System role/s) and then use the Assistant response as solution. These prompts may include examples of correct solutions, which help the model to understand what it is expected to do. Depending on the number of examples provided, prompts/tasks are categorized as *zero-shot* (no examples), *one-shot* (one example), or *few-shot* (a small number of examples). Both LLM capabilities and prompt quality have a significant influence on response accuracy.

Early studies in the usage of LLMs for text anonymization (Bubeck et al., 2023; Patsakis and Lykousas, 2023; Singhal et al., 2024) suggest that they might offer better protection than previous methods. However, due to the lack of a common and objective evaluation framework, the extent to which LLM-based text anonymization outperforms existing methods, as well as the influence of prompt variations on protected outcomes, remains unclear.

A related concern stems from the tendency of LLMs to memorize and potentially reproduce parts of their training data verbatim, which raises important privacy risks when that data includes personal information. This brings attention to a concept closely linked to text privacy: digital forgetting. Digital forgetting refers to procedures used to remove information from software systems. This term originated from privacy laws related to the right of erasure or the *right to be forgotten* (*e.g.* GDPR, Art. 17, (European Commission, 2016)), which enables data subjects to request service providers to delete information related to them from their systems. However, forgetting can be extended to any request (by data subjects, regulators, or the entity managing the software system) to remove any information or unwanted behavior from a software system. In the case of LLMs, this translates to any modification in the model or the training procedures to remove any influence of specific training data or specific training procedures on the model's capabilities or behavior. This is typically achieved through a process known as *unlearning*.

2.4 Datasets

In this thesis, we utilize three distinct datasets to conduct empirical experiments on text anonymization and its privacy and utility assessment. A summary of these datasets is provided in Table 2.1. In addition to describing the documents to anonymize, the table includes a column detailing the employed background knowledge. As will be discussed in Chapter 3, this background knowledge is essential for our privacy evaluation.

WikiActors: Presented in (Hassan, Sánchez, and Domingo-Ferrer, 2023) and available in a public GitHub repository[3], it consists of 18,672 Wikipedia articles corresponding to the "20th-century actors" category.

[3] `https://github.com/fadiabdulf/automatic_text_anonymization`

TABLE 2.1: Datasets

Name	#Protected documents	#Background knowledge documents	Chapter/s
WikiActors (Hassan, Sánchez, and Domingo-Ferrer, 2023)	50	50, 543, 581 and 1,952	3
Wiki553 (Papadopoulou et al., 2022)	553	553	4, 5 and 6
TAB (Pilán et al., 2022)	127 and 1,268	127 and 2,536	5 and 7

Each one provides biographical details of a movie actor or actress. Wikipedia biographies are frequently employed as evaluation data for text anonymization due to the substantial and varied amount of (quasi-)identifying terms they contain, making them a challenging testbed (Lison et al., 2021; Sánchez and Batet, 2016; Sánchez, Batet, and Viejo, 2013a; Staddon, Golle, and Zimny, 2007). We considered the article abstracts as the texts to be anonymized, and the article bodies, which provide extended details on the abstract's information, as the identified publicly available documents on the individuals to be protected (*i.e.*, background knowledge). This enables to define a worse-case scenario for privacy, where the background knowledge may be a superset of the data to protect. A subset of 50 article abstracts corresponding to 50 contemporary, popular and English-speaking actors was selected from this corpus as the set of documents to be anonymized. We also leverage the manually anonymized version of these abstracts, produced by the three authors of Hassan, Sánchez, and Domingo-Ferrer, 2023 using sound privacy-oriented annotation guidelines. The final annotation was agreed upon via majority voting, replacing annotated words with the term "SENSITIVE". For the background knowledge, we defined multiple subsets of the $\sim$19K articles' bodies, differing in size and overlap with the individuals to protect. Further details are provided in Section 3.3.2 of Chapter 3.

Wiki553: We leverage the general-purpose Wikipedia-based dataset presented in (Papadopoulou et al., 2022), which is available in a public GitHub repository[4]. It contains 553 articles, which have been randomly chosen from the set of biographies presented in (Lebret, Grangier, and Auli, 2016). The random sampling resulted in 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 comparatively small (such as WikiActors (Hassan, Sánchez, and Domingo-Ferrer, 2023)) and/or domain-specific (*e.g.*, medical texts (Johnson, Bulgarelli, and Pollard, 2020; Meystre et al., 2010; Yang and Garibaldi, 2015; Neamatullah et al., 2008) or the legal-oriented TAB corpus (Pilán et al., 2022)). As for the previous Wikipedia-based dataset, the abstracts are considered the sensitive text to be protected and the bodies are the publicly available background knowledge. Abstracts have been manually annotated by following guidelines[5] very similar to those described in (Pilán et al., 2022). Annotations encompassed the detection of disclosive terms, categorization into semantic groups (*e.g.*, person, date, organization, etc.), and determining whether its masking was necessary for concealing individual's identity. For the first two tasks, annotations of the spaCy NER tool[6] were given as the starting point, but annotators were instructed to make any necessary corrections. For the third task, annotators had to classify the detected text spans into identifiers, quasi-identifiers or non-identifiers, the latter being left in clear. Manual anonymizations were generated by replacing the terms marked for masking by their semantic category, as typical done in NER-based anonymization. The privacy goal provided to the annotators was in line with the standard for manual anonymizations, which involves masking the text to the extent that the corresponding individual becomes unrecognizable.

[4] `https://github.com/anthipapa/bootstrapping-anonymization`

[5] `https://github.com/anthipapa/bootstrapping-anonymization/blob/main/text-anonymization/annotation_guidelines.pdf`

[6] `https://spacy.io/api/entityrecognizer`

The attainment of that goal is defined by the annotator's own judgment and common sense. Documents were annotated by five annotators (one NLP researcher and four undergraduate students in law) using the *TagTog*[7] tool. Importantly, annotations achieved high character-level inter-annotator agreement, with Cohen's κ reaching 0.81 and Krippendorff's α measured at 0.73, both strong indicators of reliability and annotation quality. As far as we know, this is the largest general-purpose manually annotated dataset for text anonymization.

Text Anonymization Benchmark (TAB): We employed TAB (Pilán et al., 2022), a domain-specific dataset available in a public GitHub repository [8]. It comprises 1,268 English documents derived from European Court of Human Rights (ECHR) case files. Concretely, the *Procedure* section of each ruling, which is abundant in (quasi-)identifiers, was considered the sensitive text to be protected. Documents were manually annotated by 12 law students, who identified sensitive text spans and classified them into entity types, adhering to annotation guidelines very similar to those used for the previous Wikipedia-based dataset (Papadopoulou et al., 2022). Annotated terms were replaced with the corresponding entity types. For background knowledge, we employ the whole case's document except the *Procedure* section. This dataset has been leveraged in Chapters 5 and 7. However, chapters differ in the exact configuration of documents to anonymize and background knowledge. Chapter 5 utilizes the test subset of TAB, constituted by 127 *Procedure* sections to be anonymized. The corresponding 127 rulings (except the *Procedure* section) were employed as background knowledge (see Section 5.3.1 for details). Chapter 7 leverages the whole 1,268 rulings of the TAB dataset (including test, dev and train sets) as documents to anonymize. For background knowledge, a strict superset of 2,536 rulings (except the *Procedure* section) is leveraged (see Section 7.3.1 for details).

[7] `https://www.tagtog.net/`

[8] `https://github.com/NorskRegnesentral/text-anonymization-benchmark`

Chapter 3

Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack

As discussed in Section 2.2, the standard approach to evaluating privacy protection –based on recall from human annotations– has notable limitations. A comprehensive privacy assessment should instead focus on measuring the re-identification risk posed by the documents.

To tackle the limitations of recall and human-dependent evaluation of anonymized texts, in this chapter we propose TRIA, a *Text Re-Identification Attack* that adapts the principles of the record linkage technique –which was designed and exclusively applied to structured databases– to textual documents. Under this framework, we also propose TRIR, a *Text Re-Identification Risk* empirical metric based on the re-identification accuracy of TRIA, which overcomes the drawbacks of the non-risk-specific

recall-based privacy evaluation.

We define TRIA as a (highly-dimensional) multiclass classification task, where anonymized texts must be associated to individuals (classes) known by the attacker. To maximize re-identification capabilities, TRIA leverages state-of-the-art neural language models (Devlin et al., 2019) to aggregate and characterize the data sources that attackers may use as background knowledge to conduct re-identifications. These models have been shown to achieve human or above human proficiency in many language-related tasks, thus making our approach a realistic depiction of the work of an ideal human attacker. However, because the unconstrained use of large language models can be computationally expensive, we have carefully considered the means and computational resources expected to be available at the attacker's end in order to simulate a feasible re-identification attack.

We illustrate the effectiveness of our proposal by evaluating the level of privacy protection attained by a variety of text anonymization methods, together with a sample of human-based anonymization (Hassan, Sánchez, and Domingo-Ferrer, 2023) under several (ideal or more realistic) attack configurations. The empirical results reported here illustrate the deficiencies of NER-based anonymization (already discussed at a theoretical level in (Lison et al., 2021; Hassan, Sánchez, and Domingo-Ferrer, 2023)), and show substantial discrepancies against recall-based evaluations.

The contributions in this chapter have been published in (Manzanares-Salor, Sánchez, and Lison, 2024). Additionally, a preliminary version of this work was published in (Manzanares-Salor, Sánchez, and Lison, 2022).

The remainder of this chapter is organized as follows. Section 3.1 presents and formalizes TRIA and TRIR. Section 3.2 details the implementation details for TRIA. Section 3.3 reports and discusses empirical results on a variety of text anonymization methods and attack configurations. The last section gathers the conclusions.

3.1 Re-identification attack and risk metric for anonymized text

In the following we describe TRIA, a re-identification attack for (anonymized) texts, and TRIR, a metric based on TRIA's accuracy. Both aim to provide a practical solution to evaluate the actual privacy protection offered by anonymization methods for textual data. Compared to recall, our approach brings the following advantageous features:

- *Self-supervision*: TRIR does not require costly manual annotations as anonymization ground truth, which can be imperfect, potentially biased (if coming from a single annotator) or inconsistent (if coming from multiple annotators). Once the protected and background documents are set, our metric is fully automatic. Supervised learning only requires the label indicating the individual to be protected for each document.
- *Focus on risk*: Contrary to recall, our metric directly measures the empirical re-identification risk of (protected) documents as a whole, rather than assessing the coverage of individual annotations.
- *Flexibility*: Our metric can be tailored to take into account different resources available to attackers (see Section 3.1.2) thereby allowing to configure either ideal or more realistic attackers.

TRIA tries to re-identify the subjects referred in a collection of anonymized texts by leveraging a multiclass classifier, with one class for each individual known by the attacker. A set of identified and publicly available texts is employed for training the classifier. These texts encompass a population of known subjects in which the individuals referred to in the anonymized set are expected to belong. For instance, to re-identify anonymized medical reports from a city hospital, the attacker may use publicly available social network posts from the residents of that city. The set of known individuals should therefore be a

superset of the anonymized subjects. Anonymized documents would contain masked quasi-identifiers (*e.g.*, age intervals) and confidential attributes (*e.g.*, diagnoses) from unidentified individuals, whereas public texts would contain identifiers (*e.g.*, names and surnames) and clear quasi-identifiers (*e.g.*, specific ages) from known individuals. Deficiencies in anonymization would enable unequivocal matchings of the (quasi-)identifiers of both types of documents, allowing to re-identify the subjects in the protected documents and, subsequently, disclose their confidential information.

According to the above, our approach can be seen as an adaptation of the record linkage attack (designed for structured databases) to textual data, where words or text spans correspond to attribute values, documents correspond to records, and the classifier performs the linkage between the anonymized and public data.

3.1.1 Formalization

Let A_D be the set of anonymized (*i.e.*, non-identified) documents and B_D the set of identified public documents (*i.e.*, background knowledge). Each document refers to a specific subject, thus defining the sets of individuals A_I and B_I, and the mapping functions $F_A : A_D \rightarrow A_I$ and $F_B : B_D \rightarrow B_I$.

As in the original record linkage attack, we assume that $A_I \subseteq B_I$, which means that a re-identification function $F_C : A_D \rightarrow B_I$ exists, which matches protected documents with the corresponding known individuals. From the attacker's point of view, A_D, B_D, B_I and F_B are known, and A_I, F_A and F_C are unknown. Consequently, the goal of the attack is to exploit the similarities between A_D and B_D to obtain ${F_C}'$, an approximation of F_C. Figure 3.1 shows a graphical representation of this setting.

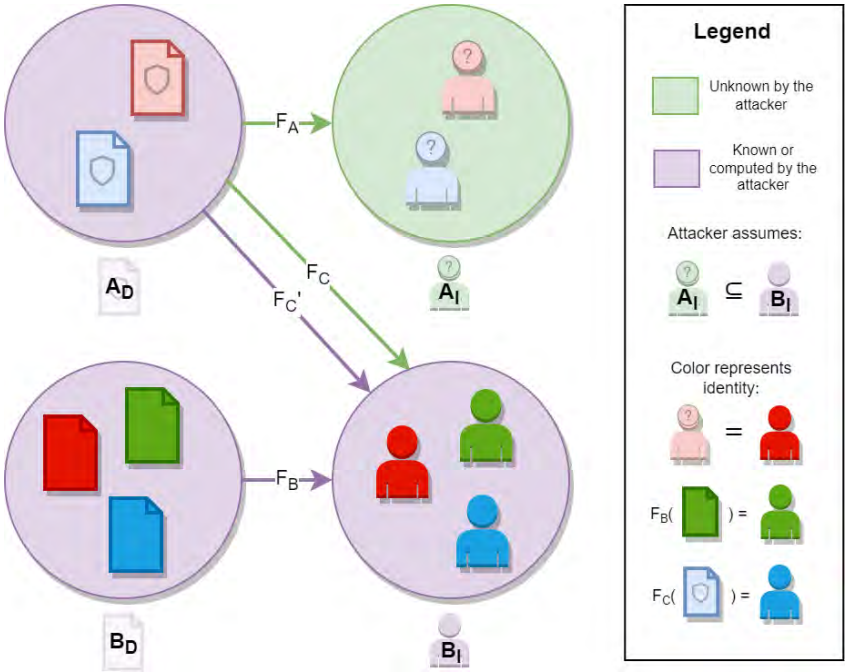


FIGURE 3.1: TRIA scenario, including individuals and documents sets from the point of view of the attacker.

Using the medical example above, the anonymized documents (A_D) could be medical reports from patients, and the individuals to be protected (A_I) would be those patients. Identified documents (B_D) are publicly available documents associated with the identities of individuals, such as posts by such patients about themselves on their social networks. Individuals linked to public documents (B_I) are known to the attacker (social networks are non-anonymous and publicly accessible), while patients (A_I) and their connection to specific medical reports – which correspond to F_C, the re-identification function–, are initially unknown.

Our proposal is formalized in Algorithm 1, which outputs the number of correct re-identifications achieved by TRIA for a collection of anonymized documents.

Algorithm 1: Re-identification risk assessment for anonymized text documents

Require: A_D, B_D, B_I, F_B, F_C

1: $classf \leftarrow build_classifier(B_D, B_I, F_B)$
2: $numReIds \leftarrow 0$ /* Correct re-identifications */
3: **for each** d in A_D **do**
4: $pred_ind \leftarrow classf.predict(d)$
/* TRIA predicted individual */
5: **if** $pred_ind == F_C(d)$ **then**
6: $numReIds \leftarrow numReIds + 1$
7: **end if**
8: **end for**
9: **return** $numReIds$

First, a multiclass classifier is built and trained to predict F_C (line 1, see Section 3.2.4 for details). Following the formal notation above, the classifier would implement $F_C{}'$ by learning which individuals from B_I correspond to the documents in B_D based on the knowledge (*i.e.*, data sources) available to the attacker. Afterwards, the classifier is tested in the anonymized set of documents A_D (line 4, refer to Section 3.2.5 for

details). A correct re-identification would happen if the prediction (*i.e.*, $F_C{}'$) matches F_C (lines 5-6). The total number of re-identifications is finally returned in line 9.

As in the record linkage method (Eq. 2.3), we assess the TRIR of A_D as the accuracy of TRIA:

$$TRIR = \frac{Correct}{|A_I|} \tag{3.1}$$

where *numReIds* is the number of correct re-identifications achieved by Algorithm 1, and $|A_I|$ is the number of protected individuals from the A_I set.

3.1.2 Attacker's resources

One of our goals for TRIA is to be flexible enough to encompass both ideal or more realistic attackers. This means considering the resources (*i.e.*, background knowledge and available computational resources) that attackers may (ideally or realistically) devote to the attack. This is in line with the GDPR[1], which states that, to assess the residual risk of anonymized data, one should account the reasonable means that can be employed to achieve re-identifications.

We next characterize and discuss the resources that attackers may use to conduct re-identification attacks, and that should be considered when building TRIA's classifier:

- *Background knowledge breadth*: Corresponds to the number and type of individuals (and corresponding documents) considered as candidates for re-identification or, in other words, to the size of the

[1]Extract from the GDPR's Recital 26: "[...] To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of processing and technological developments. [...]"

population to which the individuals to be protected belong. Formally, this is the B_I set (and the corresponding B_D) that the attacker determines as a superset of A_I. It is important to note that the assumption $A_I \subseteq B_I$ may not hold in practice because A_I is usually unknown to the attacker. In this case, the re-identification risk is limited by the proportion of overlap between A_I and B_I. Furthermore, the difficulty of the re-identification task would depend on the number of individuals and the intra-class similarity of both A_I and B_I. On the one hand, increasing the size of B_I would complicate the re-identification, since there are more individuals (classes) to distinguish. This behavior is inherent to any multiclass classification task. In addition, the more individuals in B_I not present in A_I the more prone to false positives the classifier would be; this increases the cost of the attack (because more classes must be learned) without expected benefits. On this basis, the goal of an attacker would be to define a set B_I with a size as close as possible to that of A_I while trying to fulfill $A_I \subseteq B_I$. On the other hand, a high intra-similarity between the individuals in B_I not present in A_I does not affect the accuracy of the attack since there are no instances for these individuals in A_D. Finally, with respect to intra-similarity between the individuals in B_I also present in A_I, and those not present in A_I, the lower the better, since we do not want the classifier to be confused by the individuals outside A_I.

- *Background knowledge depth*: Compasses the quantity and quality of the background information available to attackers for each individual or, in other words, the (quasi-)identifiers available in each document of B_D. Generally, the more information the attacker has for an individual, the easier the re-identification will be. However, if the individual in B_I is not represented in A_I (*i.e.*, a background knowledge breadth issue), the information will probably render

irrelevant and misleading to the classifier.

- *Computational capabilities*: The processing power available to attackers influences their ability to make inferences. Greater computational capacity would enable them to use more complex models to build the classifier and/or more costly training strategies (see next section), potentially increasing the success rate of the attack.

In the next section, we discuss the choices that attackers can make to balance inference accuracy and computational cost. Moreover, the experiments we report in Section 3.3 are defined so that the influence of these resources on the TRIA's accuracy can be observed in practice.

3.2 Choosing and implementing the classifier model

The classifier model is a fundamental element of our approach because it determines the accuracy of the attack according to the resources available to execute it. Although the design of TRIA is general and can be enforced using any machine learning model capable of classifying documents, in this section, we present our specific design choices and implementation details adopted for the *build_classifier* and predict methods (lines 1 and 4 of Algorithm 1).

3.2.1 Selecting the language model

Our goal is to reproduce as realistically as possible the approach that an attacker may employ for re-identification. Under this premise, we expect attackers to take advantage of pre-trained language models, which have been introduced in Section 2.3, as state-of-the-art for NLP. The idea is to exploit the general understanding of text brought by language models devise a re-identification attack by additionally pre-training and fine-tuning those models with the background knowledge available.

TABLE 3.1: Comparison of widely-used BERT-based language models

Name	#parameters	Pre-training data
BERT (Devlin et al., 2019)	110M (`bert-base-uncased`) 340M (`bert-large-uncased`)	Wikipedia + BookCorpus (Zhu et al., 2015)
DistilBERT (Sanh et al., 2019)	66M (`distilbert-base-uncased`)	Same as BERT
RoBERTa (Liu et al., 2019)	125M (`roberta-base`) 355M (`roberta-large`)	Same as BERT + CCNews (Mackenzie et al., 2020) + Open WebText (Gokaslan and Cohen, 2019) + Stories (Trinh and Le, 2018)

On this basis, we consider BERT (Devlin et al., 2019) or its variations (Sanh et al., 2019; Liu et al., 2019; Jiao et al., 2019; Wang et al., n.d.; Rasmy et al., 2021) as reasonable language models candidates for the attack. Table 3.1 compares widely used language models grounded on BERT in terms of size and pre-training data, which we will later use in our empirical experiments. In particular, DistilBERT (Sanh et al., 2019) is a distilled version of the original BERT which reduces a 40% the model's size while keeping a 97% of its performance in multiple tasks, while RoBERTa (Liu et al., 2019) is a replication study of BERT pre-training, with a careful search of hyperparameters and trained with ten times more data (including also BERT's training set). Due to its larger size, RoBERTa tends to provide better results than BERT in multiple tasks. BERT-based models can obtain human-level results with neither a huge cost nor unfeasible knowledge assumptions from the attacker. They can also be easily run locally, thereby eliminating the need to outsource potentially sensitive documents to a remote server for processing. Furthermore, these models are well-established and publicly accessible, with numerous domain-specific fine-tuned versions available[2].

The following subsections explain our approach for using a BERT-based language model for re-identification. This approach can be applied to any BERT-based language model with negligible code modifications thanks to the HuggingFace's Transformers library (Wolf et al.,

[2] `https://huggingface.co/models?other=bert`

2020) we employ, which is the *de facto* standard library for the implementation, pre-training, fine-tuning and usage of language models.

3.2.2 From language model to classifier model

To adapt a BERT-based language model for a classification task, we apply the standard procedure described in (Sun et al., 2019) already implemented in the Transformers library[3]. It consists of appending a fully connected classification layer to the language model. This layer takes as input the output vector corresponding to a special classification token named [CLS] that is prepended to the input sequence. The classification layer then produces a logit vector with one value per class (which, in our case, corresponds to an individual from B_I). By applying the normalization SoftMax activation to the logits, the final probability for each class is obtained. As this new layer is designed to identify the individual from B_I related to the document, we will refer to it as the *individual classification layer*.

Transformer-based models such as BERT have a limited input length. In the case of BERT (and its variations) this limit is 512 tokens, which is shorter than the length of many real-world documents. To address this limitation, we divide each document into acceptable-length splits and process each split using the BERT-based classifier. Details of the splitting criterion are provided in the next section. During training, the BERT-based classifier is fine-tuned to classify document splits individually (see Section 3.2.4). For prediction, we aggregate the classification output of each split to obtain the final prediction. More details on this are provided in Section 3.2.5.

3.2.3 From documents to model input

For each input document, we perform the three subsequent steps:

[3] `https://huggingface.co/docs/transformers/v4.40.1/en/model_doc/bert#transformers.BertForSequenceClassification`

- *Pre-processing*: This step is done to unify semantically equivalent text spans and remove non-meaningful ones. Concretely, we perform lemmatization (*e.g.*,"changing" replaced by "change") and remove special characters (*e.g.*, the "#" of a hashtag) and stop words (*e.g.*, "the" or "that"). This can be done with any NLP library such as spaCy[4]. On the one hand, lemmatization gets rid of morphological variations of words, making the classification less dependent on the text syntax and more focused on the ((quasi-)identifying) facts conveyed by the document. On the other hand, the removal of special characters and stop words brings three benefits. First, it discards non-meaningful (and, therefore, non-identifying) information that may confuse the model. Second, since the classifiers' input length is limited, removing those elements increases the density of quasi-identifiers at each split/input, which should be beneficial for re-identification. Finally, it reduces the number of splits (explained at the sliding window step) generated per document that, in turn, decreases the training and prediction runtimes.

- *Tokenization*: Transforms the sequence of words of the input text into a sequence of tokens, as required for a language model. For BERT-based models, the standard WordPiece subword segmentation algorithm (which BERT relies upon (Devlin et al., 2019)) is employed by leveraging the implementation available in the Transformers' library[5].

- *Sliding window*: To accommodate BERT-based models' input length limitations (512 tokens), tokenized documents sometimes need to be divided into smaller splits. Inspired by (Pappagari et al., n.d.), we split the documents by using a sliding window with a length of N tokens (being N smaller or equal than the model's limit),

[4] `https://spacy.io/`

[5] `https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertTokenizerFast`

and an overlap of M tokens with the previous window (equivalent to a stride of N-M). The overlap enables the model to never lose the context of consecutive tokens, that is, it ensures that the model can observe the previous and following M tokens for every token (except the document's first and last) in at least one split. Setting an appropriate N and M is crucial, since they determine how long the detected long-term dependencies can be (which affects the model's accuracy), and the number of text splits generated (which influences training and prediction runtimes, which depend on the number of inputs rather than their length). These N and M hyperparameters can be empirically adjusted through hyperparameters search, as we detail in Section 3.2.6.

3.2.4 Building the classifier

By leveraging the steps outlined in the previous subsections, we now detail the implementation of the *build_classifier* method from Algorithm 1. As shown in the algorithm, it receives as input the background knowledge documents (B_D), the corresponding individuals (B_I) and the mapping function (F_B) that connects both. The output of the method is a classifier trained for document re-identification. In Figure 3.2, a diagram of the workflow for this method is depicted, which we explain below.

As introduced in Section 2.3, adapting language models to specific tasks often involves additional pre-training and/or fine-tuning steps on task data. Following (Sun et al., 2019), we apply both steps to the language model for obtaining a corpus-specific BERT classifier adapted to the domain of the documents to re-identify. For both steps, the procedures defined in Section 3.2.3 are applied to all documents identically, with the exception of sliding window splitting.

In the additional pre-training step, the model is further optimized on a masked language modeling (MLM) objective. This is the standard

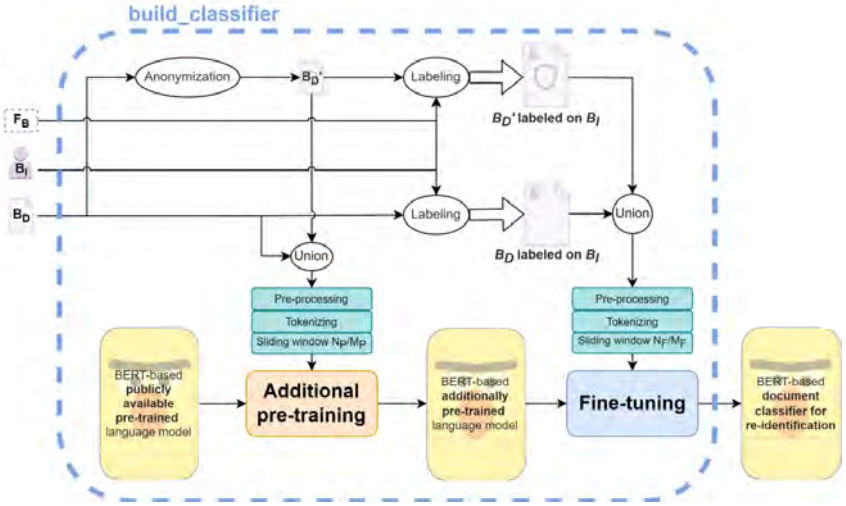


FIGURE 3.2: Workflow for the build_classifier method from Algorithm 1.

pre-training task on which BERT was initially trained, and implementations[6] and guides[7] are provided in the Transformer's library. In MLM, a vocabulary classification layer is appended to the language model, and all the parameters of this extended model are optimized. At the end, the vocabulary classification layer is removed, and we preserve the additionally pre-trained language model. For this step, increasing the sliding window length and/or overlap (N_P and M_P in Figure 3.2) should be beneficial, so that the context available for each token is maximized.

For the fine-tuning step, the language model resulting from the additional pre-training step is further optimized on labeled data[8], which in our case corresponds to the identity of the individuals referred in the background documents. More specifically, an individual classification

[6] `https://huggingface.co/docs/transformers/v4.40.1/en/model_doc/bert#transformers.BertForMaskedLM`

[7] `https://huggingface.co/docs/transformers/tasks/masked_language_modeling`

[8] `https://huggingface.co/docs/transformers/es/training`

layer is appended to the language model (see Section 3.2.2), and the resulting classifier model is trained with a cross-entropy loss[9] to predict the individual being referred to in each document split. For this step, increasing the length of the sliding window (N_F in Figure 3.2) may not be beneficial: while a longer window provides more information to assist the classification, it can also be too specific to the document. Moreover, for a fixed overlap (M_F in Figure 3.2) ratio, the longer the window, the fewer splits are created for each document, thus reducing the variability of the samples per individual. As a result, a too large window may lead to overfitting, and make the model focus on document-specific (sub)sentences rather than on more general (quasi-)identifying words or text spans. The same sliding window configuration selected for fine-tuning will be used at prediction (see Section 3.2.5). Regarding the data employed for this two-step training, it should be stressed that the attackers' knowledge is limited to B_D, B_I and F_B. Documents in B_D provide knowledge of the individuals' specific vocabulary, which improves understanding of domain-specific words. Furthermore, B_D can be labeled on B_I by using F_B, thereby providing useful information on the relationship between publicly available information and the individuals' identity. This can lead to the detection of (quasi-)identifying attributes (*e.g.*, the person's name, or demographic attributes), which form the basis of the re-identification attack.

An intuitive approach would be to perform additional pre-training using B_D and fine-tuning with B_D labeled on B_I. This results in a corpus-specific model capable of mapping documents to B_I, as is needed for the attack. However, it is important to note that the goal of the attack is to correctly classify documents from A_D, which stem from a different data distribution than B_D documents. In particular, B_D are clear texts (*e.g.*, identified publications in social networks) whereas A_D are anonymized documents (*e.g.*, non-identified medical reports with quasi-identifying

[9] `https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html`

terms masked via suppression or generalization). This could impair the re-identification accuracy, since machine learning algorithms are sensitive to differences between training and test data distributions.

Our solution is to create $B_D{}'$, an anonymized version of B_D, by using any practical text anonymizer that attackers may have at their disposal. Using the same method employed for A_D would be the ideal but, since such method would be rarely known, a standard NER-based approach (given that NER is the most used and directly applicable technique for text anonymization) can be used instead. In our experiments, we utilized spaCy NER for this purpose, which is one of the evaluated anonymization techniques (refer to Section 3.3.1). Because the resulting document set $B_D{}'$ is more similar to A_D, it will provide a better approximation of how data are (typically) anonymized. Moreover, $B_D{}'$ can be labeled on B_I (given that $B_D{}' \rightarrow B_D$ is known), which facilitates discovering the identities associated to the masked documents; for example, by discovering identifying words neglected by the anonymization method (*e.g.*, a specific location) that also appear in A_D documents. Under these considerations, we propose using B_D and $B_D{}'$ for additional pre-training, and the union of B_D and $B_D{}'$ labeled on B_I for fine-tuning, thereby obtaining a classifier model better adapted to the expected anonymized documents.

3.2.5 Prediction

The classifier resulting from the *build_classifer* method is employed in the *predict* method of Algorithm 1 to infer the individual from B_I corresponding a (presumably anonymized) document. This final classification step is depicted in Figure 3.3. First, the document is pre-processed, tokenized and split by using the same sliding window configuration employed for fine-tuning (see Section 3.2.4). Afterwards, each split is processed by the trained BERT-based classifier (one at a time) and the

results are aggregated into a single global prediction using the element-wise sum of all the logits vectors. The distribution of class (in our case equivalent to individuals in B_I) probabilities is obtained by applying a SoftMax function to those results. Finally, the class/individual with the maximum probability is chosen as the predicted individual for the document.

We favored the sum of logits instead of the most frequent prediction or the sum of the classes' probabilities (*i.e.*, logits after SoftMax) to account for the magnitude of the output. If the model is able to clearly identify an individual based on one text split, this prominent output would be reflected in the sum of logits. The motivation for this design is to mimic human reasoning for re-identification, which gives greater importance to parts of the text with clear re-identification clues (such as (part of) the person's name left unmasked), than to those with less clear evidences. This is also in line with (Li, Li, and Venkatasubramanian, 2006; Mozes and Kleinberg, 2021), which point out that missing a single identifying attribute during anonymization may be enough to reveal an individual's identity.

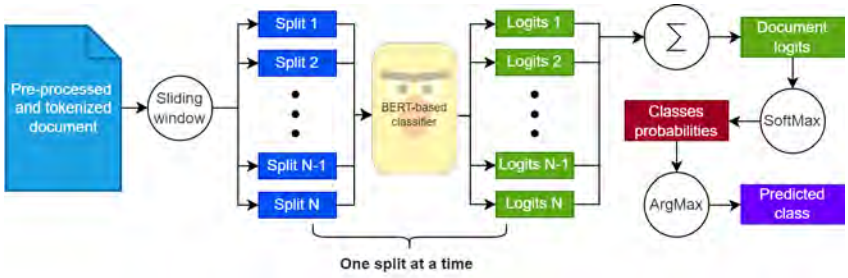


FIGURE 3.3: Our proposed workflow for the predict method.

3.2.6 Selection of hyperparameters

The classifier construction relies upon the definition of several hyperparameters common to neural networks training (*i.e.*, *number of epochs*, *learning rate* and *batch size*) and TRIA-specific (*i.e.*, *sliding window length* and *sliding window overlap*, see Section 3.2.3). To this end, we propose to empirically set this hyperparameters based on a separate development set, as it is standard in machine learning. That is, by defining a development set disjoint but similar to the evaluation set and performing multiple tries with different hyperparameter values in order to select the combination that provided the best results.

In our case, the evaluation set to be mimicked by the development set are the documents in A_D labeled in B_I (*i.e.*, re-identification ground truth) and a try is the complete re-identification attack presented in Algorithm 1. We propose to define a fixed number of epochs for additional pre-training, and fine-tune it until a pre-defined maximum number of epochs is achieved or the development accuracy stops improving (early stopping). As development set, we propose to create a random subset from the documents in B_D, which we call C_D, and transform it to be as similar as possible as the data distribution of A_D. For this, a straightforward approach would be to anonymize C_D; nevertheless, this would result into identical texts to those in ${B_D}'$, which are already present in training data. Thereupon, a previous step is required aiming to differentiate C_D texts from B_D ones and, if possible, to assimilate them to those in A_D prior anonymization. We propose to apply a summarization-like procedure on documents from C_D, which would result in a set that we call $\acute{C}_D$. To this end, abstractive or hybrid summarization methods are preferred to extractive ones (El-Kassas et al., 2021), so the resulting summarizations do not include sentences identical to those present in B_D. After summarization, documents in $\acute{C}_D$ are anonymized (producing ${\acute{C}_D}'$) as done for ${B_D}'$. The resulting ${\acute{C}_D}'$ is used as the development

set. The specific hyperparameters employed in our experiments are detailed in Section 3.3.3.

3.3 Evaluation results

In the following we report empirical results on using TRIA and TRIR to evaluate several text anonymization methods on a common corpus of text documents. To offer a broad comparison, we assessed the residual re-identification risk of NLP-based, PPDP-oriented and manual anonymizations. Our experiments are configured to provide a comprehensive analysis of the influence of the resources available to attacker resources in the risk assessment. We also report a comparison against the standard recall metric based on ground truth human annotations. To ensure reproducibility, all the code and data are available on: `https://github.com/BenetManzanaresSalor/TextRe-Identification`

3.3.1 Evaluated methods

As stated in Section 2.1, automatic methods usually approach anonymization as a NER task (Meystre et al., 2010; Aberdeen et al., 2010; Chen et al., 2019; Dernoncourt et al., 2017; Elazar and Goldberg, 2018; Hassan, Domingo-Ferrer, and Soria-Comas, 2018; Huang et al., 2020; Johnson, Bulgarelli, and Pollard, 2020; Liu et al., 2017; Mamede, Baptista, and Dias, 2016; Neamatullah et al., 2008; Reddy and Knight, 2016; Szarvas, Farkas, and Busa-Fekete, 2007; Xu et al., 2019; Yang and Garibaldi, 2015; Yogarajan, Mayo, and Pfahringer, 2018; Bubeck et al., 2023), in which words belonging to allegedly re-identifying categories (*e.g.*, locations, names, dates, etc.) are masked. In this case, masking consists of replacing the detected entities by their semantic categories. We considered the following NER-based tools, which have been in employed the literature to anonymize texts (Hassan, Sánchez, and Domingo-Ferrer, 2023; Lison et al., 2021; Pilán et al., 2022):

- *Stanford NER* (Manning et al., 2014): 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); and *NER7*, which detects ORGANIZATION, MONEY, DATE, PERCENT, PERSON and TIME.

- Presidio[10]: an anonymization-specific tool based on NER. Among the diversity of entity types supported by Presidio, we enabled those corresponding to quasi-identifying information: PERSON, LOCATION, NRP –person's nationality, religious or political group– and DATE_TIME.

- spaCy NER[11]: we employed the en_core_web_lg[12], neural model trained on the Ontonotes v5 corpus (Weischedel et al., 2011), which is capable of detecting named entities of DATE, CARDINAL, EVENT (*e.g.*, wars), GPE (*e.g.*, countries, cities), FAC (*e.g.*, buildings), LANGUAGE, LAW (named documents made into laws), MONEY, LOC (non-GPE locations such as mountain ranges), NORP (nationalities or religious or political group), ORDINAL, PERCENT, ORG, PERSON, PRODUCT (*e.g.*, vehicles), TIME (times smaller than a day), QUANTITY and WORK_OF_ART (*e.g.*, books titles, songs) types.

For PPDP-grounded text anonymization methods, as discussed in Section 2.1, the only practical method we found is Hassan, Sánchez, and Domingo-Ferrer, 2023. This method employs word embeddings (e.g., Word2Vec (Mikolov et al., 2013)) to detect text spans semantically linked to the individual requiring protection. Hereafter, we will refer to this method as *Word2Vec*, as it is built upon this model. The model is trained in an unsupervised manner from scratch by using the documents to be

[10] `https://github.com/microsoft/presidio`
[11] `https://spacy.io/api/entityrecognizer`
[12] `https://spacy.io/models/en#en_core_web_lg`

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 an user-defined threshold t are considered (quasi-)identifiers and are subsequently masked. Unlike NER-based anonymization –which operates in a non-configurable manner– this tunable threshold allows Word2Vec to balance privacy and utility, making it unique among the evaluated methods. In our experiments, we employed two threshold values: $t = 0.25$, which was observed by the method's creators to maximize privacy protection, and a more relaxed $t = 0.5$, which would result in less masking and, therefore, weaker protection. For this method, we leveraged its implementation available in GitHub[13]. 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 addition to the methods listed above, we also considered the manual anonymization effort conducted in (Hassan, Sánchez, and Domingo-Ferrer, 2023), which was based on sound privacy-oriented annotation guidelines. This enables us to evaluate the robustness of manual anonymization under TRIA, but also makes it possible to compute the recall of the above-described methods –by considering the manual anonymization as the ground truth–, and compare this metric against the risk assessment of TRIR.

Finally, we also report the TRIR for the non-anonymized versions of the documents in A_D. This defines the baseline risk that anonymization methods should (substantially) reduce.

[13] `https://github.com/fadiabdulf/automatic_text_anonymization`

3.3.2 Evaluation corpus and background documents

We employed as evaluation data the corpus of Wikipedia biographies from actors and actresses presented in (Hassan, Sánchez, and Domingo-Ferrer, 2023) and described in Section 2.4 under the name *WikiActors*. To simulate the scenario depicted in Figure 3.1, we considered the article abstracts as the texts to be anonymized, and the article bodies, as the publicly available background knowledge. A subset of 50 abstracts corresponding to 50 contemporary, popular and English speaking actors was selected from this corpus (as done in (Hassan, Sánchez, and Domingo-Ferrer, 2023)). We considered them the set of documents to be anonymized. In attack's formal notation, the 50 extracted actors constitute A_I, the 50 abstracts anonymized with a specific method m define ${A_D}^m$, and the article bodies in the corpus define B_D with a population of B_I actors. As introduced in Section 3.1.2, TRIA (and subsequently TRIR) depends, among other attacker's resources, on the *background knowledge breadth*; that is, the overlap between B_I (the complete set of known individuals in the background knowledge) and A_I (the set of individuals referred to in the anonymized documents, which are unknown to the attacker), and the relative size and similarity of those two sets. To evaluate the background knowledge breadth, we defined several scenarios by setting increasingly larger B_Ds and corresponding B_Is with varying distributions:

- *50_eval*: this is a worst-case scenario for privacy, in which B_I exactly matches A_I. Subsequently, B_D comprises the 50 article bodies from Wikipedia corresponding to the 50 anonymized abstracts from those same articles. This scenario constitutes the easiest re-identification setting because there are no individuals in B_I absent from A_I and, therefore, defines an ideal attacker with respect to background breadth.
- *500_random*: a scenario consisting of 500 article bodies randomly extracted from the full corpus plus those corresponding to the 50

actors in A_I that were not included in the random selection (thereby ensuring that $A_I \subseteq B_I$). Compared to the *50_eval* scenario, B_I is larger, with most of its individuals absent from A_I. Nevertheless, the similarity between the randomly selected individuals not present in A_I and the popular English-speaking actors of A_I is expected to be low, thereby facilitating re-identification. This scenario defines a more feasible but still powerful attack.

- *500_filtered*: a more realistic scenario with 581 article bodies obtained by filtering the full dataset of 18,672 articles according to several attack-oriented criteria. The criteria aim at maximizing the number of individuals in A_I present in B_I (even without knowing A_I, as it would happen in practice). In particular, the filtering discarded article bodies corresponding to non-native English speakers, dead individuals, non-actors (*e.g.*, directors or producers), actors born before 1950 or after 1995 (latter included) and those whose article included less than 100 links and was available in less than 40 languages. The latter two criteria aim to capture the "popularity" of the actor. As a result, only 40 out of the 50 actors in A_I appeared in B_I, thereby limiting re-identification accuracy to 80%. Higher intra-similarity than for the *500_random* scenario is expected due to the tailored (*i.e.*, non-random) filtering. This defines a more realistic scenario, because the attacker knows that A_I individuals are popular, English-speaking and alive actors. The attack is considered a powerful yet credible threat.

- *2000_filtered*: a more realistic scenario with 1,952 article bodies selected by using the same filtering as in *500_filtered* but omitting the criterion related to the number of languages per article. As a result, 41 out of the 50 actors in A_I appeared in B_I, thereby limiting re-identification accuracy to 82%. This scenario is a superset of *500_filtered*, with 1,371 new individuals added due to the less strict filtering, most of them not present in A_I. The intra-similarity is

expected to be slightly lower than between the 581 individuals of *500_filtered*, but still significantly higher than for *500_random*. This defines a weaker but probably more feasible attack, with a more realistic size for the population in the background knowledge.

As we latter report, for larger cardinalities of B_D the training time of the model is likely to take longer than 12 hours with the (reasonably powerful) hardware configuration we employed. Specifically, ten fine-tuning epochs using the whole 18,672 articles takes approximately 21 hours. In such cases in which the number of background documents available for the attacker outweigh its computational capabilities to train and conduct an exhaustive attack (and provided that the attacker does not know A_I), the strategy employed to define the *500_filtered* and *2000_filtered* scenarios would be the most reasonable choice. This reaffirms the fact that these two latter scenarios are more realistic than *50_eval* and *500_random*, which define worse-case (from a privacy perspective) but more ideal (from the perspective of the attacker) scenarios. After defining B_D for each scenario, the corresponding $B_D{}'$, C_D, $\c{C}_D$ and $\c{C}_D{}'$ sets needed to build the training and development sets should be created as detailed in Section 3.2.6. For $B_D{}'$, the documents in B_D are anonymized using spaCy NER. After that, $\c{C}_D$ is defined as a subset of the abstracts corresponding to the bodies in B_D. Since the abstracts are summaries of the article bodies, this procedure can be considered equivalent to the summarization-based approach proposed in Section 3.2.6, which means that C_D does not need to be explicitly built. Finally, the same method employed for creating $B_D{}'$ is applied to the documents in $\c{C}_D$. This results in the $\c{C}_D{}'$ set that constitutes the development set. The size of $\c{C}_D{}'$ was set to 30% for *50_eval*, and 10% for the *2000_filtered*, *500_filtered* and *500_random* scenarios.

3.3.3 Testing environment and hyperparameters

To simulate the implementation of TRIA by potential attackers, we selected Google Colaboratory[14] as execution environment (see details in Section 2.4).

As discussed in Section 3.2.1, a BERT-based classifier was employed to conduct the attack. We considered the popular pre-trained language models based on BERT outlined in Table 3.1. For the experiments varying the background knowledge, we opted for DistilBERT[15] (Sanh et al., 2019), because it offers an outstanding trade-off between accuracy and computational cost.

The remainder of our implementation was performed as described in Sections 3.2.3, 3.2.4 and 3.2.5. Hyperparameters search was performed by using the re-identification accuracy at the development set as selection criteria, as outlined in Section 3.2.6. Ideally, this search should be performed independently for each B_D defined in the previous subsection. Nevertheless, given the number of tests to be conducted, their runtime, and the similarities of the scenarios, we only applied it to the *50_eval* scenario and then employed the obtained parameters for all the evaluated scenarios. Concretely, the hyperparameters that obtained the best accuracy at the development set were: *learning rate=5e-5* (for both additional pre-training and fine-tuning), *batch size for additional pre-training=8*, *batch size for fine-tuning=16*, *sliding window length/overlap for additional pre-training=512/128* and *sliding window length/overlap for fine-tuning=100/25*. Training relied on an Adam optimizer with betas set to 0.9 and 0.999. The additional pre-training was performed for 3 epochs and the fine-tuning for a maximum of 20 epochs. Using early stopping with a patience of 5 epochs and the accuracy on the development set as criteria, fine-tuning was run for $\sim$20 epochs for the *50_eval*, *500_random* and *500_filtered* scenarios, and for $\sim$10 epochs for the *2000_filtered* scenario.

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

[15] We specifically relied upon the `distilbert-base-uncased` pre-trained model available on HuggingFace (`https://huggingface.co/docs/transformers`).

3.3.4 Influence of attacker's resources

Hereafter we perform a comprehensive analysis of TRIA and TRIR according to the resources assumed to be available to attackers. According to the characterization of such resources introduced in Section 3.1.2, we considered:

- *Background knowledge breadth*: By using the different background sets presented in Section 3.3.2, we assessed the role of the size and type of the population in the background knowledge available to attackers.
- *Background knowledge depth*: We considered two different sizes for the background documents (*i.e.*, article bodies): full bodies, which were used in most experiments, and the first 25% of the body contents, which were employed to analyze the influence of the background knowledge depth.
- *Computational capabilities*: We evaluated how the complexity of the attacker's inference affected TRIA's accuracy and runtime. This involved using different document splitting strategies (*e.g.*, sliding windows of different sizes), usage of anonymized B_D (*i.e.*, $B_D{}'$), and varying the size of the base language model employed to build the classifier (*i.e.*, DistilBERT, BERT or RoBERTa).

By configuring these aspects, we can set ideal (*i.e.*, most powerful) or more feasible (but less powerful) attack scenarios.

Background knowledge breadth Results depicted in Figure 3.4 show the influence of the different background knowledge breadths (corresponding to the B_Ds described in Section 3.3.2) in the re-identification of the anonymized documents (A_D) resulting from the methods introduced in Section 3.3.1. Notice that the "average" column in this and the forthcoming figures has been computed without considering the results

for $A_D{}^{\text{Clear text}}$. Results show that increasing background knowledge breadths significantly reduce the attack's accuracy, which is in line with the discussion in Section 3.1.2. The only exception is the *500_random* scenario that, despite having a B_I set much larger than A_I, resulted in just slightly less re-identification risk than *50_eval*. This was caused by the low similarity between the randomly selected individuals and the 50 protected ones, which made the latter easily distinguishable within the random set. In comparison, the risk of the filtered B_Ds (*500_filtered* and *2000_filtered*) was significantly lower because *i)* not all the protected subjects were present in B_D, and *ii)* those present were closer to the other individuals in B_D, thereby being more difficult to differentiate.

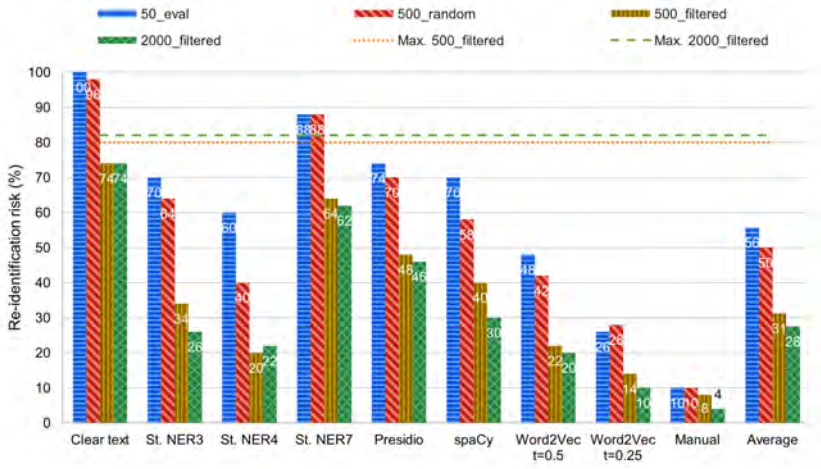


FIGURE 3.4: TRIR using different background knowledge breadths for several anonymization approaches. Horizontal lines depict the upper bounds for *500_filtered* and *2000_filtered*.

It is worth noting that the TRIR of $A_D{}^{\text{Clear text}}$ (*i.e.*, non-anonymized documents) is very close to the maximum risk of each scenario, being 100% for *50_eval*, 98% for *500_random* and 74% for *500_filtered* and

2000_filtered. This demonstrates the capability of the additionally pre-trained and fine-tuned DistilBERT model as classifier for the re-identification task. For anonymized documents, TRIA was able to re-identify individuals even from ${A_D}^{\text{Manual}}$, with accuracies greater than the random guess, which were 2% for *50_eval*, 0.2% for *500_random*, 0.17% for *500_filtered* and 0.05% for *2000_filtered*. The fact that ${A_D}^{\text{Manual}}$ did not achieve perfect protection (*i.e.*, its re-identification accuracy was above random guess), highlights the susceptibility of human annotations to omissions and errors, as discussed in Sections 1.1 and 2.2. This is caused by the difficulty of accounting for all the background knowledge and, therefore, all the (quasi-)identifying information (and their combinations) to be considered during manual anonymization.

Results for methods based on NER show weak protection, with re-identification risks greater than 45% for the *50_eval* worst-case scenario and, still, greater than or equal to 20% for *2000_filtered*. Results for ${A_D}^{\text{NER7}}$ are the worst in all cases, probably due to the lack of a LOCATION and MISC types, which encompass a large variety of quasi-identifying information. In contrast, the Word2Vec instantiations achieved the lowest risks among all automated methods and across all B_Ds. As expected, ${A_D}^{\text{Word2Vec t=0.25}}$ produced a better protection than ${A_D}^{\text{Word2Vec t=0.5}}$ thanks to its stricter threshold. The risks figures of ${A_D}^{\text{Word2Vec t=0.25}}$ are also quite similar to those obtained by the manual anonymization. The outstanding protection of this method is the result of not limiting masking to a pre-defined set of categories (as NER-based methods do).

The runtime of TRIA mainly depends on the classifier fine-tuning step, which in turn is influenced by the number of samples. On this basis, 20 epochs of fine-tuning on *50_eval* took 30' while for *500_random* it required 1h 40'. For *500_filtered*, the same 20 epochs took 4h 10', 2.5 times more than for *500_random*. This is in line with the length of the documents, which is 3 times larger for *500_filtered* than for *500_random*, because the popularity-related filters applied resulted in longer articles

(*i.e.*, popular actors have longer articles). Lastly, 10 epochs with *2000_filtered* required 5h.

Background knowledge depth To analyze the influence of the background knowledge depth available to attackers, we created a reduced version of the B_Ds sets presented in Section 3.3.2 encompassing the first 25% of the content in the corresponding Wikipedia article bodies. Figure 3.5 depicts the results for all the anonymization methods with the depth-reduced background knowledge sets.

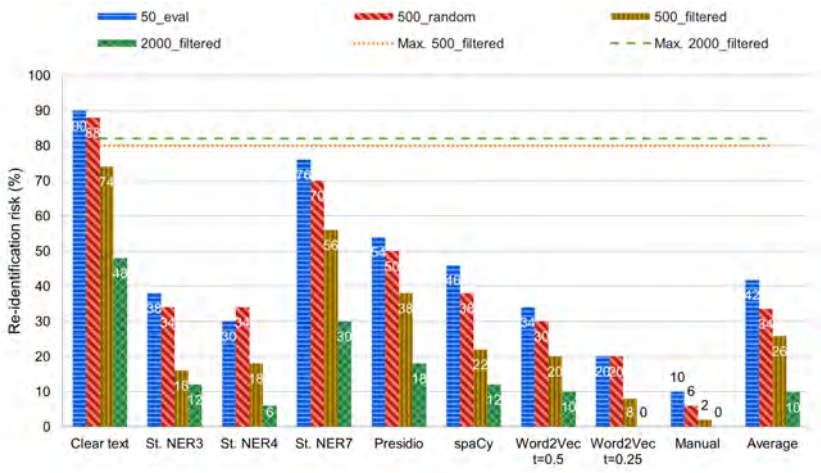


FIGURE 3.5: TRIR using the first 25% of the content from the background knowledge texts.

Comparing these results with those in Figure 3.4 we observe significantly lower risk figures, with values near zero for the manual and Word2Vec approach in some configurations. Differences were more pronounced for the largest background knowledge set (*2000_filtered*), which indicates that, for larger background populations (breadth), the amount of background knowledge (depth) available is crucial to differentiate individuals and enable accurate re-identifications.

As the overall risk has decreased, the gap between the different automatic anonymization methods also narrowed. Nonetheless, their relative rankings remain largely unchanged. An exception to this trend is seen for ${A_D}^{\text{NER3}}$ and ${A_D}^{\text{NER4}}$, which obtained risks comparable with those of ${A_D}^{\text{Word2Vec t=0.5}}$. This suggests that the quasi-identifiers missed by those NER-based methods (and exploited by TRIA) were mostly located beyond the first 25% of content of the background documents.

Computational capabilities The most computational-intensive component of TRIA is building the classifier (see Section 3.2.4), particularly, during the fine-tuning step when the model is trained for re-identification. We next report experiments to evaluate how the different settings that can be tuned by attackers to improve training (and, therefore, improve inference quality at the cost of computation time) affect TRIA's accuracy and runtime. To maximize the observable differences to changes to those settings, all tests were conducted on the *50_eval* worst-case scenario, which is the least affected by the size of B_I and its divergence with A_I.

Figure 3.6 illustrates the effect of the splitting technique (detailed in Section 3.2.3) during the fine-tuning that, as stated in Section 3.2.5, is the same used for prediction. Sliding-window N/M defines a window of length N and an overlap M with the previous window. On the other hand, sentence-based stands for one split per sentence, an alternative splitting technique that produced splits with an average length of 23.5 tokens.

As shown on Figure 3.6, there is a general trend whereby smaller splits sizes produce more accurate results. This is in line with the hypothesis made in Section 3.2.4, which associated longer window lengths with overfitting. During the fine-tuning step, using long splits may result in samples too specific to the publicly available sources of the individual (*i.e.*, the article's body in B_D and its anonymized version in ${B_D}'$), while reducing the number of samples from which learn the class

(since, in general, the longer the split, the fewer the number of splits available). Subsequently, the model is likely to overfit to the training documents rather than learning generalizable (quasi-)identifiers that facilitate re-identification. This makes it preferable to use shorter splits, but these would increase fine-tuning runtime. In fact, fine-tuning time scales almost linearly to the number of splits and, therefore, runtimes are significantly larger for smaller split sizes (*i.e.*, 15,873 splits and 2h 30' for sentence-based, 4,217 splits and 30' for 100/25 sliding-window, and 2,088 and 15' for 200/50 sliding-window). For this reason, we used the sliding window 100/25 configuration in all the other tests as a well-balanced trade-off between fine-tuning time and accuracy.

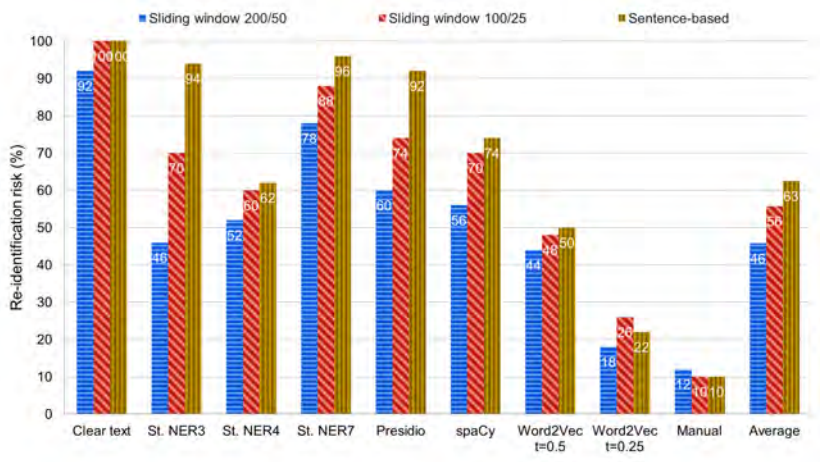


FIGURE 3.6: Comparison between different split sizes at fine-tuning for the *50_eval* scenario.

Figure 3.7 presents an ablation study in which the classifier is only trained with the raw bodies B_D or the anonymized bodies $B_D{}'$, instead of the combination of both sets (as we proposed in Section 3.2.4 3.2.4). These ablations almost halve the training data (and consequent runtime) of the additional pre-training and fine-tuning steps. Figure 3.7 shows

that, in most cases, important benefits are obtained by using $B_D{}'$ instead of B_D, and that their combination further improves the results. The improvements are considered worth the increase of training time (30' instead of 20'). This is in line with the arguments given in Section 3.2.4, probably with $B_D{}'$ helping the model to deal with anonymized documents, and B_D providing the information corresponding to the masking replacements. Specifically, documents in $B_D{}'$ allow the model to adapt to the data distribution of anonymized documents, which differs from that of B_D due to the masking replacements. Moreover, they allow the model to discover text spans neglected by the anonymization method that are also present in documents from A_D. When B_D and $B_D{}'$ are combined, the model can learn useful information on how documents have been anonymized, thereby facilitating re-identification.

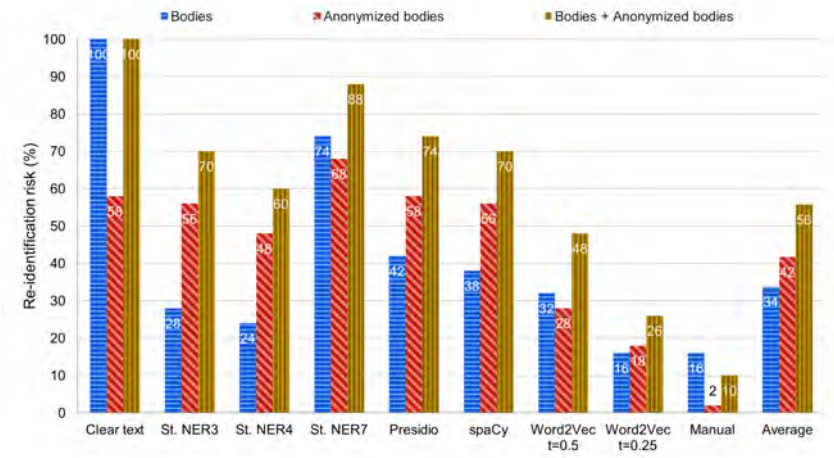


FIGURE 3.7: Ablation study on the inclusion of anonymized bodies in training for the *50_eval* scenario.

Finally, Figure 3.8 illustrates the effect of the neural language model employed for the re-identification. We compare the results obtained with DistilBERT (`distilbert-base-uncased`, with ∼67M parameters)

with those obtained by using the standard BERT model (`bert-base-uncased`, with ∼110M parameters), and RoBERTa (`roberta-base`, with ∼124M parameters). Notice that the combination of the narrowest knowledge breadth (*50_eval*), deepest background knowledge (complete article bodies) and largest model (RoBERTa) defines the most unconstrained and powerful attacker configuration considered so far. Fine-tuning runtimes in this experiment were 30′ for DistilBERT, 1h for BERT and 1h 15′ for RoBERTa. The accuracies averages show that, in comparison with DistilBERT, the improvement provided by BERT is small, whereas RoBERTa performed slightly worse for automated methods. The minor improvement of BERT is in line with the negligible loss of accuracy (∼3%) observed by the authors of DistilBERT in other tasks (Sanh et al., 2019). The relatively poor performance of RoBERTa is, on the other hand, surprising. One possible explanation might come from the training data employed for its pre-training. All BERT, DistilBERT and RoBERTa use the Wikipedia as training data, which therefore matches our evaluation dataset; but, whereas for BERT and DistilBERT this represents 75% of the training data, for RoBERTa it is just 7.5%. Exceptionally, the re-identification accuracy for ${A_D}^{\text{Manual}}$ seems to significantly benefit from the (usually) better-performing models (*i.e.*, RoBERTa > BERT > DistilBERT). This may be motivated by the stricter manual anonymization, which requires re-identification to focus on learning the few highly specific (quasi-)identifiers that may remain in the anonymized documents. In that case, re-identification depends more on the amount of knowledge available to build the model, rather than on Wikipedia-specific knowledge.

3.3.5 Comparison with recall and human annotations

As mentioned in Section 2.2, the goal of empirical risk assessment is to evaluate whether the level of protection achieved by the anonymized documents meets the privacy requirements of the data holder. The

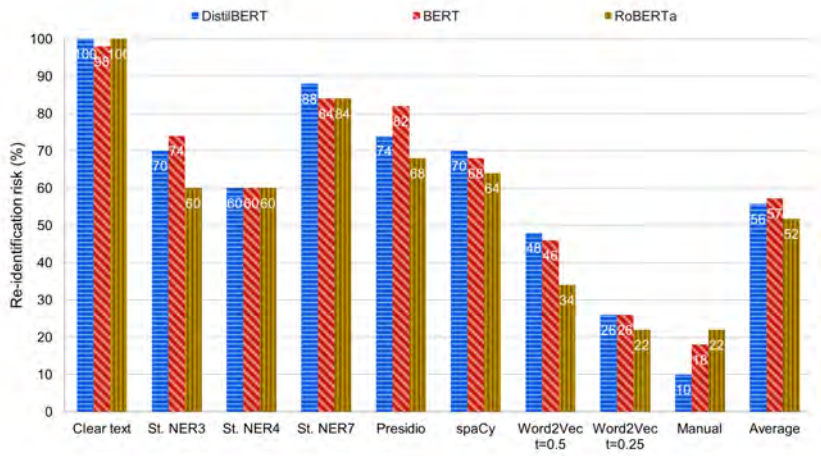


FIGURE 3.8: Effect of different language models on TRIR.

recall metric (Eq. 2.4) employed in the literature assesses risk by comparing automatic anonymization with manual annotations, which are considered the anonymization ground truth. Consequently, it evaluates a single (and presumably strict) privacy requirement encompassed by that manual ground truth. In contrast, TRIA (and, therefore, TRIR) can be configured for evaluating distinct privacy requirements according to the assumptions made on the attacker's resources. As shown in the previous section, by calibrating those resources, one may simulate ideal/powerful attackers that would define a worst-case scenario for privacy, or more feasible (and less powerful) attackers with more realistic access to data sources and computational capabilities.

In the following, we compare recall and TRIR under different attacker configurations with varying background knowledge breadths and depths that, as shown in the previous section, are the factors with the greatest influence in the risk assessment. To compute the recall for each

TABLE 3.2: Correlation between recall and TRIR with different attacker's resources for the different anonymization methods.

Background knowledge breadth	100% background knowledge depth	25% background knowledge depth
50_eval	0.871	0.648
500_random	0.820	0.664
500_filtered	0.699	0.573
2000_filtered	0.690	0.694

method, we used its standard implementation[16] and the manual annotations from (Pilán et al., 2022) as ground truth. Given that recall quantifies the inverse of the disclosure risk, we leveraged its complementary (*100% - Recall*) to enable a direct comparison with TRIR. The comparison was done in terms of the Pearson correlation between both metrics for the different anonymization methods considered. Results are reported in Table 3.2.

The correlation figures vary substantially, with the highest value corresponding to the most powerful attacker configuration (*50_eval* background breadth and 100% background knowledge depth). This is consistent with the manual annotations conducted in (Hassan, Sánchez, and Domingo-Ferrer, 2023), which were based on sound privacy-oriented annotation guidelines and which, as shown in the figures reported in the previous section, resulted in the strongest anonymization. Inversely, the lowest correlation corresponded to one of the weakest (but also more feasible) attacker configurations.

Figure 3.9 depicts a detailed comparison between recall and TRIR for the most and least correlated attacker configurations. We can see that, even though recall was strongly correlated with TRIR's most powerful attack scenario, it also systematically underestimated the observed re-identification risk. This is the result of recall relying on (imperfect) manual annotations, which exhibited a non-negligible re-identification

[16] `https://github.com/NorskRegnesentral/text-anonymization-benchmark/blob/master/evaluation.py`

risk (10%) under this attacker configuration. Differences with respect to the least correlated configuration were larger, which shows the inability of single annotations to encompass different privacy requirements or attack scenarios. These results showcase TRIR as an unsupervised alternative to recall, which is free from its dependencies (on human annotations) and that is more flexible to encompass varying privacy requirements.

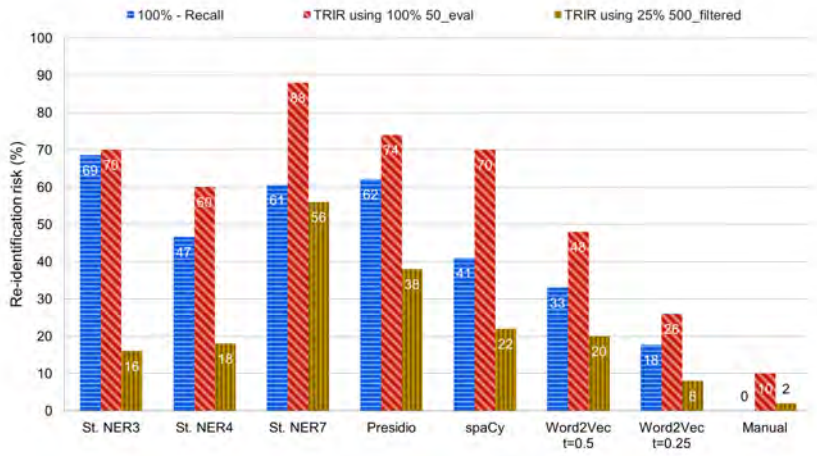


FIGURE 3.9: Comparison between the standard recall metric and the most and least correlated TRIR's attack configurations.

As a complement to the former results, and to better understand the limitations of manual annotations, we also measured the *precision* of the automatic anonymization methods. Precision computes the percentage of automatically masked text spans that were also annotated by humans, and it is used in the literature as an estimate of the utility preserved by anonymized documents. Low precision indicates that terms were unnecessarily masked, thus hampering the utility of anonymized documents. Figure 3.10 reports the precision values for the different methods using the same implementation as for recall.

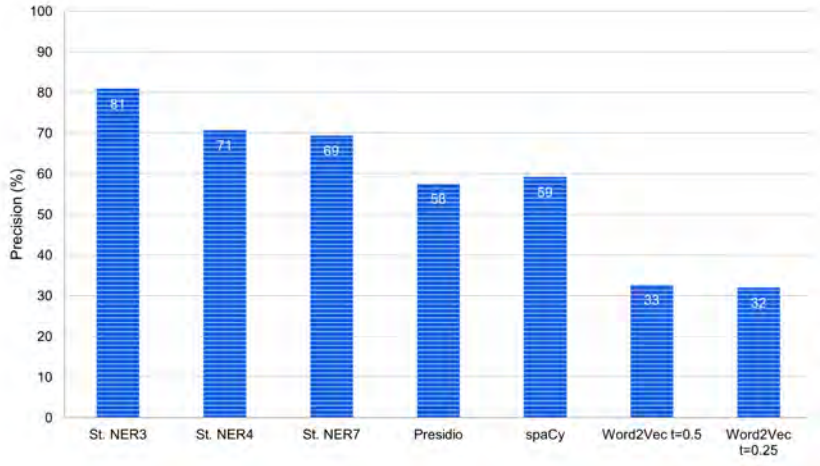


FIGURE 3.10: Precision of automatic anonymization methods.

Consistent with previous privacy assessments, there are notable differences between the NER-based approaches and the Word2Vec-based method. On the one hand, the higher precisions of NER-based methods derive from the fact that most of the (limited) NE types they are able to detect actually correspond to (quasi-)identifying information. The decrease in precision among these methods is also related to the diversity of named entity types each method encompasses, with a broader coverage resulting in a lower precision. Because this broader coverage not only did not provide better privacy, but resulted in the highest risk (see NER7, Presidio and spaCy in Figure 3.9), we can conclude that more extensive NER coverage does not necessarily lead to better anonymization.

On the other hand, the Word2Vec-based method resulted in the lowest precision. However, the fact that both Word2Vec instantiations obtained nearly identical precision but significantly different recall/TRIR, makes for an interesting analysis. First, this indicates that most of the

identifying text spans detected in ${A_D}^{\text{Word2Vec t=0.25}}$ but not in ${A_D}^{\text{Word2Vec t=0.5}}$ were also detected by humans in ${A_D}^{\text{Manual}}$, thereby increasing recall. Those are the text spans whose embeddings obtained a similarity in the range (0.25, 0.5] with the individual's name, probably corresponding to slightly disclosive quasi-identifiers. Subsequently, the 67% of text spans detected in ${A_D}^{\text{Word2Vec t=0.5}}$ that were not detected by humans in ${A_D}^{\text{Manual}}$, obtained similarities in the range (0.5, 1] that, because their proximity to the individual's identity, are expected to correspond to highly disclosive quasi-identifiers. If the Word2Vec criterion is minimally correct (which it certainly seems to be because of its strong protection), this would imply that the human annotators missed some disclosive text spans. This observation aligns with the non-negligible risk measured for the manual anonymization in most attack configurations, and underscores the humans' difficulties to detect non-obvious (quasi-)identifiers that can be leveraged by attackers to conduct re-identifications.

3.4 Conclusions

We have proposed two privacy evaluation tools for text anonymization: a *Text Re-Identification Attack* (TRIA) and an associated *Text Re-Identification Risk* (TRIR) metric. Compared to the recall-based risk assessment ubiquitously employed in the text anonymization literature, our approach provides an unsupervised alternative that does not rely on costly manual annotations, and that is flexible enough to encompass a wide vary of privacy requirements/attack configurations. Moreover, whereas recall is a completeness measure that has been used as a proxy of true risk assessment, our metric measures privacy in the same manner it would be threatened: performing re-identification attacks by exhaustively exploiting the available background knowledge. Empirical results highlighted the limitations of recall due to its dependency on (imperfect) human annotations. This is consistent with the limitations of the

use of absolute recall values as a measure of protection/residual risk highlighted in recent studies (Lison et al., 2021; Pilán et al., 2022; Mozes and Kleinberg, 2021). Moreover, we provided additional empirical evidence to the criticisms raised in (Lison et al., 2021; Hassan, Sánchez, and Domingo-Ferrer, 2023) on the drawbacks of using NER methods for text anonymization.

Chapter 4

Unsupervised utility evaluation of text anonymization methods via neural language models

The standard approach for assessing utility preservation in text anonymization consists in comparing protected documents with manual annotations using the IR-based precision metric. Nevertheless, as outlined in Section 2.2, this human-based approach presents several drawbacks. An effective utility evaluation should focus on measuring the actual information loss resulting from anonymization.

To overcome the limitations of precision-based utility evaluation, in this chapter we present TPI (*Text Preserved Information*), our 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 a 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 Chapter 3, thereby offering an complete framework for text anonymization evaluation. 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 rest of the chapter is organized as follows. In Section 4.1, we propose and formalize our metric. Section 4.2 reports and discusses empirical results obtained by applying our metric on a variety of anonymization methods, and compares it with the standard precision metric. Section 4.3 offers a comprehensive evaluation of all the considered anonymization methods, reporting their empirical privacy and comparing their privacy/utility balance. Finally, Section 4.4 gathers the conclusions of the chapter.

4.1 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 on manual annotations.

Given that semantics play the most critical role in determining text utility (Batet and Sánchez, 2018), 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 (Resnik, 1995). 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 (Resnik, 1995; Sánchez and Batet, 2016; Batet and Sánchez, 2020). In this chapter, 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.

Our utility metric can be seen as an adaptation of the methodology employed for utility evaluation in the SDC area (Hundepool et al., 2012), where documents correspond to records, TIC can be considered as an abstract characterization of the utility the data convey, and TPI computes the dissimilarity between TIC of original and anonymized documents.

4.1.1 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 repetitions of terms). Let P be the protected version of O. P also consists of an ordered list of M terms, $P = [k_1, \ldots, k_M]$. Note that the number

of terms in P might differ from those in O due to the term suppressions performed by the anonymization method. As introduced above, the information content (IC) is leveraged to measure the terms' semantics/utility (Resnik, 1995):

$$IC(t) = -log(prob(t)) \tag{4.1}$$

Details on the calculation of $prob(t)$ will be provided in Section 4.1.2.

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:

$$TIC(D) = \sum_{t \in D} IC(t) \tag{4.2}$$

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:

$$TPI(P, O) = \frac{TIC(P)}{TIC(O)} \tag{4.3}$$

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. 2.5) 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 is different from 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, and would give practitioners a clearer idea on the actual utility they can expect. The practical differences between TPI and precision will be empirically evaluated in Section 4.2.1.

4.1.2 Probability calculus

Gathering accurate term probabilities is crucial for TIC calculation and, subsequently, for the TPI (Eq. 4.3) 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 (Resnik, 1995). 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 (Sánchez, Batet, and Viejo, 2013a; Sánchez and Batet, 2016). 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) (Seco, Veale, and Hayes, 2004; Batet and Sánchez, 2020). 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 dependency of 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), with 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 depend 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 the advances in neural language models to obtain contextualized term probabilities. As mentioned in Section 2.3, 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. For instance, for the input sentence "I drink apple `[MASK]`" the model learns to predict the word "juice" that corresponds to the `[MASK]` tag. The model outputs the probability for each vocabulary token to be placed at the `[MASK]` position that, in this example, is expected to be the maximum for the token "juice".

Taking this into account, we propose to replace each term for which probabilities should be calculated by the `[MASK]` tag, input this masked text into the MLM model and use the model's prediction probabilities for these terms as their occurrence likelihood. A high probability value would indicate that the token is predictable from the document context and, therefore, that it has a low information content. Conversely, a low probability value indicates that the token can hardly be inferred from the remainder of the document, and that it therefore has a high information content. 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 (Devlin et al., 2019), which are freely available at (Wolf et al., 2020).

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 from 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 sub-parts. For example, if the WordPiece tokenization (Wu et al., 2016) 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 pre-processing, 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 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 (Anandan and Clifton, 2011)).

Secondly, replacing all the terms to be assessed by the `[MASK]` tag

at once is not suitable, because the model will not have 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 2, a detailed description of the procedure is depicted. On the other hand, in N-TA, one term of each N is masked, keeping the remainder (including surrounding terms and terms from other sentences) as context. Algorithm 3 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.

Algorithm 2: Sentence-Term Alternating (S-TA)

Require: *terms*

termsProbs ← *Vector*[*terms.length*]
for (*sentence*, *startIndex*) in *sentenize*(*terms*) **do**
  for *i* in [1, . . . , *sentence.length*] **do**
    maskedSentence ← *copy*(*sentence*)
    maskedSentence[*i*] ← "[*MASK*]"
    probs ← *MLM*(*maskedSentence*)
    termsProbs[*startIndex* + *i*] ← *probs*[*i*].*forTerm*(*terms*[*i*])
  end for
end for
return *termsProbs*

Algorithm 3: N-Term Alternating

Require: *terms*, N

```
termsProbs ← Vector[terms.length]
for i in [1, ..., N] do
    maskedTerms ← copy(terms)
    maskedIndices ← ∅
    for j in [i, ..., terms.length] do
    /* Terms masking */
        if j mod N == 0 then
            maskedTerms[j] ← "[MASK]"
            maskedIndices.insert(j)
        end if
    end for
    probs ← MLM(maskedTerms)
    /* Get probabilities */
    for k in maskedIndices do
        termsProbs[k] ← probs[k].forTerm(terms[k])
    end for
end forreturn termsProbs
```

Intuitively, S-TA is expected to provide more reliable probabilities than N-TA with a small N, because it keeps a 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, N-TA 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 example, 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 as sentence terms, while

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

The behavior of both strategies is depicted in Figure 4.1 in a sample text.

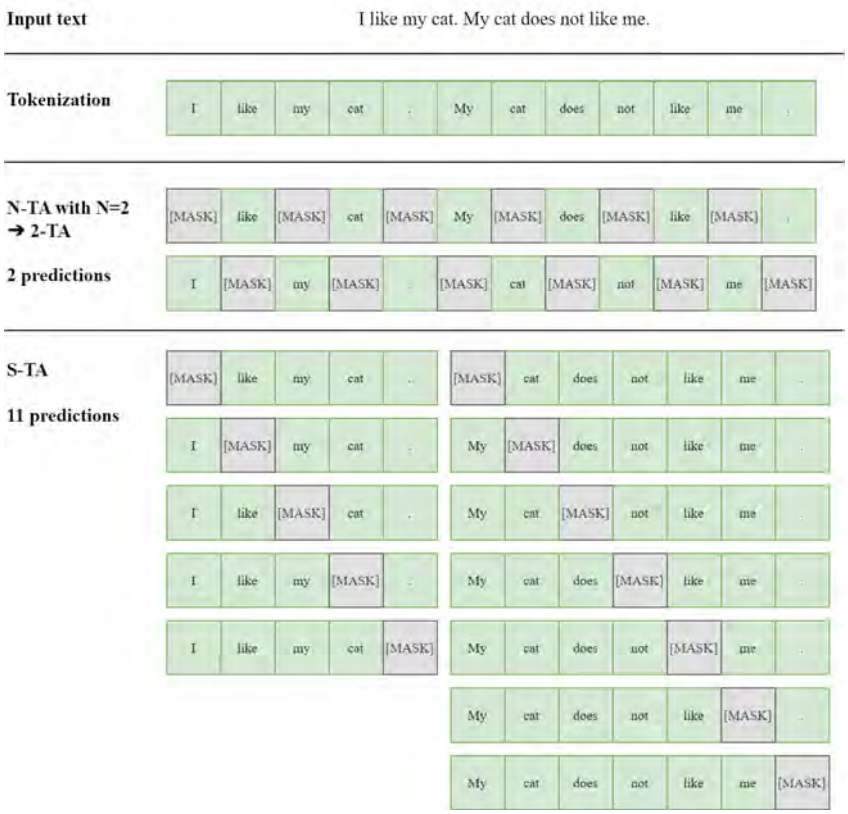


FIGURE 4.1: Comparison between N-TA and S-TA on an sample sentence

Using either of these two strategies, the probability of all terms in the original document O can be obtained. Nonetheless, 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 example, because a language model trained in clear text does not expect that 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, 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 in obtaining suitable generalizations that preserve utility to replace sensitive terms (Hassan, Sánchez, and Domingo-Ferrer, 2023). Because of 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 term probabilities computed for O by simply subtracting the IC of the protected terms from $TIC(O)$ to get $TIC(P)$. Equation 4.4 formalizes this computation, where $IsProtected(t)$ is a binary function that returns 1 if the term is protected and 0 otherwise.

$$TIC(P) = TIC(O) - \sum_{t \in P} IsProtected(t) \cdot IC(t) \tag{4.4}$$

By incorporating this calculus into Eq. 4.3, we obtain the following detailed expression for TPI:

$$TPI(P, O) = 1 - \frac{\sum_{t \in P} IsProtected(t) \cdot IC(t)}{TIC(O)} \tag{4.5}$$

4.2 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-focused dataset introduced in (Papadopoulou et al., 2022) and described in Section 2.4 under the name *Wiki553*. It is constituted by 553 randomly selected Wikipedia biographies' abstracts, with individuals ranging from football players to World War II generals.

4.2.1 Anonymization methods

The methods evaluated in the following experiments extend those assessed in Chapter 3 (see Section 3.3.1 for details), forming a superset with the inclusion of an additional PPDP-oriented approach.

For NER-based anonymization, this implies *Stanford NER* (Manning et al., 2014), *spaCy NER*[1], and *Microsoft Presidio*[2].

Regarding PPDP-oriented methods, we again include the practical approach presented in (Hassan, Sánchez, and Domingo-Ferrer, 2023), referred as *Word2Vec* due to its reliance on this embedding model. Concretely, in our experiments we trained the Word2Vec model using the 553 abstracts to be protected and an external corpus of 19,000 Wikipedia biographies of movie actors that was employed in the experiments of the

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

original paper (Hassan, Sánchez, and Domingo-Ferrer, 2023). The similarity threshold t was empirically set to 0.5 to balance the privacy/utility trade-off.

As a new addition, we include the PPDP-oriented approach presented in (Papadopoulou et al., 2022). This method exploits a knowledge graph that captures the distribution of terms in a large corpus (*i.e.*, Wikidata[3]) to detect terms or text spans that appear in fewer than k documents. According to this, the method enforces k-anonymity (Samarati, 2001) on the set of documents to be protected by incrementally suppressing terms until no combination of terms is present in less than K documents. Masking can be carried out in a *greedy* manner, where 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'. Note that effective protection using this PPDP-oriented approach requires a knowledge graph that not only covers the domain of the documents but also the (quasi-)identifiers of the individuals to protect. Developing such a resource beyond the comprehensive and well-structured Wikipedia setting poses significant difficulties and remains an open challenge for this particular method.

In addition to the automatic methods listed above, we also considered the manual anonymization conducted in (Papadopoulou et al., 2022) (see Section 2.4 for details). Beyond evaluating the utility derived from these manual annotations, we utilized them as ground truth to compute the precision and recall metrics for the aforementioned methods.

[3] `https://www.wikidata.org/`

4.2.2 Results

We next report the results of the application of TPI to the above-described methods. The calculation of both original and anonymized documents' TIC is implemented as described in Section 4.1 (Eq. 4.5), including the chunking pre-processing (using spaCy's noun chunks) and the S-TA and N-TA 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 by using the BERT language model (Devlin et al., 2019) with its standard pre-training for MLM on Wikipedia and the BookCorpus (Zhu et al., 2015). TPI calculation was performed on the cloud-powered IDE Google Colaboratory[4]. All experiments were executed in a stable environment consisting of an Intel Xeon CPU, 12GB of RAM and an Nvidia Tesla T4 GPU 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 4.2 reports TPI values for the different anonymization methods and compares them with the standard precision metric. Error bars correspond to the standard deviation for each metric and method across all the documents. The *average* column represents the mean of all automatic methods (*i.e.*, it does not consider manual anonymization).

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 to assess the preserved utility without requiring from costly human annotations. The fact that precision values are deterministically higher than those of TPI is consistent with the claim made in Section 4.1.1: whereas the upper bound for precision represents a perfect agreement between automatic and manual anonymizations, for TPI

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

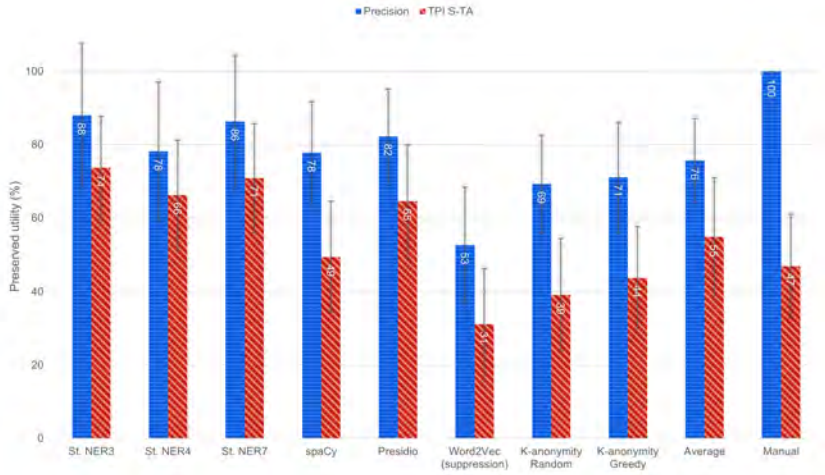


FIGURE 4.2: TPI and precision for different anonymization methods

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 specially 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 it 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 Word-2Vec approach, what 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. On the contrary, 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 more successful in preserving utility than the random technique.

4.2.3 Task-oriented utility evaluation

The following experiment tries to answer the following 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 was unsupervised (therefore, it was not dependent on task-tailored models or human criteria), and because all the documents' contents were leveraged during clustering. Specifically, we considered the clusters obtained from the original documents as 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) as similarity metric, which produces values in the [0..1] range.

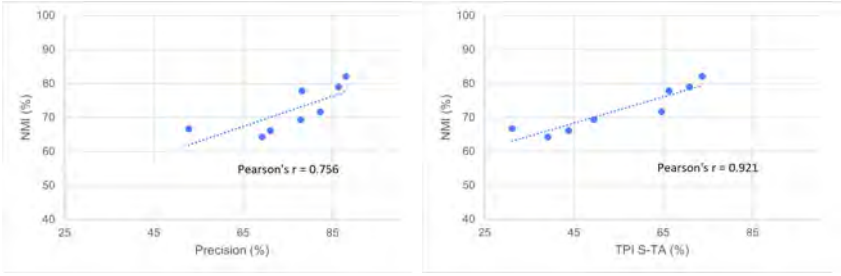
The clustering methodology that we used was inspired by the approach presented in (Subakti, Murfi, and Hariadi, 2022). First, a BERT-based[5] feature extraction was performed, which produces embeddings for all the document's tokens. Averaging them, we can obtain the document embedding. In the case of anonymized texts, the embeddings of masked text spans (being many of them very general or even meaningless deletion masks) were excluded from the averaging. Afterwards, the document embeddings were clustered with the K-means++ algorithm (Arthur and Vassilvitskii, 2007) with $K = 4$. K was chosen based on the best inertia obtained when clustering the original documents. To minimize the effect of 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 4.3 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.

[5]For our implementation, we leverage the `bert-base-cased` model available at `https://huggingface.co/google-bert/bert-base-cased`



(A) (B)

FIGURE 4.3: Correlation between the NMI cluster similarity for the different automatic anonymization methods and (A) Precision and (B) TPI S-TA

4.2.4 Comparing utility calculation strategies

The next experiments are designed to evaluate the influence of the context masking strategies discussed in Section 4.1.2. This is equivalent to measuring the effect of changing the context used by the language model to predict terms, which directly influences the TIC and TPI calculation.

Figure 4.4 compares the TPI values with the S-TA masking strategy (sentence-level context) against those obtained by 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 the MLM pre-training of BERT (Devlin et al., 2019). 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 the sentence level (that is, the usually most useful context), and masking the sentence terms one by one.

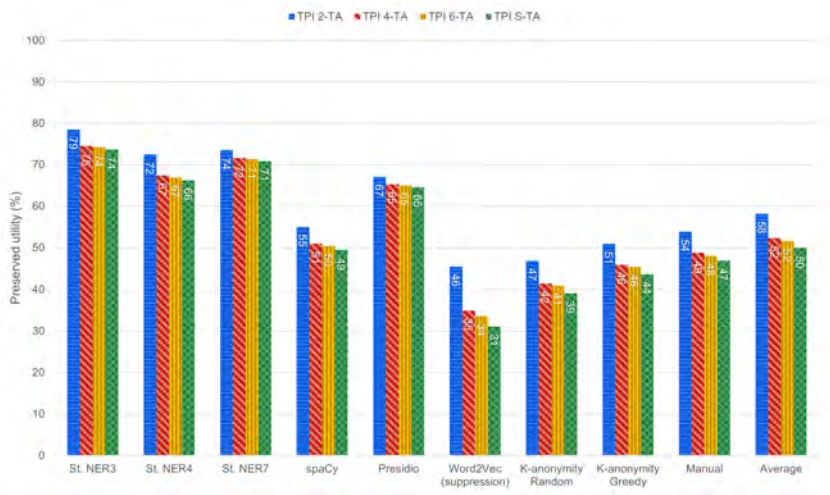


FIGURE 4.4: TPI values calculated with different context masking strategies

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 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 4.1, we report the runtime of the TIC calculation for the different context masking strategies. Notice that the runtime with S-TA is around 10 times greater than with 6-TA. Nevertheless, Figure 4.4 shows that 6-TA provides very similar results to S-TA. This is in line with the reasoning made in Section 4.1.2, which postulates N-TA as a lower-cost technique that approximates S-TA's results.

TABLE 4.1: TIC calculation runtime with different context masking strategies

Context masking strategy	Runtime
2-TA	39″
4-TA	61″
6-TA	65″
S-TA	10′ 21″

4.3 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 Chapter 3, in order to offer a comprehensive evaluation of the anonymization methods introduced in the previous section. TRIR and TPI 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, recall and F-scores.

4.3.1 Privacy assessment

As introduced in Chapter 3, 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 on 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 Chapter 3, 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 (Papadopoulou et al., 2022), we scraped the Wikipedia. We observed that around 8% of the biographies lacked a body section, consisting only of an abstract without 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 4.5 we report, in percentage, the TRIR-based privacy assessment of the different methods measured as the inverse of the TRIR-based re-identification risk, that is, $1 - TRIR$. For reference, we also report the recall –which is standardly used to measure privacy– computed w.r.t. the manual anonymization. Consequently, the recall for the original unmasked documents is 0%, whereas for the manually anonymized ones is 100%. Note also that, while 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 4.1.1 when comparing precision and TPI.

Focusing on TRIR, as expected, the lowest privacy level (*i.e.*, highest TRIR) corresponds to the original documents. Nonetheless, it is quite above the minimum theoretical privacy 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

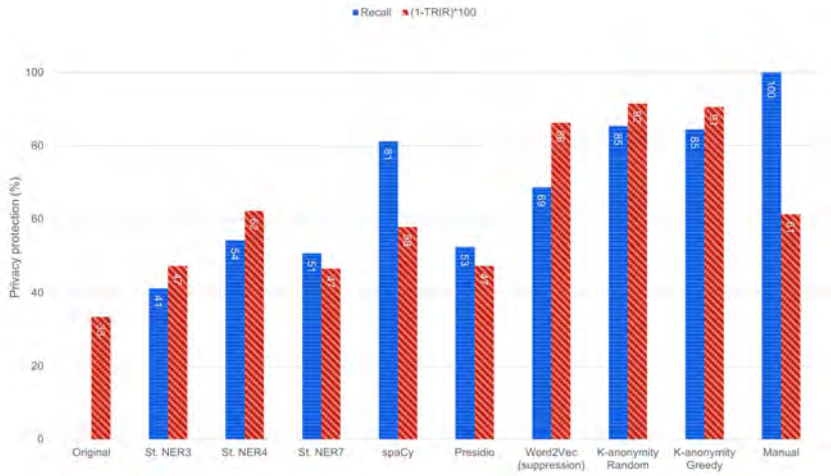


FIGURE 4.5: Privacy protection attained for the different anonymization methods according to recall and TRIR

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 (Lison et al., 2021), which argues 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, the Stanford's 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 *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 (Hundepool et al., 2012). 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.

Surprisingly, according to TRIR, the manual anonymization attained a level of privacy protection as low as the best NER-based method. This illustrates the shortcomings of manual annotations in an even more exacerbated way than the 10% TRIR observed in Section 3.3.4 of Chapter 3. The relatively low privacy observed in manual anonymization can be (at least partially) attributed to two non-exclusive causes. First, it is probably the result of the manual annotations being too much based on the typically privacy-lacking NER-based pre-annotations (Papadopoulou et al., 2022). 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, albeit not as noticeable, can be seen for TPI vs. precision in Figure 4.2. 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. Additionally, 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 over PPDP-oriented ones. The second possible cause is missed combinations of quasi-identifiers. While human annotators can proficiently identify and mask direct identifiers, such as names or addresses, they struggle recognizing the co-occurrence of multiple quasi-identifiers across documents. For example, if a document reveals that the individual to be protected is a male, plays football, comes from Spain, and has blond hair, an annotator will probably mask the term "Spain" based on the awareness that it is the sole individual from that country in the dataset. However, the annotator will more easily overlook the fact that this individual is also the only male footballer with blond hair in the dataset, especially

if this information is scattered throughout the document. This oversight is significant, because it is precisely those unique combinations of quasi-identifiers that are exploited by re-identification attacks. Moreover, according to the annotation guidelines in (Papadopoulou et al., 2022) and (Pilán et al., 2022), manual anonymization only aimed to preclude strict re-identification or, in other words, annotators never aimed for a re-identification probability larger than 50% (*i.e.*, $k = 2$ in terms of k-anonymity). All in all, this illustrates the limitations of manual anonymization efforts –which are prone to errors and omissions (Lison et al., 2021)–, and defines a relatively privacy-lacking ground truth. This not only compromises precision and recall as reliable evaluation metrics, but also results in an unfair assessment of more privacy-robust methods (Word2Vec and k-anonymity-based ones, in this case).

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) 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 employed parameters, the best TRIR was obtained by the k-anonymity random method, even though being less than a 1% better than its greedy counterpart. However, this comes at the expense of a lower utility, as reported in Figure 4.2, which therefore causes the greedy method to have a better privacy / utility trade-off. The Word2Vec-based approach is the closest, but with significantly worse utility (see Figure 4.2). The better performance of the k-anonymity-based methods can be explained by the high quality statistics provided by the knowledge

graph in which they rely. 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 a Wikipedia level. Nonetheless, 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 pre-processing nor tagging.

4.3.2 F-scores

As discussed in Section 2.2.2, 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.

F_1 and F_2-scores are computed as instantiations of F_β (Eq. 2.6). Notice that, as we have done in Section 4.3.1, privacy has been computed as $(1 - TRIR)$ in order to be properly aggregated with utility (TPI). Figures 4.6 and 4.7 report, respectively, the F_1 and F_2-scores resulting from our metrics and precision and recall.

We can see that the relative ranking of automatic methods obtained by sorting them by 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 automatic methods, as discussed in Section 4.2.2), divergences mainly come from the privacy evaluation (r = 0.8 for the results of automatic methods reported in Figure 4.5). In this respect, in Section 4.3.1 we observed that the manual annotations were privacy lacking, rendering the metrics relying on them as ground truth unreliable.

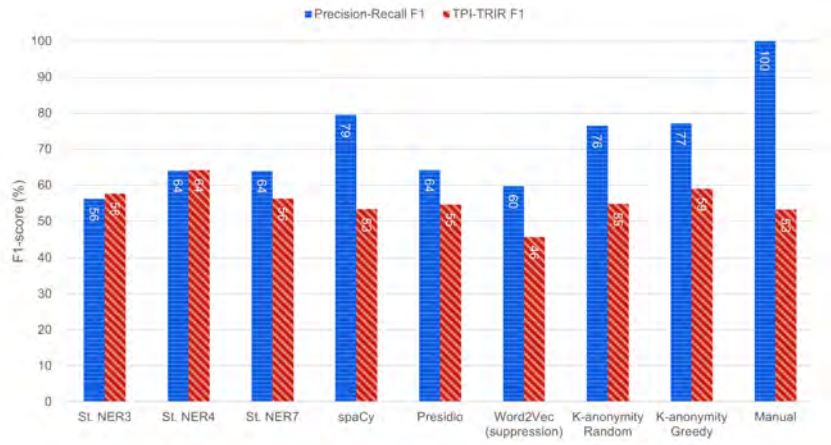


FIGURE 4.6: TPI and TRIR-based F_1-score vs. precision and recall-based F_1-score for the different anonymization methods

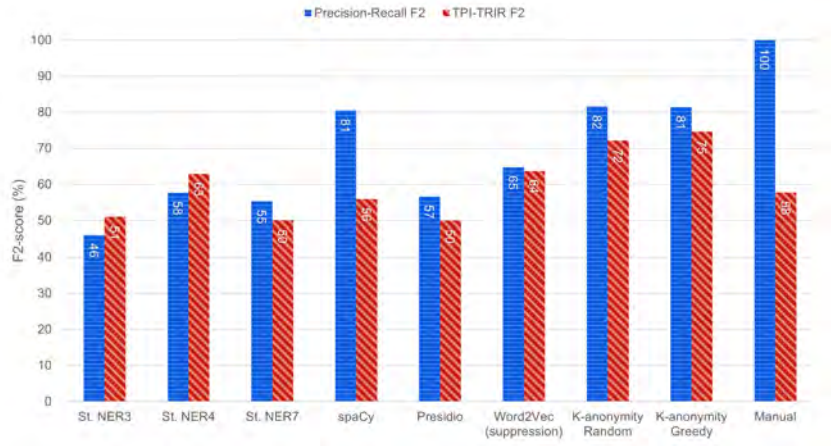


FIGURE 4.7: TPI and TRIR-based F_2-score vs. precision and recall-based F_2-score for the different anonymization methods

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 4.2.2), whereas TRIR's distribution is in line with recall's (as shown in Section 4.3.1), 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 matches with 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 (Ferrández et al., 2012; Mendels, 2020), should also be preferred for our metrics.

4.4 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 from costly and error-prone manual annotations. Experiments depict a strong correlation between TPI and precision, thereby 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 privacy metric presented in Chapter 3, TPI defines a complete evaluation framework that is free from the issues that

arise when employing precision, recall, and manual annotations to evaluate anonymization methods. The suitability of our metrics is illustrated through the comprehensive evaluation of a variety of text anonymization methods reported in Section 4.3. The reported results provide empirical evidence for the criticisms raised in (Lison et al., 2021; Hassan, Sánchez, and Domingo-Ferrer, 2023) on the drawbacks of using NER methods for text anonymization. Moreover, they also show that privacy-oriented methods (Hassan, Sánchez, and Domingo-Ferrer, 2023; Papadopoulou et al., 2022) produce more robust anonymizations, whereas manual anonymization may be hampered by inaccuracies and bias, and may introduce compromises when used as ground truth.

Chapter 5

Enhancing text anonymization via re-identification risk-based explainability

As discussed in Section 2.1.2, NER-based methods remain the predominant approach to text anonymization, offering some of the few practical and scalable solutions available. Nevertheless, these techniques exhibit notable privacy limitations, which undermine their ability to fully safeguard sensitive information.

Considering the limitations of NER for text anonymization, in this chapter we propose PETRE (*Privacy Enhancement for Text via Risk-oriented Explainability*), a privacy-grounded method that is able to improve NER-based anonymizations by enforcing a user-defined reduction of the empirical re-identification risk.

Our method brings the following contributions:

- Its risk assessment raises awareness to its users on the (usually

non-negligible) residual disclosure risk of well-established and widely used text anonymization mechanisms, specifically, NER-based ones.

- Even though it has been conceived for NER-based approaches, it is general enough to be applied to any starting method, regardless of its operating principle.

- On the contrary to NER-based methods, our approach empowers the user to intuitively configure the desired level of privacy by specifying the subjects to be protected and the maximum re-identification risk allowed for those subjects. Our method is grounded on probabilistic k-anonymity (Soria-Comas and Domingo-Ferrer, 2012) –a relaxation of the indistinguishably guarantee of the well-known k-anonymity privacy model (Samarati, 2001)–, which restricts the re-identification probability to be lower or equal than $1/k$.

- In contrast with the standard application of k-anonymity to structured datasets, in our method the aggregated background knowledge considered to assess disclosure risks can be made subject-specific. Therefore, anonymization can be fine-tuned to each individual to be protected, which reduces unnecessary masking.

- Anonymization is enforced in an iterative manner, by which the empirical re-identification risk is re-evaluated after each masking until reaching the user-specified privacy threshold. This also contributes to minimize unnecessary masking.

- Because the anonymization of each document is guided by the observed risk (rather than employing a predefined masking criterion such as NER-based methods), the anonymized outcomes will present an empirical re-identification risk lower or equal than $1/k$ regardless of their content and the starting anonymization method.

Thus, no *a posteriori* manual evaluations of the attained level of privacy would be required.

- In addition to defining the privacy requirements for anonymization, our method does not require human intervention and operates fully automatic and self-controlled.

To demonstrate the effectiveness of our proposal, we conducted several experiments on a general-purpose dataset of Wikipedia biographies (Papadopoulou et al., 2022) with a variety of starting anonymization methods encompassing not only NER-based approaches, but also more sophisticated privacy-grounded methods. We also show the generalizability of our approach in a domain-specific dataset of legal documents (Pilán et al., 2022). Results show that our approach consistently and proportionally reduces the re-identification risk of all the starting anonymizations, and competes against state-of-the-art methods while being free from their complexities, costs and external dependencies.

The contributions in this chapter have been published in (Manzanares-Salor and Sánchez, 2025).

The rest of this chapter is organized as follows. Section 5.1 presents and formalizes our approach. Section 5.2 details the implementation of our method. Experiments and results are reported in Section 5.3. Finally, Section 5.4 provides the conclusions.

5.1 Privacy Enhancement for Text via Risk-oriented Explainability

In this section we present PETRE (*Privacy Enhancement for Text via Risk-oriented Explainability*), a PPDP-oriented framework that takes the (privacy-lacking) anonymizations produced by NER and automatically and iteratively enhances them until reaching a user-defined privacy level, which is intuitively defined in terms of re-identification risk.

More specifically, our approach is able to enforce probabilistic k-anonymity (Soria-Comas and Domingo-Ferrer, 2012) on top of the starting NER-based anonymizations, thereby attaining an intuitive and user-configurable privacy guarantee. For this, the user will specify the desired value of k and the background knowledge for the individuals to be protected. Then, our method will iteratively enhance the starting anonymizations until the observed probability of re-identification of each document to be protected is lower or equal than $1/k$.

This probability is assessed through a re-identification model that leverages the stated background knowledge to infer the individuals referred in the documents to be anonymized. Our approach then uses explanations from this re-identification model -—specifically, how much each term in the document contributes to re-identification– to iteratively mask the most disclosive terms. This incremental process is repeated until each document meets the probabilistic k-anonymity requirement. By targeting only the terms that increase the re-identification risk, and by re-evaluating the residual risk at each masking iteration, our method avoids over-masking by design.

The architecture and workflow of our proposal are detailed in Section 5.1.1, while Section 5.1.2 formalizes it.

5.1.1 Architecture

In the following we provide a high-level overview of the PETRE framework by introducing its architecture, core components and workflow. These elements are depicted in Figure 5.1, with its core components highlighted in blue and bold text:

- **Starting anonymized documents**: The set of documents to be privacy-enhanced by PETRE. They are the result of the application of a basic anonymization mechanism on the original (non-protected) documents. Each document (*e.g.*, a medical record or a CV) is labeled with the name of the individual whose identity is to be

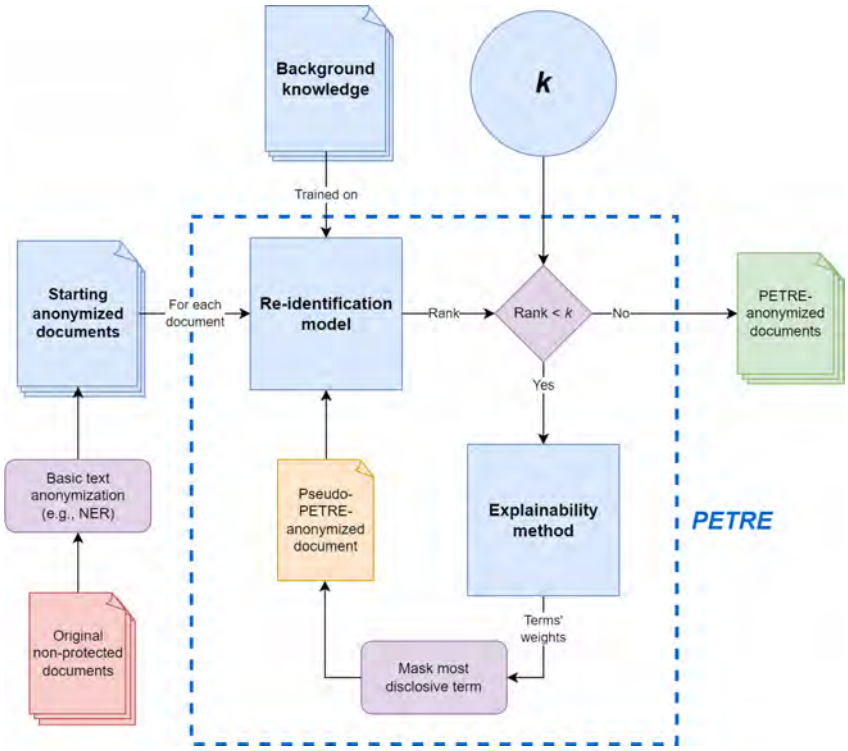


FIGURE 5.1: PETRE architecture and workflow

concealed. Our approach can be applied to any method that suppresses or generalizes text spans, but it is specifically designed to improve NER-based anonymizations. This is because it addresses their main weakness, that is, poor privacy due to their limited coverage of entity types. On the other hand, PETRE's performance will be constrained by starting anonymizations that might provide higher privacy protection at the cost of over-masking. This is because PETRE cannot recover unnecessary masking incurred by the starting anonymization method.

- **Background knowledge**: A collection of (assumed) publicly available identified documents containing information on the individuals to be protected. These documents are used for training a re-identification model (defined below), which is used to assess the practical re-identification risk of the documents to be protected. The higher the overlap between the individuals' information in this background knowledge and that present in starting anonymized documents, the higher the re-identification risk will be because more (quasi-)identifiers can be leveraged to conduct re-identification attacks. In turn, a higher re-identification risk will make anonymization stricter, as more (quasi-)identifiers will require to be masked. A straightforward way to define this background knowledge is to consider the original non-protected documents as background knowledge. This constitutes a worst-case scenario, which will result in the strictest privacy protection (at the cost of utility). For a better balanced privacy/utility trade-off, data holders may carefully define which background knowledge could be realistically available for each individual. In particular, the scarcer the background knowledge for an individual, the less masking (and subsequent information/utility loss) is required for achieving the desired level of privacy protection.

- **Re-identification model**: A classifier trained on the aforementioned background knowledge to predict the individuals' identity. This component evaluates the privacy level of a (protected) document. More specifically, for any given document, it predicts a probability distribution over the set of possible individuals to be re-identified. PETRE uses the produced rank –that is, the position of the corresponding individual's identity within the predicted probability distribution– to assess the re-identification risk: the lower the rank, the higher the probability of successful re-identification and, therefore, the higher the risk.

- k: A positive integer that defines the desired level of privacy in terms of maximum allowed risk of re-identification ($1/k$, as stated by probabilistic k-anonymity). PETRE uses it as a reference for the rank produced by the re-identification model for each document to be protected. In this way, PETRE can ascertain whether a document necessitates further masking ($rank < k$) or if it has already reached the desired level of privacy ($rank \geq k$). In other words, it defines the ending condition of the iterative anonymization. The desired level of protection, as well as the balance between privacy and utility, is determined by this parameter. Users should select k according to the characteristics of the dataset being anonymized, and the applicable regulatory requirements. First, the value of k should be proportional to the dataset's size. For instance, smaller datasets with only a few hundred samples, would not allow choosing a very large k without a significant loss of data utility. Second, the acceptable level of re-identification risk should correspond to the intended purpose of the data. While any $k \geq 2$ guarantees that no individual can be uniquely re-identified, as required by Recital 26 of the GDPR(European Commission, 2016)[1], lower re-identification risks are usually desired. For example, as highlighted in (El Emam, 2013), a re-identification risk of approximately 0.09% or less (equivalent to $k > 10$) is considered reasonable, as it has been explicitly proposed by two independent health regulators (European Medicines Agency, 2014; Health Canada, 2019).

- **Explainability method**: A module capable of estimating the influence of the document's terms in the probability predicted by the re-identification model for the individual it refers to. The output of this component are the terms' weights in the re-identification success. Weights' positive or negative magnitudes indicate their

[1] `https://gdpr-info.eu/recitals/no-26/`

influence (favorable or unfavorable, respectively) on the classification probability of the individual. PETRE uses these weights to mask the terms with the greatest influence in the re-identification risk at each iteration. In terms of explainability, this module obtains a local explanation for the probability predicted for a specific class/individual. Note that, for our application, the fidelity of the weights takes precedence over their interpretability, as our approach is automatic and does not involve human interpretation. Valid explainability methods can be model-agnostic or those model-specific that are compatible with the design of the re-identification model (*i.e.*, a classifier). Model-agnostic approaches grounded on additive feature attribution such as SHAP (Lundberg and Lee, 2017) or LIME (Ribeiro, Singh, and Guestrin, 2016) are adequate.

PETRE's design is general and allows each of its functional components (*i.e.*, the re-identification model and the explainability method) to be replaced, *e.g.*, to improve their performance as new developments arise. Together with the possibility of configuring the anonymization parameters (*i.e.*, k and background knowledge) for specific privacy/utility needs, this defines a more flexible and customizable approach than existing text anonymization methods while still keeping it practical.

Now that the core components have been elucidated, we can outline the workflow depicted in Figure 5.1. Each document from the **starting anonymized set** is processed individually by PETRE. First, privacy level of the document is assessed by the **re-identification model**, which has been previously trained with the **background knowledge**. If the rank predicted by the re-identification model is lower than the specified $\boldsymbol{k}$, the document will require further masking. In this case, the **explainability method** estimates the weights for all the terms in the document. Subsequently, the term with the largest weight is masked (*e.g.*, removed), thereby incrementally lowering the re-identification risk. Afterwards, the document is re-evaluated for re-identification risk and the process is

repeated until the document reaches or surpasses the specified k. This workflow is fully automatic and self-controlled, thereby requiring no human intervention.

5.1.2 Formalization

Let $A = \{(D_1, I_1), \ldots, (D_n, I_n)\}$ be the set of starting anonymized documents, consisting on n pairs of protected documents D and the identities of the corresponding individuals to be protected I.

Let $\Gamma(d)$ be the re-identification model, which receives a document d and outputs a probability distribution $P = (p_1, \ldots, p_n)$ summing one. Each p_i represents the probability that d actually refers to the individual I_i. This machine learning model is trained with a background knowledge set $B = \{(C_1, I_1), \ldots, (C_n, I_n)\}$, mapping the set of public identified documents C to their corresponding individuals in I.

For a given document-individual pair (D_i, I_i) from A, the document D_i is considered unequivocally re-identified if the probability p_i (corresponding to I_i) returned by $\Gamma(D_i)$ is the highest in the distribution P. This is in line with the well-known record linkage attack (Abril, Navarro-Arribas, and Torra, 2012), which data anonymization aims to prevent. Subsequently, we can assess the re-identification risk of a document D_i according to the ranked position of p_i in the distribution P. We denote $Rank(d, i, \Gamma)$ as the function that, given a document d, an individual i and the re-identification model Γ, returns that ranked position. For example, $Rank(d, i, \Gamma) = 1$ indicates that $\Gamma(d)$ returned the largest probability for the individual i. On this basis, the upper-bound for re-identification risk can be intuitively derived as follows:

$$Rank(d, i, \Gamma) = k \implies \textit{Re-identification risk}(d) \leq 1/k \tag{5.1}$$

This way of measuring risk is in line with the privacy guarantee of k-anonymity (see Eq. 2.1). To be more precise, a $rank = k$ conforms to

the principles of probabilistic k-anonymity (Soria-Comas and Domingo-Ferrer, 2012); that is, it limits the re-identification risk to $1/k$ without requiring the documents of the k individuals be indistinguishable from each other –as stated by standard k-anonymity–. It is important to note that, in text anonymization, indistinguishability is neither necessary nor advantageous, as it only hampers utility by requiring to resort in common generalizations across different documents.

Let $k \in \mathbb{N}^+$ be the privacy threshold, which defines the minimum rank acceptable for a PETRE-anonymized document. For instance, a value $k = 1$ means that no anonymization is necessary, while $k = 2$ requires that all documents have, at least, a $rank = 2$. If a document is assessed with a rank lower than the desired k, it will need further masking.

The explainability method $\xi(d, i, \Gamma)$ receives a document d, an individual I_i and the re-identification model Γ. The document d consists in a set of terms $(t_1, \ldots, t_m)$, with m varying for each d. The output of ξ are the weights($\in \mathbb{R}$) of each document term $(w_1, \ldots, w_m)$ for the probability p_i predicted for the individual I_i. Positive weights indicate a direct influence (the higher, the more the influence) of the corresponding terms towards re-identifying I_i; that is, terms with positive weights are effectively acting as (quasi-)identifiers for I_i. On the other hand, negative weights correspond to terms that hinder the re-identification, whereas weights close to zero are expected for non-meaningful terms (such as stop words) or terms non-related to the individual to be protected. We operate under the assumption that the explainability method's fidelity might be imperfect, thereby resulting in weights that may not precisely represent the impact of masking the corresponding terms. Due to this reason, weights are recalculated after each masking iteration and disclosure risk re-evaluation. In this way, we ensure that they end up being as faithful as possible, even if this means conducting more iterations.

Algorithm 4 formalizes our iterative privacy-enhancement method. First, the re-identification model Γ is trained by using the background

knowledge B (line 2). Then, for each document of the starting anonymized set A (line 3), its rank is measured (line 4) and, as long as it exceeds the privacy threshold k (line 5), the document is iteratively masked (lines 6 to 9). For each masking step, first, the explainability method is leveraged to obtain the terms influences' weights in re-identification (line 6). Afterwards, in line 7, the index of the term that contributed the most to re-identification (*i.e.*, that with the maximum weight) is obtained. That index is then employed in line 8 to mask the corresponding term. Masking only the most disclosive term at each step aims to minimize the amount of masked terms, which will contribute to retain data utility. At the end of the masking loop (line 9), the rank of the document is re-evaluated. Once the document is sufficiently anonymized, it is stored in A^{PETRE} (line 11), which is the PETRE-enhanced set of documents produced by the algorithm (line 13).

Algorithm 4: PETRE

Require: A, B, Γ, ξ, k
1: $A^{PETRE} \leftarrow \emptyset$
2: $\Gamma \leftarrow Train(\Gamma, B)$
3: **for** $(document, individual)$ in A **do**
4: $\quad rank \leftarrow Rank(document, individual, \Gamma)$
5: $\quad$ **while** $rank < k$ **do** /* Mask until privacy threshold is met */
6: $\quad\quad weights \leftarrow \xi(document, individual, \Gamma)$
7: $\quad\quad termIdx \leftarrow ArgMax(weights)$
8: $\quad\quad document \leftarrow Mask(document, termIdx)$
9: $\quad\quad rank \leftarrow Rank(document, individual, \Gamma)$
10: $\quad$ **end while**
11: $\quad A^{PETRE} \leftarrow A^{PETRE} \cup document$
12: **end for**
13: **return** A^{PETRE}

5.2 PETRE implementation

In this section, we provide a comprehensive account of the practical implementation of the general framework introduced above. Based on the architecture described in Section 5.1.1, we focus on the re-identification model and the explainability method. In the following, we determine and discuss the technologies used for both components. Algorithm 5 depicts the specific adaptations made to the general PETRE framework outlined in Algorithm 4 to accommodate the chosen solutions.

For the re-identification model, we considered the TRIA (Text Re-Identification Attack) method proposed in Chapter 3, which trains a language model-based classifier specifically for text re-identification. TRIA perfectly matches the Γ function defined in Section 5.1.2, as it can be trained on background knowledge set B and infer a probability distribution P over the set of possible individuals I for an input document d. Consequently, in Algorithm 5 the parameter Γ has been renamed as $TRIA$.

It is worth noting that TRIA's predictions are done per document splits, and those are later aggregated to obtain the document-level probability distribution (see Section 3.2.5 for details). For our implementation, we configured TRIA code to define one split per sentence, using default values for the other settings. The model then produces sentence-level logits, which are then summed and passed through a softmax function to compute the final probability distribution. As backbone pre-trained language model upon TRIA builds on, we leverage DistilBERT (Sanh et al., 2019), a smaller version of the well-known BERT model that offers similar accuracy, but any other general model could be employed.

On the other hand, we adopted SHAP (Lundberg and Lee, 2017)[2] as the explainability method. SHAP is a state-of-the-art model-agnostic approach that estimates the influence of each input feature in the prediction (local explanation) of a regressive model. In our framework, the

[2] `https://github.com/shap/shap`

re-identification model is a classifier, which in turn can be abstracted as a multi-regression method (*i.e.*, one regression per class). Subsequently, SHAP can be used to measure the weight (*i.e.*, influence) of each document's term (*i.e.*, input feature) in the re-identification probability (*i.e.*, regression value) of a specific individual (*i.e.*, class). This aligns with the individual-specific explainability function defined as ξ in Section 5.1.2 and employed in Algorithm 4. As a consequence, Algorithm 5 uses $SHAP$ in place of ξ. However, making SHAP usable for PETRE requires from some modifications.

Firstly, a SHAP explanation entails performing multiple model predictions by randomly replacing combinations of input features (*i.e.*, terms). The amount of combinations scales quadratically with the number of terms. This causes a scalability issue that, in conjunction with the potentially non-negligible cost of the re-identification model prediction, can lead to impractical execution times, especially when dealing with long documents. To address this problem, we leverage TRIA's particularity of processing documents per splits (prior to aggregation) and knowledge on SHAP foundations.

SHAP's need of evaluating features combinations comes from the assumption that the features influence may not be independent. That is, the influence of a feature may vary depending on other features of the same input. This is in line with the *compositionality* property of language, being each term's meaning (and subsequent influence in prediction) affected by its context. Nevertheless, TRIA processes documents by splits, so the terms from different splits cannot influence each other. This nullifies the necessity of evaluating their combinations. Therefore, SHAP can be applied to each split (rather than to the whole document prediction), thereby drastically reducing the number of term combinations to evaluate while not losing explanation quality. For instance, if having a document with m terms divided into s equally-lengthen splits, weights' computing cost is reduced from $O(m^2)$ to $O(s * (\frac{m}{s})^2)$.

Unfortunately, we observed that even this cost reduction may not be

enough for avoiding unpractical runtimes. To tackle this issue, we recall that PETRE only uses the weights to ascertain the most disclosive terms of the document (see line 7 of Algorithm 4). Because it is expected that these terms are contained in the most disclosive split, the explanation can be limited to this split (*i.e.*, lines 7 and 8 in Algorithm 5). This further reduces explanation cost from the aforementioned $O(s * (\frac{m}{s})^2)$ to $O((\frac{m}{s})^2)$.

Finally, it is relevant to establish a clear definition of what constitutes a term, particularly when it comes to the detection of (quasi-)identifiers. SHAP calculates its weights at the token-level[3], being tokens the input features for language models. However, in the context of text anonymization, the minimum meaningful units are typically words and/or noun chunks. This holds true for our framework as well, since attempting to mask individual tokens instead of complete words (*e.g.*, "ina" in "Angelina") or noun chunks (*e.g.*, "New" in "New York") would be ineffective. Hence, we designate both noun chunks and individual words (those not part of chunks) as terms. For noun chunks detection, we leverage the spaCy library[4]. Subsequently, we calculate terms' weights as the summation of the weights of the constituent tokens. In this way, masking becomes independent of the language model vocabulary, which is optimized for performance on NLP tasks rather than anonymization. Specifically, the vocabulary is based on tokens instead of terms, the latter being the relevant entities for anonymization.

5.3 Experiments

We next assess the performance of PETRE on multiple text anonymization mechanisms and datasets. All tests were conducted on the implementation described in Section 5.2 running on a computer equipped with a Nvidia 4080 GPU, an i7 10700 CPU, 32GB of RAM, and a 1TB

[3] `https://github.com/shap/shap/blob/master/notebooks/text_examples/translation/Machine%20Translation%20Explanations.ipynb`

[4] `https://spacy.io/usage/linguistic-features#noun-chunks`

Algorithm 5: PETRE with TRIA and SHAP

Require: $A, B, TRIA, SHAP, k$
$A^{PETRE} \leftarrow \emptyset$
$TRIA \leftarrow Train(TRIA, B)$
for $(document, individual)$ in A **do**
  $rank \leftarrow Rank(document, individual, TRIA)$
  while $rank < k$ **do** /* Mask until privacy requirement is met */
    $splits \leftarrow Split(document)$
    $mdSplit \leftarrow GetMostDisclosiveSplit(splits, TRIA)$
    $weights \leftarrow SHAP(mdSplit, individual, TRIA)$
    $termIdx \leftarrow ArgMax(weights)$
    $mdSplit \leftarrow Mask(mdSplit, termIdx)$
    $rank \leftarrow Rank(document, individual, TRIA)$
  end while
  $A^{PETRE} \leftarrow A^{PETRE} \cup document$
end for
return A^{PETRE}

NVME SSD. Code and data for all the reported experiments are available at `https://github.com/BenetManzanaresSalor/PETRE`.

5.3.1 Evaluation datasets

We leverage the general-purpose Wikipedia-based dataset presented in (Papadopoulou et al., 2022) and described in Section 2.4 under the name of *Wiki553*. It contains 553 abstracts of randomly selected Wikipedia biographies, with individuals ranging from football players to World War II generals.

We also assess the ability of our approach to generalize by applying it to a domain-specific dataset. For this purpose, we used the Text Anonymization Benchmark (TAB) (Pilán et al., 2022), derived from European Court of Human Rights (ECHR) case files (see Section 2.4 for details). Concretely, we employed the test subset of this dataset, consisting of 127 documents with corresponding annotations.

5.3.2 Anonymization methods

We tested the performance of PETRE on a variety of state-of-the-art automatic anonymization methods. These include not only NER tools (which would be the natural target for PETRE), but also PPDP-oriented approaches. The latter will show the capability of PETRE to provide consistent results regardless of the type and operating principles of the starting anonymization.

The evaluated NER-based approaches are *Stanford NER* (Manning et al., 2014) and *Microsoft Presidio*[5]. Detailed descriptions of these methods are provided in Section 3.3.1.

For PPDP-oriented methods, we considered the approach referred as *Word2Vec* (Hassan, Sánchez, and Domingo-Ferrer, 2023) (presented in Section 3.3.1) and the method based on *k-anonymity* and knowledge graphs (Papadopoulou et al., 2022) (described in Section 4.2.1 of Chapter 4). For the Word2Vec implementation, we trained the embedding model using both the target corpus of 553 abstracts requiring protection and an external corpus comprising 19,000 Wikipedia biographies of movie actors, consistent with the experimental setup of the original work (Hassan, Sánchez, and Domingo-Ferrer, 2023). Following the parameters of the original study, we employed the restrictive similarity threshold of $t = 0.25$.

It is worth noting that both PPDP-oriented methods, although being more practice-oriented than most related approaches in the literature (Pilán et al., 2022), would still be quite difficult to apply in a real-life scenario. On the one hand, Word2Vec requires training a word embedding model from scratch, which can incur in a substantial cost and requires the compilation of suitable external corpora to avoid overfitting. Moreover, because masking is based on an empirically-tuned similarity threshold, it does not offer a clear guarantee on how the documents have been protected. On the other hand, the method from (Papadopoulou

[5] `https://github.com/microsoft/presidio`

et al., 2022) requires building a knowledge graph, which, for most domains, would imply bespoke construction by humans. The associated effort of creating such a knowledge graph on demand could rival or even exceed that of manual anonymization.

Given the iterative and self-adaptive design of PETRE, intuition suggests that the privacy improvement it enables will be inversely proportional to the quality of the starting anonymization. That is, the more privacy-lacking the starting anonymization is, the larger the enhancement will be. Under this assumption, we expect the protection of the (privacy-lacking) NER-based methods to improve substantially with PETRE. For PPDP-oriented approaches, which are far more complex in their application but have shown to provide much more robust empirical privacy (see Section 3.3.4 in Chapter 3 and Section 4.3.1 in Chapter 4), we expect more subtle improvements; but we still decided to include them in order to test the generality of our approach.

Consistently with the masking strategies enforced by the methods above, the additional masking performed by PETRE consisted in *removing* sensitive terms from the documents.

5.3.3 Background knowledge

Because we use manual annotations as ground truth, PETRE's background knowledge should closely align with what human annotators had access to. For both datasets, annotators likely considered the own documents to protect as their background knowledge. Therefore, it is reasonable to set the whole sets of 553 (Wikipedia) and 127 (TAB test) documents as background knowledge for PETRE. This setting also corresponds to a worst-case scenario for privacy, where we assume perfectly knowledgeable attackers. This background knowledge was utilized to train the re-identification model employed in the experiments. This training process, which consists in further pre-train and fine-tune the DistilBERT language model, and is done once for the whole set of

documents to anonymize, took 104 minutes for the Wikipedia dataset and 121 minutes for the TAB test dataset.

5.3.4 Evaluation metrics

To gauge the improvement brought by PETRE over the baseline methods, we employed the standard evaluation methodology adopted in the text anonymization literature, which consists in comparing automatically anonymized outcomes to a manually anonymized ground truth (Johnson, Bulgarelli, and Pollard, 2020; Meystre et al., 2010; Yang and Garibaldi, 2015; Hassan, Domingo-Ferrer, and Soria-Comas, 2018; Neamatullah et al., 2008; Yogarajan, Mayo, and Pfahringer, 2018; Huang et al., 2020; Mamede, Baptista, and Dias, 2016; Sánchez and Batet, 2016; Sánchez and Batet, 2017; Hassan, Sánchez, and Domingo-Ferrer, 2023; Pilán et al., 2022; Papadopoulou et al., 2022). The agreement between both anonymizations is measured through the IR-based metrics *recall* (Eq. 2.4) and *precision* (Eq. 2.5), which correspond to privacy protection and utility preservation assessment. The overall anonymization quality is evaluated with *F-score* (Eq. 2.6), which is the weighted harmonic mean of precision and recall. F_1-score gives equal weight to both precision and recall, while F_2-score doubles the weight for recall. Further details of these metrics are available in Section 2.2.2 of Chapter 2.

In addition the standard measures above, we also evaluated the extend to which the anonymized documents retained their analytical utility in a downstream task. We focused on document clustering, which relies on the semantic intra- and inter-similarity among documents. Analytically useful anonymized documents mean that clusters created from those documents closely align with those derived from the original texts, ensuring that similar insights can still be extracted.

This evaluation follows the methodological framework established in Section 4.2.3 of Chapter 4. For the quantitative assessment of analytical utility preservation, we utilized the *Normalized Mutual Information*

(NMI) metric (Fred and Jain, 2003), which provides a robust measure of clustering similarity.

5.3.5 Setting the privacy parameter *k*

The privacy protection attained by PETRE directly depends on the value of the k parameter, which states the desired level of (probabilistic) k-anonymity. To enable a fair comparison against the human anonymizations employed as ground truth, in the following we try to ascertain which k value matches the best with the level of protection offered by those manual anonymizations.

For this, we measured the effective document rank provided by the re-identification model employed by PETRE for each manually anonymized document. As explained in Section 5.1.2, a $rank = k$ states an actual re-identification risk equal or lower than $1/k$ or, in other words, that the document is probabilistically k-anonymous. In particular, we measured the rate of documents that are compliant with 2-anonymity or above (that is, the percentage of documents that obtain a rank equal or higher than 2). As a reference, the percentage of original unprotected documents that were already 2-anonymous was just 1% for both datasets, which shows that the documents in those datasets are privacy challenging. Surprisingly, manual anonymizations got a 2-anonymous compliance rate around 19% for both datasets, which also means that 100%-19%=81% of the documents had a $rank = 1$ or, in other words, that a 81% of the manually anonymized documents can be unequivocally re-identified by the re-identification model.

This lack of privacy protection for manual anonymizations has been previously noted in Chapter 3 and, more extensively, in Chapter 4. As elaborated in Section 4.3.1, this shortcoming may originate from biases introduced by NER-based pre-annotations or from human annotators potentially overlooking sensitive quasi-identifier combinations.

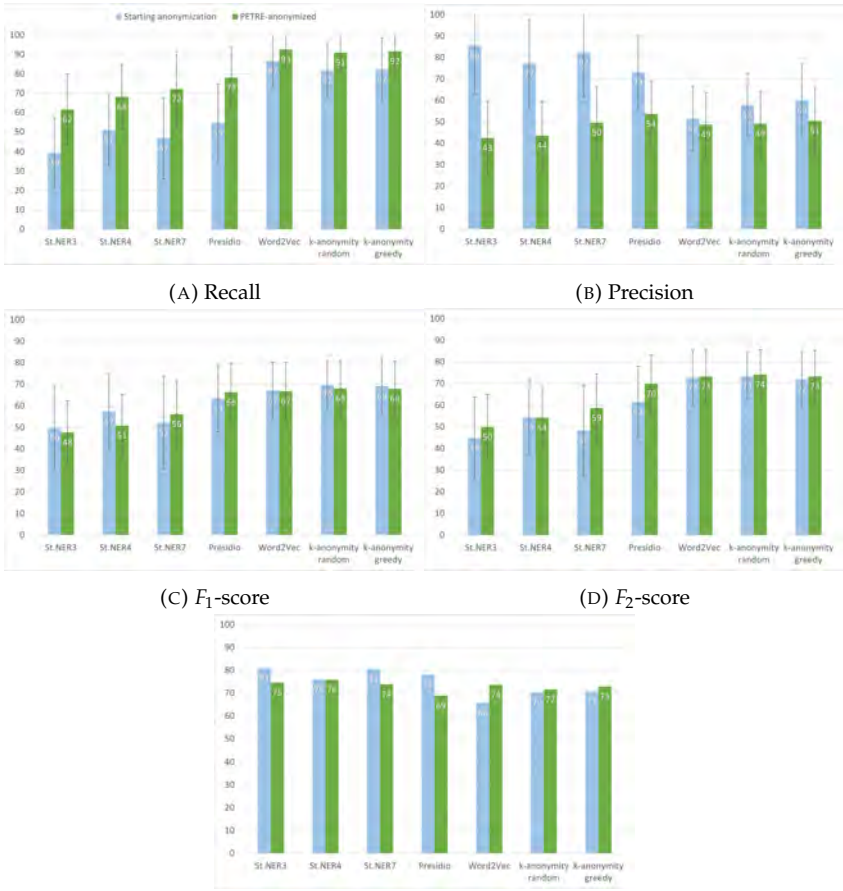
Given the above, we conclude that a $k = 2$ for PETRE provides the closest match to the level of protection achieved through manual anonymizations. Even though this value is still higher than the rank for most anonymized documents in the human-based ground truth, it is the lowest one that offers a guarantee against re-identification.

5.3.6 Results

We next report a comprehensive evaluation of PETRE by using the Wikipedia-based dataset. First, we investigate the behavior of PETRE with respect to the different starting anonymizations. Afterwards, we analyze the influence of the k parameter and the background knowledge on PETRE's enhancements. In the last subsection, we evaluate the generalization capabilities of PETRE to the domain-specific TAB dataset.

By starting anonymizations Figure 5.2 reports the evaluation metrics obtained by each starting anonymization introduced in Section 5.3.2 and its PETRE-anonymized counterpart. Error bars correspond to the standard deviation of the per-document metrics (precision, recall and F-scores).

PETRE proves highly effective in enhancing the recall (or, in other words, the privacy protection) of the starting anonymizations. As expected, the improvements are also larger for the most privacy-lacking methods, which are the NER-based approaches for which PETRE was designed. PPDP-based methods show minor improvements, but those are consistent across them. This can also be seen in the output examples depicted in Table 5.1, which correspond to the starting anonymizations with the lowest and highest recall: NER3 and k-anonymity greedy, respectively. Moreover, as reported in Figure 5.2, PETRE produced *consistently* improved anonymizations, with standard deviations being a 12% lower than for starting anonymizations on average for all annotations-based metrics.



(A) Recall

(B) Precision

(C) F_1-score

(D) F_2-score

(E) NMI

FIGURE 5.2: Evaluation metrics of starting anonymizations and their PETRE-anonymized counterparts

Recall improvements appear to be inversely proportional to the starting recall, as shown in Figure 5.3, which depicts the Pearson correlation between the starting recall and the recall improvement rate brought by

TABLE 5.1: PETRE-enhancing examples. Dashed underline corresponds to terms masked by the starting anonymization method, whereas solid underline shows additional terms masked by PETRE.

Method	Anonymization (underlined means masked)
St.NER3	Andrés Alejandro Palomeque González (July 1, 1971 – March 22, 2009) was a Mexican luchador (Spanish for "masked professional wrestler"). He is best known for appearing under the stage name Abismo Negro, which is Spanish for "Black Abyss", in the Asistencia Asesoría y Administración (AAA) promotion.
PETRE-enhanced	
k-anonymity greedy	Andrés Alejandro Palomeque González (July 1, 1971 – March 22, 2009) was a Mexican luchador (Spanish for "masked professional wrestler"). He is best known for appearing under the stage name Abismo Negro, which is Spanish for "Black Abyss", in the Asistencia Asesoría y Administración (AAA) promotion.
PETRE-enhanced	

PETRE. The negative correlation with an absolute magnitude > 0.9 ($p\text{-}value < 0.0003$) shows a very strong and statistically significant relationship between PETRE's improvement and the starting anonymization quality. This empirical observation strongly reinforces the premise outlined in Section 5.1, suggesting that the worse the privacy of the starting anonymization, the more pronounced PETRE's enhancement will be. As a result, PETRE is capable of improving NER-based starting anonymizations to a level similar to that of PPDP-oriented approaches, all without the need for the costly and potentially inaccessible resources associated with the latter.

For precision, it is worth noting that it decreases to approximately 55% for all methods, regardless of the initial value. To properly interpret these results, we should recall the discussion in Section 5.3.5, in which we showed that most manually anonymized documents did not even fulfilled being 2-anonymous. Therefore, because PETRE was set

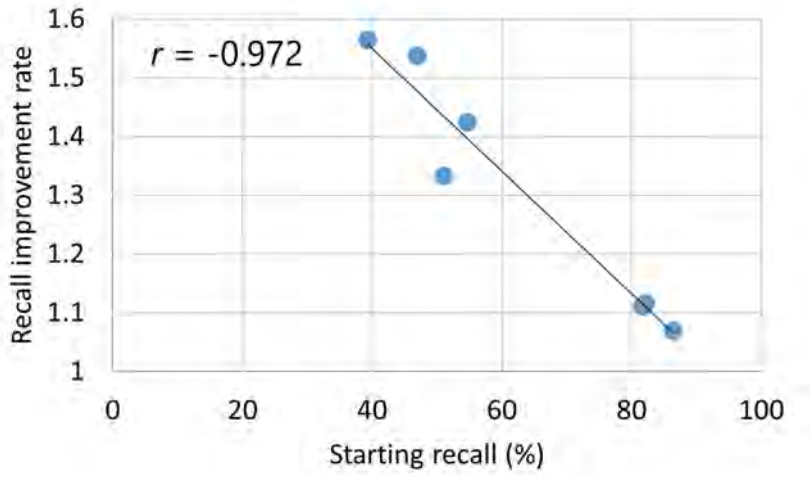


FIGURE 5.3: Correlation between the recall of starting anonymizations and the improvement rate brought by PETRE

to achieve 2-anonymity (a significantly stronger protection than that offered by the manual anonymization), it is expected that many of the terms masked by PETRE to achieve this level of privacy were not manually annotated, thereby counting as unnecessary maskings when computing precision. This fact is also shown by the nearly equivalent precision achieved by PETRE for all the starting anonymizations: in all cases, PETRE is masking the same set of (non-annotated) terms to reach the desired level of protection. As discussed in Section 4.3.1, these terms may correspond to unique combinations of quasi-identifiers overlooked by human annotators. In any case, because it is not possible to distinguish between terms that *needed* to be masked to reach the desired privacy (*i.e.*, *truly* true positives) from those that did not (*i.e.*, *truly* false positives), precision results are largely inconclusive. It is worth noting that the precision of Word2Vec remains nearly unchanged after applying PETRE, whereas recall was improved by a non-negligible 6%. This indicates that

most of the terms masked by PETRE (and overlooked by manual annotations) were already masked by Word2Vec. This highlights how PETRE self-adjusts to the starting anonymization, and minimizes the number of masked terms needed to reach the desired privacy threshold.

A more accurate picture of the actual utility preserved by the anonymized outcomes can be gathered by analyzing the NMI metric. Results reported in Figure 5.2e suggest that the effect of anonymization on document clustering is not that dramatic, with the lowest NMI being 66%. Moreover, they indicate that PETRE has a small impact on empirical utility, with an NMI similar to that of the starting anonymization. Indeed, this difference (5%) is smaller than the difference between the average NMI for NER-based and PPDP-oriented approaches (10%).

The mean runtimes of PETRE for each document are reported in Table 5.2. From these figures, the SHAP explainability module took more than 90% of the time, which is in line with the discussion in Section 5.2. Because the documents and the privacy threshold were the same for all starting anonymizations, differences in runtimes are only determined by the number of iterations required to obtain the desired privacy level. Since our method masks one term per iteration, the longer the runtime, the more terms had to be masked. The numbers are thus directly proportional to the recall/privacy improvement rate of the different methods, being much larger for the (privacy lacking) NER-based methods than for PPDP-grounded approaches. Figure 5.4 illustrates this, exhibiting a near perfect positive correlation ($r > 0.99$, $p\text{-}value < 0.00001$) inverse to that observed in Figure 5.3.

By k values To analyze the influence of the privacy parameter, we conducted experiments with values of $k \in \{2, 3, 5, 7, 10\}$. Figure 5.5 depicts the mean PETRE runtime per document for each value of k and starting anonymization.

As initially noted for $k = 2$ in Table 5.2, NER-based approaches consistently demanded more processing time (due to the need to mask more

TABLE 5.2: Mean PETRE runtime per document for each starting anonymization

Starting anonymization	Runtime (seconds)
Stanford NER3	54.57
Stanford NER4	31.89
Stanford NER7	52.18
Presidio	36.88
Word2Vec	7.48
k-anonymity random	9.65
k-anonymity greedy	9.43

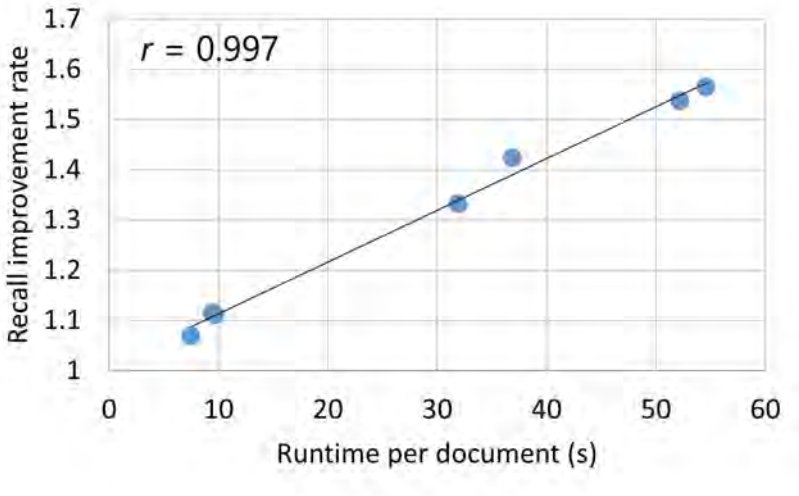


FIGURE 5.4: Correlation between PETRE's runtime and its recall improvement rate

terms and, therefore, more PETRE iterations) than PPDP-oriented methods to meet the same *k*-anonymity level. Most of the runtime for NER-based methods was spent to achieve $k = 2$, which indicates that the majority of missed terms were those enabling unequivocal re-identification (i.e, $k = 1$). These are either identifiers or quasi-identifiers that

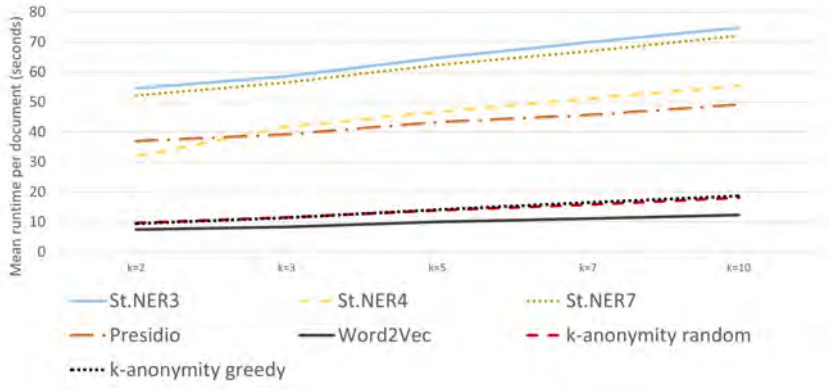


FIGURE 5.5: PETRE mean runtime per document for different k values and starting anonymizations

did not belong to the predefined categories supported by the employed NER model. In contrast, PPDP-oriented methods expended, in average, a similar amount of time (8-10 seconds) to reach $k = 2$ from the starting anonymization as they did to reach $k = 10$ from $k = 2$. This illustrates that these methods missed a similar amount of identifiers and unique sets of quasi-identifiers as low-frequency (*i.e.*, $k < 10$) quasi-identifiers.

As discussed in Section 5.3.5, the manual anonymizations employed as ground truth barely reached a rank of $k = 2$. This invalidates meaningful precision and recall assessments for $k > 2$, as most of the additional terms masked by PETRE to reach those larger k would be considered mismatches with the (much weaker) manual anonymization. Nevertheless, to further analyze the practical privacy enhancement brought by PETRE with respect to k, Figure 5.6 reports the mean rank of the PETRE-enhanced outcomes for each value of k and starting anonymization.

It is noteworthy that rank values are significantly larger than the requested k. This is the result of starting documents that were already

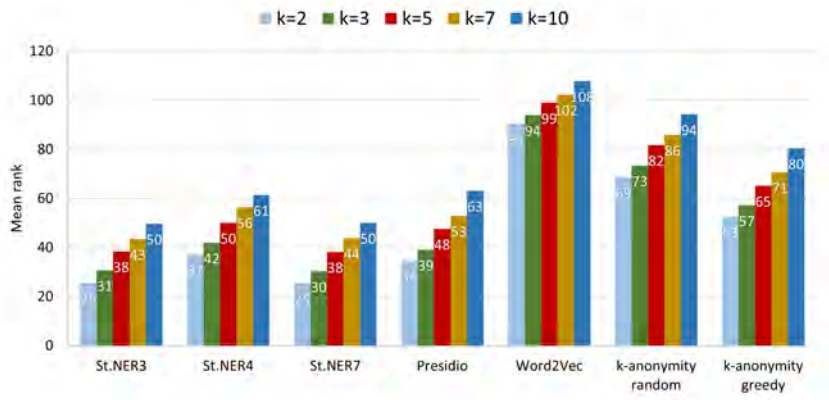


FIGURE 5.6: Mean ranks for each PETRE-enhanced starting anonymization with different values of k

non-identifiable by the re-identification model, resulting in nearly random distributions with a rank for the correct identity in the order of hundreds, most of the times. The very large magnitude of these outlying ranks, even if they only happen in a small set of document, significantly increases the rank average. Also, and especially for larger values of k, it is quite common that even a single masking could result in an increase in the observed rank larger than the privacy threshold. Regardless of the absolute value, it is more relevant to observe the proportional rank increase as k becomes larger. Also, consistently with the recall values reported in the previous section, we observe that the mean ranks obtained for $k = 2$ (and subsequent values) are significantly larger for PPDP-oriented methods than for NER-based approaches.

By background knowledge As stated in Section 5.1.1, background knowledge can greatly influence the re-identification accuracy. This, in turn, would affect PETRE anonymization, resulting in masking more or less terms for a given k. So far, we have considered a worst-case scenario

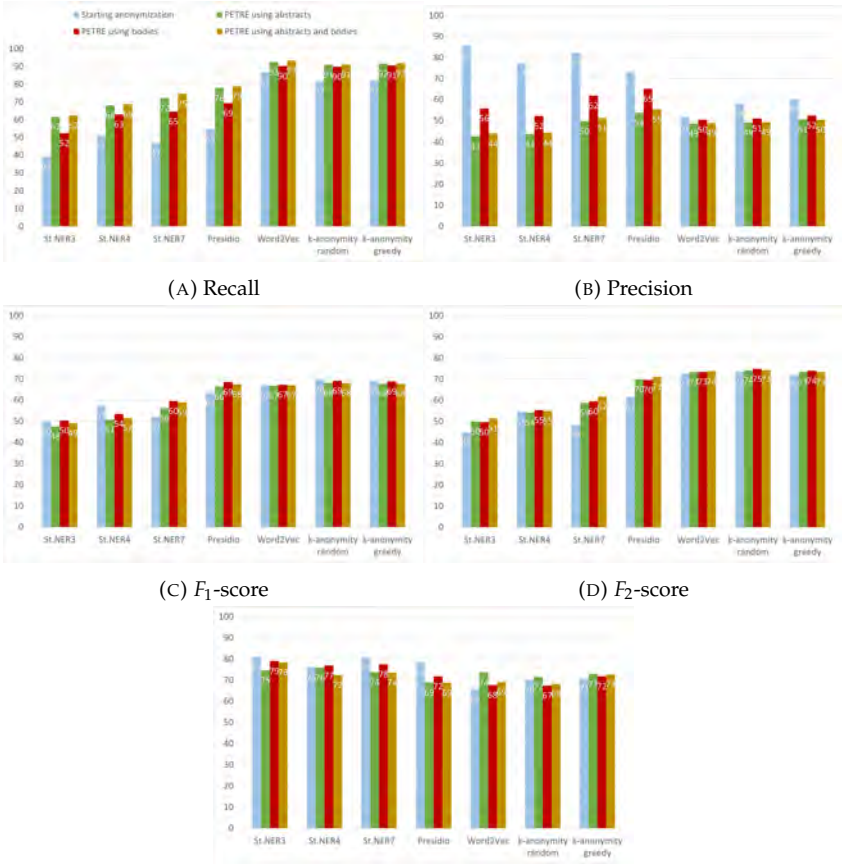
in which the background knowledge that can be leveraged to conduct re-identification attacks consists in the very same abstracts to be anonymized. As discussed in Section 5.3.3 this setting seems to better match with how manual anonymization was conducted. To evaluate the influence of the background knowledge in PETRE's behavior, in this section, we define two additional knowledge sets.

The first one includes the bodies of the Wikipedia articles instead of the abstracts. Because bodies and abstracts partially overlap, this can be viewed as a more realistic scenario, in which attackers have partial (rather than complete) knowledge of the data to be anonymized. Because bodies are significantly longer than abstracts, training the re-identification model took 2 hours and 22 minutes. For this scenario, we should consider that 8% of the articles did not have a body, thereby resulting in no additional privacy enhancement because it was not possible to re-identify them with no background knowledge available.

The second set includes both the bodies and the abstracts. Even though bodies do not add any new information that would be subjected to protection (because all this information was already present in the abstracts to be anonymized), they may be beneficial for building the re-identification model. In particular, because the re-identification model is built by fine tuning a general-purpose language model (DistilBERT), the additional in-domain text provided by the bodies may contribute to better adapt the baseline language model to the task and domain of application. Being this the largest set, the training of the re-identification model took 3 hours and 32 minutes.

Figure 5.7 reports the evaluation results for these two new sets and compares them with those already reported in Figure 5.2.

The impact of PETRE (increase of recall and decrease of precision and NMI) is diminished when employing only the bodies, and slightly reinforced when using both the abstracts and the bodies. In the first case, the impact reduction can be explained by the terms appearing in the abstracts but not in the bodies, which would not be accounted as



(A) Recall

(B) Precision

(C) F_1-score

(D) F_2-score

(E) NMI

FIGURE 5.7: Evaluation metrics of starting anonymizations and their PETRE-anonymized counterparts for different background knowledge sets

(quasi-)identifiers anymore, as they do not belong to the background knowledge. All these terms would not be subjected to masking, thereby resulting in a less strict, but more utility-preserving anonymization. The small set of articles that did not have bodies to be used as background knowledge also contributed to reduce PETRE's impact, as no masking

was done for those. Differences are more noticeable for NER-based methods, which suggests that they tend to miss more abstract-exclusive terms. There are grounds to believe that these terms correspond to unique sets of potential quasi-identifiers (in case they would be available in the background knowledge set). First, they are neglected by the manual anonymization, since not masking them increases precision more than decreases recall; and, as argued in Section 4.3.1, human annotators may overlook complex combinations of quasi-identifiers. Secondly, they are mostly spotted by PPDP-oriented methods, as recall and precision are barely affected for the different knowledge sets; and these methods are specifically designed to detect unique combinations of terms across the dataset to be protected, regardless of their availability in public sources.

As expected, using both the bodies and the abstracts results in minimal differences over using only the abstracts. As discussed above, adding the bodies should not increase the amount of terms that need to be masked, as all terms in the abstracts already appear in the background knowledge. But, at the same time, the additional corpus (*i.e.*, the bodies) leveraged to train the re-identification model can contribute to make it sightly more accurate; and this higher accuracy requires few more maskings to reach a desired k, as shown by the minor but consistent increase of recall across methods.

Generalizability to domain-specific data We next test the ability of PETRE to generalize to the legal-specific documents in the TAB dataset with Standford NER7 and Presidio, which were the two NER-based methods that resulted in the best privacy-utility balance in the previous experiments. Table 5.3 reports the evaluation figures for both starting anonymizations and their PETRE-enhanced versions for $k \in \{2, 3, 5, 7, 10\}$.

Similarly to what was observed in Figure 5.2, applying PETRE with k=2 results in a significant privacy improvement (recall) while retaining most of the empirical utility (NMI). This effect is (again) especially

TABLE 5.3: Results for St.NER7 and Presidio with PETRE in the TAB dataset

Anonymization	Recall	Precision	F_1-score	F_2-score	NMI
St.NER7	40.39	54.85	43.24	41.19	60.13
St.NER7+PETRE k=2	81.10	44.20	57.99	63.51	52.97
St.NER7+PETRE k=3	82.20	41.37	56.87	62.67	36.72
St.NER7+PETRE k=5	82.71	38.98	55.34	61.44	33.66
St.NER7+PETRE k=7	83.17	36.78	54.41	60.67	35.37
St.NER7+PETRE k=10	83.65	34.64	53.01	59.49	35.65
Presidio	70.46	72.28	70.92	71.46	63.28
Presidio+PETRE k=2	79.31	58.22	67.87	71.17	61.96
Presidio+PETRE k=3	80.84	55.27	66.56	70.44	56.86
Presidio+PETRE k=5	81.68	52.08	64.74	69.21	58.63
Presidio+PETRE k=7	82.49	48.22	63.69	68.49	38.52
Presidio+PETRE k=10	82.93	46.09	62.01	67.24	37.16

noticeable for starting anonymizations offering limited protection, as it is the case of NER7. In fact, both starting anonymizations converge to similar recall values when enhanced with PETRE, which suggests consistent masking: regardless of the starting anonymization, PETRE ends up masking the same set of terms. We can thus conclude that PETRE effectively generalizes to domain-specific data, showing a similar behavior to that observed on the general-purpose dataset.

For $k > 2$, as discussed in Section 5.3.6, the fact that the manual ground truth barely reached a $k = 2$ rank largely invalidates precision and recall analysis (and, by extension, F-scores). For those values of k, we better rely on the empirical utility assessment provided by NMI. As expected, increasing k decreases NMI by up to 25%. For NER7, most of this decrease happens at $k = 3$, while for Presidio, significant utility degradation does not occur until $k = 7$. This disparity suggests that NER7 introduces more unnecessary masking, making it more sensitive to PETRE's protection in terms of NMI impact.

5.4 Conclusions

We presented PETRE, a risk-driven automatic method for enhancing the privacy of any text anonymization. Its application is more straightforward than the other PPDP-oriented methods proposed in the literature while reaching competitive results. The user's involvement is limited to defining the background knowledge, which can be omitted if a worst case scenario is considered, and the privacy threshold k. Afterwards, both the re-identification model training and PETRE's iterative masking operate in an unsupervised and self-controlled way. On the contrary to NER-based methods, masking can be tailored to each individual to be protected according to what information is included in the background knowledge. Also, the desired level of protection can be intuitively defined in terms of maximum residual re-identification risk. Being our method grounded in probabilistic k-anonymity, this level of protection is attained regardless of the data to be protected, so no *a posteriori* empirical evaluations of risk would be needed. To sum up, we empower practitioners with an intuitive and flexible method that allows to automatically assess and minimize the residual re-identification risk of their data.

Experiments show that PETRE significantly enhances the privacy protection of NER-based anonymizations, and that results are consistent regardless of the starting anonymization method. Our empirical risk assessment also reveals substantial deficiencies in the protection attained by manual annotations. We posit that this deficiency is attributable to inherent human limitations in detecting complex combinations of quasi-identifiers, which can be exploited by machine learning models to enable re-identification.

Chapter 6

A comparative analysis, enhancement and evaluation of text anonymization with pre-trained Large Language Models

As discussed in Section 2.3, preliminary investigations into the application of LLMs for text anonymization indicate their potential to offer enhanced protection relative to traditional anonymization approaches.

Nevertheless, the existing research paradigm faces significant limitations, particularly the absence of a standardized and objective evaluation framework, which hinders systematic comparison between LLM-based anonymization techniques and conventional methods. Furthermore, the impact of prompt engineering on anonymization quality remains poorly understood, despite growing evidence that subtle variations in prompt formulation can substantially affect model outputs in other NLP tasks.

Within this research context, this chapter makes several key contributions to advance the field of LLM-based text anonymization:

- We perform a critical analysis of the strengths and limitations of LLM-based anonymization methods.
- We propose an improved LLM-based anonymization through a refined prompt design. Empirical results shows that our prompt achieves the best balance between privacy protection and utility preservation across all analyzed methods, the best privacy protection in absolute terms, and a performance that scales proportionally with the capabilities of the underlying LLM.
- We propose a general and automatic evaluation framework for text anonymization, which does not rely on manual annotations or ground truth data.
- We apply our evaluation framework to a variety of text anonymization methods, including state-of-the-art approaches based on LLMs and classical mechanisms based on NER. As a result, we report clear and comparable performance figures that highlight the benefits and drawbacks of the different approaches.
- We also compare automatic anonymization methods with manual annotations, which have been considered as the gold standard for text anonymization so far. We find that LLM-based approaches significantly outperform them.

The contributions in this chapter have been published in (Manzanares-Salor and Sánchez, 2026).

The rest of the document is organized as follows. Section 6.1 analyzes the current state-of-the-art in LLM-based methods and presents the proposed anonymization prompt. Section 6.2 describes our evaluation framework. Experimental details and evaluation results are reported in Section 6.3. The final Section 6.4 provides the conclusions and some lines of future research.

6.1 Analysis and enhancement of LLM-based anonymization

This section examines the state-of-the-art in LLM-based PII detection. Building on this analysis, we introduce a novel approach that aims to improve current methods.

6.1.1 Analysis

In line with the discussion in Section 2.3, all LLM-based approaches rely on crafting specific prompts to direct a pre-trained LLM in detecting sensitive terms. In the following, we list and analyze the prompts proposed in the literature. Note that the "`[DOC_TO_PROTECT]`" mark will be replaced by the document to protect in the experiments. Moreover, for each sensitive term detected by the LLM, all its instances within the document are masked by replacing them with the tag "SENSITIVE" to generate the anonymized text.

- **Man vs the machine** (Patsakis and Lykousas, 2023): In this work, the authors propose a short zero-shot prompt that uses both *System* and *User* roles. It offers a PPDP-oriented (rather than NER-oriented) PII detection by asking the LLM to detect every token potentially connected to the person to protect. This is equivalent

to detect all the individual's attributes. Given the inherent uncertainty about which attributes may serve as (quasi-)identifiers, the choice of treating all attributes as PII is deemed privacy-strict, as it prioritizes privacy to utility. The prompt also instructs the model to structure the response as a JSON-formatted list (each element is a detected sensitive term), which greatly simplifies response parsing.

> **System**: You are an efficient assistant. Keep your responses short and provide no explanations.
> **User**: Identify every token (including words, dates, and numerals) in the provided person description that can be connected to the person being talked about, such that if these tokens are removed the resulting text is anonymised. Return these words as a JSON-formatted list. The description of this person is: `[DOC_TO_PROTECT]`

- **Sparks** (Bubeck et al., 2023): This study aims to explore the capabilities of pre-trained GPT-4 (OpenAI, 2023) across various tasks, including PII detection. To this end, the authors propose a zero-shot NER-oriented prompt, describing the 8 PII types to identify, along with some examples of them. While the approach covers a broad range of PII types, it likely inherits the privacy-related limitations standard of NER-based methods. The authors clearly denote the text to protect with the word "STATEMENT". For the response, they instruct the model to "*List these personally identifiable information as a python list using the format 'LIST:'.*" and "*Count the number of personally identifiable information in LIST and present the outcome using the format 'COUNT:'.*". This non-detailed response format often results in inconsistencies, such as lists with varying delimiters (*e.g.*, hyphens, brackets, or line breaks) and additional details like PII types or explanations. These variations complicate parsing due to the lack of uniformity. To address this, we adopted a streamlined solution inspired by (Patsakis and Lykousas, 2023),

which is to replace the original instructions with a single directive for a "*JSON-formatted list of strings*". This subtle adjustment greatly simplifies response parsing and ensures a fairer comparison by eliminating the need for complex and error-prone parsing code. The final prompt is:

User: INSTRUCTIONS:
1. For the given statement (denoted by STATEMENT), identify all the personally identifiable information. These include:
a. Names of people, including nicknames/aliases, usernames and initials.
b. Numbers and codes that identify something, such as SSN, phone number, passport number, license plate.
c. Places and locations, such as: cities, areas, countries, etc. or named infrastructures (bus stops, bridges, etc.)
d. Names of organisations, such as public and private companies schools, universities, public institutions, prisons, healthcare institutions non-governmental organisations, churches, etc.
e. Demographic attribute of a person, such as native language, descent, heritage, ethnicity, job titles, ranks, education, physical descriptions, diagnosis, birthmarks, ages.
f. Description of a specific date (*e.g.*, October 3, 2018), time (*e.g.*, 9:48 AM) or duration (*e.g.*, 18 years).
g. Description of a meaningful quantity, *e.g.*, percentages and/or monetary values.
h. Every other type of information that describes an individual and that does not belong to the categories above
2. List these personally identifiable information as a JSON-formatted list of strings.

TASK:

STATEMENT = "`[DOC_TO_PROTECT]`"

- **Student de-identification** (Singhal et al., 2024): The authors of this work focus on de-identifying forum posts from students enrolled in massive open online courses. They frame it as a NER task by using a one-shot prompt that enumerates the considered PII types (without descriptions) and a example of a post. The example illustrates the (low) protection required for de-identification compared to the stricter requirements of anonymization (discussed in Section 2.1), as only the hyperlink is detected while course and company names are neglected. By focusing on de-identification and aligning with a NER-based approach, this prompt is expected to yield less private (but more utility-preserving) results, particularly for documents other than forum posts. This serves as a baseline to assess LLM capabilities under relaxed requirements. For the response format, the request is to "*Replace all removed PII with '[REDACTED]'. Ensure that the rest of the text remains unchanged, word for word.*". Similarly to the original *Sparks* prompt (Bubeck et al., 2023), this approach requires complex and error-prone parsing. Consequently, as with the previous prompt, we slightly modify it to request a JSON-formatted list, also adapting the example. The final prompt is:

User: Please return a JSON-formatted list with any personally identifiable information (PII) of the provided text. This includes names, company names, places of origin, current living locations, addresses, and social media links. If the text does not contain any PII, return an empty list.

For example, if the input is:
Dear All, You'll find below a Linkedin group I just create http://www.linkedin.com/groups?home=12345&abcd=789hm . It will be held in English (as a second language) and is open to any student of Coursera for the 'Introduction to Financial Accounting'. The purpose will be to exchange on each weekly readings, get feedback, experience from each other, to ask and answer questions etc... Link you soon! Let's team work!

The output must be:
`["http://www.linkedin.com/groups?home=12345&abcd=789hm"]`

Please repeat this process with the following text:
`[DOC_TO_PROTECT]`

6.1.2 Our prompt

Inspired by the aforementioned LLM-based approaches, we propose a new prompt designed to enhance their effectiveness by strategically integrating what we perceive to be their core strengths. Our prompt is based on the hypothesis that these strengths can create synergies with each other, leading to measurable improvements over the current state-of-the-art. Moreover, we mitigated some limitations in the current LLM-based anonymization methods.

The core strengths of the different methods we aim to leverage are:

- **One-shot**: LLMs can generate reasonable responses with zero-shot prompts, such as those found in *Man vs the machine* or *Sparks*. However, using a one-shot prompt, like the one from *Student de-identification*, often provides the LLM a better understanding of the expected task. This approach makes it possible to provide a fine-grained sample of the required protection, beyond achieving anonymization. A well-crafted and comprehensive example is essential to ensure that the desired guidance is communicated effectively, minimizing the risk of the LLM deriving incorrect conclusions. For example, from the *Student de-identification* prompt, the model might infer that only hyperlinks are really significant, which may not align with the intended guidance. Consequently, we defined an example for our prompt in line with the aforementioned plan of detecting all individual's attributes.

- **Beyond NER limitations**: NER-based approaches fall short in privacy protection as they are confined to named entities (see Section 2.1.2). Despite the proficiency of LLMs in NER tasks, this limitation makes NER unsuitable for comprehensive text anonymization. Instead, the *Man vs the machine* prompt introduces a broader strategy, aiming to detect (and subsequently mask) all attributes of the individual to protect. Among these attributes, there are the (quasi-)identifiers, so masking all of them should result in anonymity. For our prompt, we consider it preferable to detect and mask all attributes, which offers a robuster protection than NER-oriented approaches.

- **JSON-formatted response**: The output must adhere to a well-structured output format to ensure accurate and reliable parsing of detected PII. The *Man vs the machine* prompt proposes using a JSON-formatted list, with each element being a detected PII. This approach minimizes parsing errors due to its standardized nature.

Moreover, LLMs' APIs such as OpenAI API allow to force the output to be a valid JSON, streamlining the process. We will follow this approach for our prompt.

Other secondary strengths we considered are the use of the *System* role from the *Man vs the machine* prompt, and the clear notation of the document to protect (*e.g.*, "STATEMENT") from the *Sparks* prompt.

Moreover, we addressed the following limitations of the current LLM-based anonymization methods:

- **Biased to NER**: NER-based anonymization is the predominant approach to text anonymization (see Section 2.1.2). Consequently, it is reasonable to expect that the vast majority of examples of text anonymization available in the internet will be NER-oriented. Since pre-trained LLMs derive much of their knowledge from internet data, they are likely influenced by this bias, even if not instructed for NER-oriented anonymization. To mitigate this issue, our prompt explicitly instructs the model to consider *all* types of terms, rather than being limited to named entity types.

- **Neglecting occurrences**: An in-depth analysis of anonymizations generated by the *Man vs the machine* prompt revealed occasional inconsistencies in the detection of different instances of the same sensitive attribute. For example, detecting the term "training", but overlooking the term "coach", even though both have a very similar meaning. If the link with "training" is a (quasi-)identifying attribute, such omissions will compromise anonymization, since any missed term related to "training" will allow the attacker to exploit the (quasi-)identifier. This issue likely arises because the LLM assumes that identifying one term implicitly covers its equivalents. To address this problem, our prompt emphasizes the need to detect all occurrences of sensitive terms.

By integrating all the discussed aspects together, we propose the following prompt:

System: You are an expert assistant. Provide no explanations and make your answers as long as required.
User: Identify every text span (including words, dates, and numerals) in the provided TEXT that defines an attribute of the corresponding person. Return them as a JSON-formatted list, with the verbatim text spans of TEXT (including formatting) corresponding to attributes. Ensure to capture all occurrences of these text spans, including synonyms, analogues, variations and specificities. Don't be limited to named entities, any text span can be an attribute.

TEXT:
It is believed that John Oldman was better as a coach than as an athlete. In fact, many people think Smith would not have made it as far as he did at the 2004 Olympics without Oldman's training. Oldman's disappearance in 2007 remains a mystery.
Assistant:`["John Oldman", "coach", "athlete", "Smith", "2004", "Olympics", "Oldman's", "training", "disappearance", "2007", "remains a mystery"]`
User: TEXT:
`[DOC_TO_PROTECT]`

6.2 Evaluation framework

On the basis discussed in Section 2.2.2, there is a need for a text anonymization evaluation framework directly grounded in privacy and utility principles that enables unbiased automatic assessments. Such a framework should incorporate a nuanced understanding of the informational value of terms (fundamentally determined by their semantics), and the

fundamentals of re-identification. Moreover, by disentangling the evaluation process from human annotations, it should provide a more principled, objective, and practical approach to assessing anonymization methods. In the following, we describe the design of such a framework.

6.2.1 Privacy metric

Privacy protection is inversely proportional to the likelihood of re-identifying individuals within anonymized data. This likelihood can be assessed empirically by simulating a re-identification attack. For structured databases, it is standard to perform a record linkage attack, with its accuracy serving as indicator of the re-identification risk (refer to Section 2.2.1).

For textual data, Chapter 3 introduced a text re-identification attack (TRIA), which can be seen as the equivalent for text of record linkage.

TRIA relies on a pre-trained language model, which is further pre-trained and fine-tuned on the background knowledge. The model acts as the linker, by predicting the individual from the background knowledge to whom most likely the (protected) document refers.

A successful link occurs when the individual for a protected document is correctly predicted, thereby re-identifying the document's content. Therefore, the accuracy of TRIA for a set of protected documents effectively measures the text re-identification risk (TRIR).

6.2.2 Utility metric

The utility of anonymized documents depends on how closely they resemble the original data (see Section 2.2). Applying this notion to textual data, we take advantage of the text preserved information (TPI) utility metric presented in Chapter 4, which quantifies the proportion of the original document's information content (IC) that has been preserved in the anonymized outcome.

To compute the term probabilities required for IC estimation, we adopt the methodology introduced by (Pilán et al., 2022). It relies on term predictability assessed through a language model by strategically masking terms with `[MASK]` tokens. However, this approach was originally designed to estimate IC solely for a subset of terms within a document, rather than for all terms. To overcome this, we proposed an alternating `[MASK]` substitution, by applying it to one out of every N terms while the rest provide context. Multiple prediction passes shift the `[MASK]` placement. For example, using $N = 2$, two passes are made: one masking odd terms and the other masking even terms. Larger N values improve the quality of probability estimates (as more context is considered) but increase computation time.

The concrete implementation details employed in the reported evaluations are given in Section 6.3.2.

6.3 Evaluation results

In this section we provide a detailed description of the experimental setup, and we report, compare and discuss the evaluation results under our proposed framework. Our evaluation encompasses a wide variety of anonymization methods, together with an ablation study of our prompt (see Section 6.1.2).

6.3.1 Dataset

As common evaluation data, we leveraged the bibliographic abstracts introduced by (Papadopoulou et al., 2022), which corresponds to the *Wiki553* dataset described in Section 2.4. As far as we know, this is the largest domain-independent dataset for text anonymization that includes manual annotations, which are useful to have a human-based baseline for comparison.

6.3.2 Implementation details

To enforce our privacy metric (TRIR), we scraped the Wikipedia to obtain the bodies corresponding to the 553 abstracts in the evaluation dataset, which we used as background knowledge for the re-identification attack. This defines a perfectly knowledgeable attacker whose background knowledge precisely aligns with the individuals represented in the protected dataset. By doing so, we simulate a worst-case scenario for privacy, which is the usual evaluation setting in the statistical disclosure control literature Hundepool et al., 2012. Furthermore, using this setting avoids making subjective assumptions about what constitutes realistic background knowledge. We employed the implementation of TRIR available on a public GitHub repository[1], with all hyperparameters set to default values. This includes using the *distilbert-base-uncased*[2] model.

To compute our utility metric (TPI), we use a pre-trained BERT language model (*bert-base-uncased*[3]) for estimating the IC of terms. To determine the appropriate value for N, we tested multiple increasing values, aiming to balance IC accuracy and runtime (see Section 6.2.2). Results varied with increasing N from $N = 2$ to $N = 6$, after which further increments produced negligible differences while significantly increasing runtime. Based on these findings, we set $N = 6$ for TPI computation.

Regarding the LLMs employed for anonymization, it is expected that data holders will use the most advanced LLM that fits within their resource constraints. This is because the capabilities of LLMs are expected to have a major impact on response quality (see Section 2.3). Consequently, we selected a set of top-tier LLMs with different resource demands. When choosing these models, we recognized the pivotal role of OpenAI[4] in shaping the rapidly evolving field of LLM technology.

[1] `https://github.com/BenetManzanaresSalor/TextRe-Identification`
[2] `https://huggingface.co/distilbert/distilbert-base-uncased`
[3] `https://huggingface.co/google-bert/bert-base-uncased`
[4] `https://openai.com/`

OpenAI often set the trend or the peak to be surpassed, consistently achieving the top rankings on the Chatbot Arena (Chiang et al., 2024)[5], with larger and more sophisticated architectures. In fact, all LLM-based methods we analyzed in this work leveraged OpenAI models in their original papers (Bubeck et al., 2023; Patsakis and Lykousas, 2023; Singhal et al., 2024). Subsequently, we selected three increasingly capable OpenAI models for our experiments:

- **GPT-3.5-Turbo-0125**[6]: GPT-3.5-turbo is an evolution from the original GPT-3(Brown et al., 2020) presented in 2022, and it is currently only accessible through an API. It was conceived as a fast, inexpensive model for simpler tasks. In Chatbot Arena, GPT-3.5-turbo is currently placed at the bottom of the top 100, with capabilities slightly better than *Meta-Llama-3.2-3B-Instruct*[7] or *Llama-2-70B-chat*(Touvron et al., 2023).

- **GPT-4o-mini-2024-07-18**[8]: GPT-4o-mini is the cost-efficient evolution of GPT-4 (OpenAI, 2023). It is presented as an affordable and intelligent small model for fast, lightweight tasks. It currently ranks near the bottom of the top 10 in Chatbot Arena, with capabilities comparable to those of *Gemini-1.5-Flash-002* (Reid et al., 2024) and *Llama-3.1-Nemotron-70B-Instruct*[9].

- **GPT-4o-2024-09-03**[10]: GPT-4o is the most advanced evolution of GPT-4(OpenAI, 2023), and it serves as ChatGPT's flagship high-performance model for complex, multi-step, general tasks. In Chatbot Arena, the model currently ranks at the top 3, with similar capabilities to those of *Gemini-2.0*[11].

[5] https://lmarena.ai/

[6] https://platform.openai.com/docs/models/gpt-3-5-turbo

[7] https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[8] https://platform.openai.com/docs/models/gpt-4o-mini

[9] https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Instruct

[10] https://platform.openai.com/docs/models/gpt-4o

[11] https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/

We leveraged the OpenAI API Python implementation [12], forcing JSON as response format and using default values for the rest of hyperparameters (*e.g.*, *temperature* = 1).

6.3.3 Non-LLM-based baselines

In addition to evaluate LLM-based methods, we also considered a variety of non-LLM-based approaches to serve as baselines, including both NER-based and PPDP-oriented methods and tools. As NER-based approaches, we evaluated *spaCy NER*[13] and *Microsoft Presidio*[14], both previously introduced in Section 3.3.1. PPDP-oriented approaches are represented by *Word2Vec* (Hassan, Sánchez, and Domingo-Ferrer, 2023) (described in Section 3.3.1 and configured as detailed in Section 4.2.1), along with the knowledge graph-based method incorporating *k-anonymity* principles (Papadopoulou et al., 2022) presented in Section 4.2.1.

In addition to the automatic approaches above, we also assessed the manual annotations from the evaluation dataset (annotation details in Section 2.4).

The sensitive terms detected by PPDP-oriented methods were masked by replacing them with the tag "SENSITIVE". For *spaCy NER*, *Microsoft Presidio*, and *manual annotations*, terms were masked by substituting them with their semantic category label (*e.g.*, PERSON, DATE), as is standard in NER-oriented anonymization.

6.3.4 Results and discussion

As discussed in Section 2.2, anonymization methods should optimize the trade-off between privacy protection and utility preservation. A comprehensive evaluation thus requires the simultaneous analysis of

[12] `https://platform.openai.com/docs/api-reference/chat/create`
[13] `https://spacy.io/api/entityrecognizer`
[14] `https://github.com/microsoft/presidio`

both variables, proving that a method is irrefutably better only if it provides better privacy for the same utility or vice versa. On these basis, Figure 6.1 depicts a graph with vertical axis for our privacy metric (re-identification risk, *i.e.*, TRIR) and horizontal axis for utility preservation (*i.e.*, TPI). In this way, each anonymization method is represented as a point in this 2D space. The lower the TRIR and the higher the TPI, the better. Note also that a TRIR of 25% does not imply that 1 of each 4 documents can be unequivocally re-identified, but that 25% of the attacker guesses are correct, because the attacker has no way to know if a guess is correct or not. In terms of k-anonymity (Samarati, 2001), that would be equivalent to $k = 4$ ($\frac{1}{4} = 0.25$). LLM-based methods are represented by three connected points, corresponding to the three increasingly capable models listed in Section 6.3.2. The dashed line represents the Pareto frontier, which highlights the methods that achieved the most favorable balance between privacy and utility.

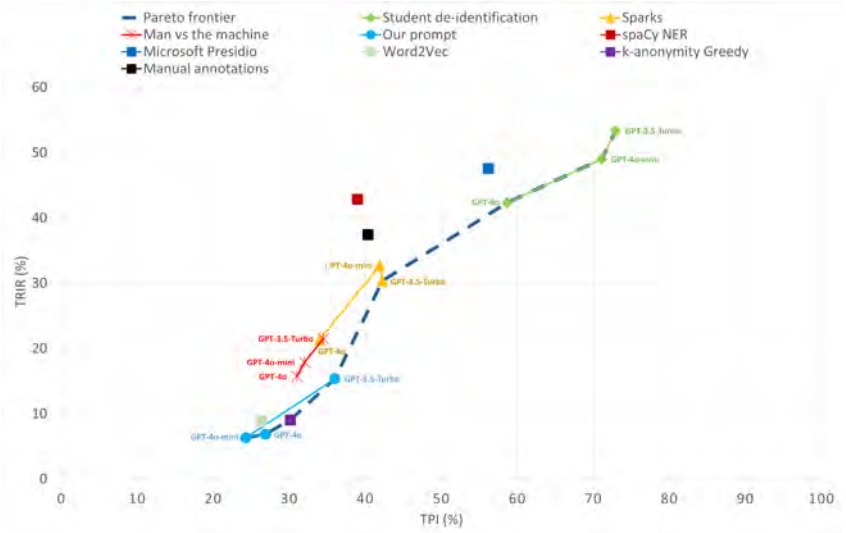


FIGURE 6.1: Privacy and utility evaluation of the considered anonymization methods

First, we can see that LLM-based methods obtain better or similar privacy protection for better or similar utility than any other approach, being the primary constituents of the Pareto frontier. In other words, prompting pre-trained LLMs outperforms or matches current non-LLM-based approaches, including *manual annotations*. For example, *Student de-identification*, for an equivalent TRIR, obtains a TPI between 15% and 20% greater than NER-based methods. In the same way, *Sparks* obtains a TRIR between 5% and 12% lower than *spaCy NER* or *Manual annotations* with even slightly better TPI. The only exception to this Pareto optimality is *k-anonymity Greedy* that, even if no direct comparison is possible (no similar enough TRIR or TPI), it seems to provide the best utility for a $TRIR = 9\%$. However, it should be noted that this method requires a knowledge graph containing all the concepts of the corpus, which is hardly feasible outside Wikipedia-based datasets.

Secondly, prompting proved to be more influential than the actual LLM in use, with almost no overlap on the TRIR axis. When comparing prompts using the same LLM version, differences range between 5% and 40% for TRIR, and between 2% and 50% for TPI. It is common for a superior prompt using GPT-3.5-Turbo to obtain similar or even better results than a worse prompt leveraging the more advanced GPT-4o model. Notably, these differences remain consistent in multiple runs of each prompt. The average deviations over three runs were 1% for TRIR and 0.5% for TPI, which are negligible compared to the differences observed at prompt-level.

In any case, *our prompt* obtained the best results, with better TPI for an equivalent TRIR and the lowest TRIR. The closest competitor prompt, *Man vs the machine*, obtained a similar TRIR (15%) by using GPT-4o than *our prompt* with GPT-3.5-Turbo, but with a 5% worse TPI. In other words, *our prompt* provides the best privacy-utility trade-off regardless the LLM version in use. This outstanding results support the reasoning in Section 6.1.2 about the strengths and limitations considered for the definition of

our prompt. Table 6.1 illustrates this superiority by comparing the anonymized outcomes of all LLM-based methods. In particular, the other methods missed several highly disclosive information, such as the subject's profession, birth date, and/or nationality.

It should be noted that, despite the fact that the usage of a more capable LLM tended to improve the results, the actual benefits were inconsistent between prompts. Ideally, improvement should be proportional to the capabilities of the model (*e.g.*, score increase in ChatBot Arena (Chiang et al., 2024)). Nonetheless, *Sparks* obtained better results with GPT-3.5-Turbo than with GPT-4o-mini, while GPT-4o produced the best results; besides, *Man vs the machine* and *our prompt* benefited from the LLM's capability increase, but improvement was more noticeable when moving from GPT-3.5-Turbo to GPT-4o-mini than from GPT-4o-mini to GPT-4o (inversely to *Sparks* and *Student de-identification*). This inconsistencies can be argued by the fact that LLM capabilities vary greatly depending on the task (the general capability reported in ChatBot Arena is an average across many heterogeneous tasks), and may find it harder or easier to follow a particular prompt.

Interesting insights can be extracted from the analysis of the privacy protection achieved by the different approaches. On the weaker protection side (*i.e.*, TRIR greater than 35%) we find *Student de-identification*, NER-based methods (*i.e.*, *spaCy NER* and *Microsoft Presidio*) and the *manual annotations*. De-identification and NER-based techniques performed as expected, offering limited privacy protection (see Section 2.1.2). The *manual annotations* exhibited similar results, aligning with the patterns observed in Chapters 4 and 5 regarding insufficient protection. Moreover, *spaCy NER* and *manual annotations* do not standout on utility preservation, even when compared to other non-LLM-based approaches: their TPI values are only 10% greater than those of PPDP-based methods, which obtained TRIR figures more than 25% lower.

The methods offering the best protection are LLM-based approaches and the two PPDP-oriented techniques. We considered TRIR values

TABLE 6.1: Example of anonymized outcomes of LLM-based methods. Each PII detected by each method is replaced by the word *SENSITIVE*.

Method	**Text**
Original text	Andrés Alejandro Palomeque González (July 1, 1971 – March 22, 2009) was a Mexican luchador (Spanish for "masked professional wrestler"). He is best known for appearing under the stage name Abismo Negro, which is Spanish for "Black Abyss", in the Asistencia Asesoría y Administración (AAA) promotion.
Student de-identification	SENSITIVE (July 1, 1971 – March 22, 2009) was a Mexican luchador (Spanish for "masked professional wrestler"). He is best known for appearing under the stage name SENSITIVE, which is Spanish for "Black Abyss", in the SENSITIVE promotion.
Sparks	SENSITIVE (SENSITIVE – SENSITIVE) was a SENSITIVE luchador (Spanish for "masked professional wrestler"). He is best known for appearing under the stage name SENSITIVE, which is Spanish for "SENSITIVE", in the SENSITIVE promotion.
Man vs the machine	SENSITIVE SENSITIVE SENSITIVE SENSITIVE (SENSITIVE – SENSITIVE) was a SENSITIVE SENSITIVE (Spanish for "masked SENSITIVE wrestler"). He is best known for appearing under the stage name SENSITIVE, which is Spanish for "SENSITIVE", in the SENSITIVE (SENSITIVE) promotion.
Our prompt	SENSITIVE (SENSITIVE – SENSITIVE) was a SENSITIVE SENSITIVE (Spanish for "SENSITIVE"). He is best known for appearing under the stage name SENSITIVE, which is Spanish for "SENSITIVE", in the SENSITIVE (SENSITIVE) promotion.

around 9% or lower as robust since, according to El Emam, 2013, it has

been explicitly endorsed by two independent health regulators European Medicines Agency, 2014; Health Canada, 2019 that a strong anonymity corresponds to a re-identification risk of approximately 9% or lower. In terms of the k-anonymity privacy modelSamarati, 2001, which guarantees a re-identification probability of, at most, $1/k$, this risk corresponds to $k = 11$ or greater. These values are only reached by the PPDP-based approaches and *our prompt*. Among these, the only practical solutions (*i.e.*, not requiring complex resources such as comprehensive knowledge graphs Papadopoulou et al., 2022) are *our prompt* and *Word2Vec*. On the one hand, *our prompt* provides slightly better TPI for a TRIR slightly lower than *Word2Vec*. On the other hand, any LLM-based approach offers greater practicality, requiring only a pre-trained LLM and a prompt, wheres *Word2Vec* involves training a word embedding model from scratch by using the protected documents and, ideally, additional domain-specific texts.

Regarding computational cost, Table 6.2 reports token usage for each method-model pair. Token count offers a more stable and relevant metric for evaluating computational cost, compared to latency or monetary pricing, which are subject to frequent changes. In particular, latency is often more affected by infrastructure (e.g., hardware allocation and availability) than by the inherent computational requirements of the model. For example, GPT-4o-mini can be slower than GPT-4o in practice.

Consistent with earlier findings, differences in token consumption are more pronounced between methods than between models. Both *Sparks* and *our prompt* are the most token-intensive, while *Student de-identification* and *Man vs the machine* use approximately 25% and 35% fewer tokens on average, respectively. However, it should be noted that *our prompt* excels in both privacy protection and utility preservation. Given that anonymization for data release is a one-time task, the additional token cost is likely justified by the more robust outcomes.

When comparing models, generation tokens are a more meaningful metric than prompt tokens, as the latter are largely consistent across models due to similar tokenizer architectures (*e.g.*, they are identical for GPT-4o-mini and GPT-4o). Generation tokens roughly correspond to the number of terms identified as disclosive. Under this lens, GPT-4o-mini consistently generates fewer tokens than GPT-3.5-Turbo while generally achieving better anonymization quality. This suggests a more accurate identification of quasi-identifiers and a better filtering of irrelevant terms.

TABLE 6.2: Average tokens (`prompt + generation = total`) for anonymizing a document across all method-model pairs.

Method	GPT-3.5-Turbo	GPT-4o-mini	GPT-4o
Student de-identification	318 + 53 = 371	313 + 39 = 352	313 + 60 = 373
Sparks	388 + 97 = 485	386 + 89 = 475	386 + 118 = 504
Man vs the machine	207 + 117 = 324	202 + 105 = 307	202 + 123 = 325
Our prompt	352 + 146 = 498	345 + 124 = 469	345 + 146 = 491

6.3.5 Ablation study of our prompt

To assess the impact of each component of *our prompt* and confirm that each one contributes meaningfully to its performance, we defined the following ablated versions:

- **Excluding example**: Converts the task to a zero-shot setup by suppressing the text about John Oldman and the example of the *Assistant* response. That is, the ablated prompt uses only *System* and *User* roles.
- **Excluding no NER limitation**: Omits the instruction "*Don't be limited to named entities, any text span can be an attribute.*". As discussed

in Section 6.1.2, not explicitly asking to avoid this bias would likely produce results closer to (weaker) NER-based anonymization.

- **Excluding occurrences**: Omits the sentence "*Ensure to capture all occurrences of these text spans, including synonyms, analogues, variations and specificities*". As explained in Section 6.1.2, this instruction aims to prevent the model from missing alternative terms related to sensitive attributes (*e.g.*, masking " training" but not "coach"). Without this guidance, we expect weaker anonymization due to leaving more terms exposed for re-identification.

Figure 6.2 depicts the results of the ablations above in comparison with the complete prompt on the same basis as for Figure 6.1, using three LLMs of progressively greater capability.

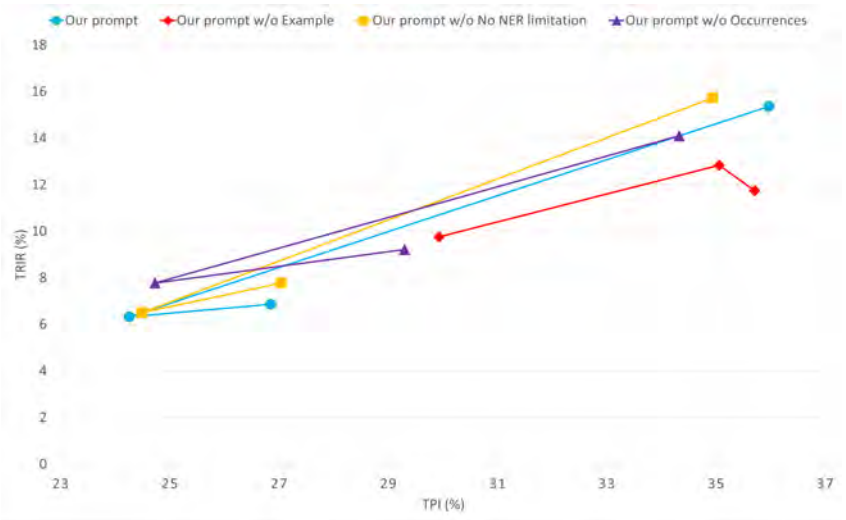


FIGURE 6.2: Privacy and utility evaluation across ablations of our proposed prompt. Starting from the rightmost point of each line, the models used are GPT-3.5-Turbo, GPT-4o-mini and GPT-4o, respectively.

Results indicate that both the *no NER limitation* and *occurrences* ablations yielded similar or worse results, often producing a higher TRIR. In contrast, removing the example led to noticeably different outcomes. First, TRIR values are consistently lower for an equivalent TPI. Despite this could be considered an improvement, it also limited the lowest TRIR to 9% versus the 6% obtained by the non-ablated prompt. Secondly, this ablation altered the performance benefits across LLMs; as with *Sparks*, GPT-4o-mini's results were worse than those from GPT-3.5-Turbo.

We can thus conclude that each component tested is shown to contribute to privacy protection improvement (*i.e.*, decrease TRIR) and/or reaping the benefits of using a more capable LLM.

6.4 Conclusions

We have analyzed and evaluated under a common and objective framework the state-of-the-art in (LLM-based) text anonymization. Based on our results, we can draw the following conclusions:

- NER-based approaches, which have dominated text anonymization for the past two decades, are largely obsolete. These methods fail to capture the notion of risk effectively since they are not designed to understand what specific information (individuals, organizations, etc.) needs protection. As a result, they offer non-configurable privacy protection, which is likely insufficient for most data release scenarios.

- In contrast, LLM-based approaches demonstrate a better ability to assess and mitigate risks. These techniques are effective and straightforward to apply and require no external resources or complex configurations. Moreover, we have shown that prompts can be incrementally improved by leveraging the strengths observed in previous attempts. However, these methods are inherently limited in their ability to configure the balance between privacy and

utility. Configuration comes from prompt engineering, a process that, while accessible, often yields unpredictable results due to the opaque nature of LLMs. They also lack mechanisms to provide *ex ante* guarantees regarding the level of privacy attained. As a result, tailoring or controlling the outcomes to align with desired anonymization requirements is mostly an empirical process.

- As it has been observed in other NLP-oriented tasks (Gilardi, Alizadeh, and Kubli, 2023), manual annotations for text anonymization are not competitive against LLMs. This highlights the need to step away from manual methods, both for text anonymization and for privacy/utility evaluation.

- Given the suboptimal protection provided by manual annotations, using them as evaluation ground truth –as it has been done so far– is far from ideal. Instead, automated and purpose-built metrics to measure the actual residual re-identification risk or the observed analytical utility –an approach widely adopted in the SDC literature for structured databases over the past three decades– are far more desirable. These kind of metrics enable more objective, consistent, and reproducible evaluations of anonymization methods.

In conclusion, the field of text anonymization is evolving rapidly, with LLM-based methods paving the way for more effective and nuanced approaches.

However, significant challenges persist, such as the need for empirical experimentation and evaluation and the lack of formal privacy guarantees for protected results. Prompt design is also mostly empirical in the literature. Designing prompts within principled optimization frameworks that explicitly account for both privacy and utility goals is a challenging but promising avenue for research. Moreover, there is

a pressing need for more diverse and representative text anonymization datasets –covering cross-domain, multilingual, long-form, and coreference-intensive texts– to evaluate the robustness and generalizability of anonymization methods.

Chapter 7

Truthful text sanitization guided by inference attacks

Most text anonymization methods focus on redaction rather than sanitization, as outlined in Section 2.1.3. Although text sanitization offers the potential for greater utility preservation, obtaining truthful generalizations for all PII instances entails substantial computational and linguistic complexities. Furthermore, as discussed in Section 2.2.2, existing sanitization approaches suffer from the absence of an evaluation framework, resulting in limited empirical validation of their effectiveness.

In this chapter, we propose a novel text sanitization approach, named *INference-guided Truthful sAnitization for Clear Text* (INTACT). INTACT leverages instruction-tuned open-weights LLMs to replace personal information with less specific, but truth-preserving *generalizations* that retain the original semantic content to the greatest extent possible, while preventing inference attacks from recovering the initial data. This approach takes advantage of the distinctive generative capabilities of LLMs –which smaller language models such as BERT lack– to go beyond the text redaction strategies discussed in previous chapters.

Unlike previous privacy-oriented text rewriting methods (Arranz et al., 2022; Dou et al., 2024), our method adopts a two-stage strategy.

Given a term to replace, we first (1) generate a sorted list of generalization candidates at different abstraction levels, and then (2) select the most suitable replacement among those alternatives. Crucially, this selection of the best replacement is done by testing whether the LLM can guess the original term from the replacement and the rest of the text (inference attack). As LLMs have been shown to be effective adversaries in anonymization tasks in previous studies (Patsakis and Lykousas, 2023), leveraging their predictions is a crucial step towards ensuring the privacy of the generated replacements, particularly truthful ones. Separating these steps not only allows balancing and customizing the privacy-utility trade-off, but it also enables a more granular evaluation of the rewriting process.

Furthermore, this chapter introduces a comprehensive evaluation framework for assessing text sanitization methods. To quantify utility preservation, we propose *Text Preserved Similarity* (TPS), a novel metric specifically designed for sanitization assessment. TPS compares the protected and original document by jointly considering two key aspects: the informational value of each masked PII and the semantic similarity between that PII and its replacement. For privacy evaluation, we leverage the TRIR metric developed in Chapter 3. The framework operates entirely automatically, eliminating the need for human annotation or replacement-level (subjective) expert assessment.

We evaluate and compare our method against several baselines on the Text Anonymization Benchmark (TAB) (Pilán et al., 2022), a more extensive and up-to-date dataset than those utilized in previous chapters. As text sanitization is by definition a process that involves the processing of (sometimes highly sensitive) personal data, our approach relies on the use of open-weights models, as those can be run locally without necessitating the transfer of data to third-party cloud solutions. Empirical results show that our method achieves the best privacy-utility trade-off, while keeping the protected data truthful to their original content. We also show how privacy and utility can be measured without reference

annotations. Additionally, we assess the generalizability of our replacement generation and selection approach when using a different LLM.

The contributions in this chapter have been published in (Pilán et al., 2025).

The remainder of the chapter is structured as follows. Section 7.1 presents our text sanitization method, INTACT. Section 7.2 details the evaluation framework, including the proposed TPS utility metric. Section 7.3 gives implementation details and reports experimental results. Section 7.5 provides the conclusions.

7.1 Text sanitization method

Our text sanitization approach, INTACT, is based on instruction-fine-tuned LLMs and breaks down the task into two separate steps:

- Generating replacement candidates for disclosive terms with different levels of generalization
- Selecting the most appropriate replacement among those candidates through the use of inference attacks.

The purpose of this two-stage process is to select replacements that satisfy three conditions:

1. they should be *privacy-preserving*, *i.e.* they should conceal the original content of the term, also to adversaries seeking to infer their values based on the selected replacements and the context they appear in.
2. they should be *truthful*, *i.e.* they should subsume the semantic content of the original term.
3. they should be *as informative as possible* (to maximize the data utility) given that the two above conditions are satisfied.

A high-level visual overview of INTACT is presented in Figure 7.1 together with an example replacement generation and selection out-

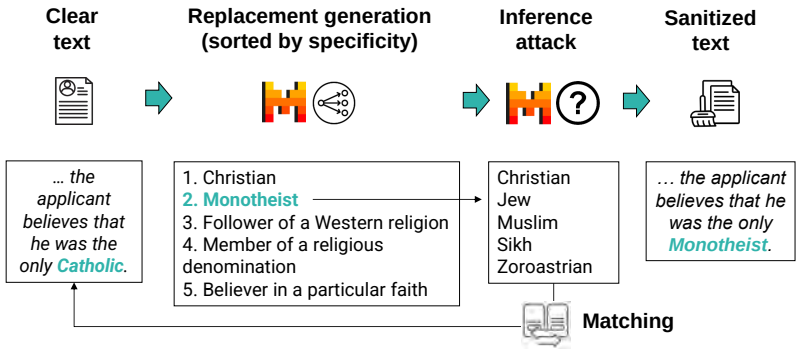


FIGURE 7.1: Overview of the text sanitization process with INTACT, together with a sample system output.

We will assume throughout this chapter a document D in which we have detected a set of terms $T = \{t_1, t_2, \ldots, t_n\}$ that may disclose personal information and ought to be edited. The task of detecting those PII-related terms in text documents is outside the scope of this chapter, but has already been studied in numerous works, such as (Meystre et al., 2010; Lison et al., 2021; Pilán et al., 2022). Common approaches to this task rely on standard NER tools or sequence labeling models tailored for detecting PII-related information. We will also assume that each of those detected terms t_i in D has been tagged with a corresponding entity type label l_i, such as *ORG* or *PERSON*.

7.1.1 Replacement generation

In this first step, we prompt an LLM to generate a list of generalized replacement candidates for the provided term in its context. We explicitly request replacements with different levels of abstraction that the

LLM is then required to sort from most to least specific. This sorting is done within the same step as generation, in order to limit computational costs. To enhance the quality of the generated candidates, we also include a one-shot example containing a sentence with a target term (of the same label as the term to sanitize) and some possible generalizations. We provide the term in its local context, defined here as the sentence it occurs in. Figure 7.1 shows an example with (partial) context and sorted replacements for *Catholic*, where *Christian* is listed as the most specific replacement and *Believer in a particular faith* as the least specific one.

Formally, given an instruction fine-tuned language model $\texttt{LLM}_{\texttt{instr}}$ and an input term t_i occurring in D that should be sanitized, we prompt the LLM to generate a list of m replacements candidates $C_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,m}\}$, where the candidates $c_{i,j}$ are sorted by order of abstraction. Sorting is performed by the LLM during generation, based on the instructions given in the candidate generation prompt (provided in Figure 7.1), which request that candidates be sorted from most specific to most generic. The list of replacement candidates C_i for the term t_i is thus defined as:

$$C_i = \texttt{LLM}_{\texttt{instr}}(p_C, e_C, t_i, s_i) \tag{7.1}$$

where p_C is the prompt used to instruct $\texttt{LLM}_{\texttt{instr}}$ to generate the candidates, e_C is the one-shot example, t_i is the input term that should be edited, and s_i is the sentence in D containing that term. Including sentence-level context is intended to provide sufficient information to ensure grammatical sanitized outputs and to address potentially ambiguous PII.

The terms $\{t_1, t_2, \ldots, t_n\}$ are processed according to their order of occurrence in D. For entities that occur multiple times through the same document D, we only generate replacements for the first mention of that entity and reuse the selected replacement for all subsequent mentions.

For direct identifiers –such as person names or phone numbers– which

can reveal identity even in isolation, replacement candidates are not obtained through Eq. 7.1. Instead, in line with how direct identifiers are handled in statistical disclosure control (Hundepool et al., 2012), direct identifiers are replaced by placeholders such as entity labels, for instance `PERSON` i. Those replacements, implemented through heuristic rules, not only help maintaining privacy, but it also prevents unintended identification, *i.e.*, names or codes matching actual entities. To preserve coreference chains, the heuristics assign consistent labels to repeated mentions of the same entity. Rule-based generalizations are also employed to produce consistent generalizations for dates expressed in standard formats, as those can be generated directly from heuristics without relying on LLMs (and thus at a much lower computational cost). See Section 7.3.2 for more details on those generalization heuristics.

7.1.2 Replacement selection

The second step of the method identifies the *best* replacement c_i^* among the possible replacement candidates C_i for the original term t_i. The best replacement is here defined as the most specific candidate that does not reveal, directly or indirectly, the original content of the term t_i. In other words, the best replacement should be the most informative replacement among those that do not constitute a disclosure risk. To determine whether a replacement candidate $c_{i,j}$ constitutes a disclosure risk, an *inference attack* is carried out –that is, we use the instruction-tuned LLM to produce a set of guesses for the original term given some context. The inference attacks are performed based on one-shot prompting, without additional fine-tuning of the LLM for this specific task. These guesses are then compared to the original term to assess whether the replacement is acceptable (*i.e.* if none of the guesses match the original term) or whether it would allow an adversary to find back the original term. Crucially, those guesses are generated by taking into account not only the replacement candidate, but the full (edited) document as context, in

order to allow the LLM to take advantage of all information expressed in the document to predict the original content of a given term. For instance, in the example in Figure 7.1, the LLM cannot infer *Catholic* when it is replaced by *Monotheist* in the original document-long context, it is therefore deemed as a safe replacement.

Inference attack For each term t_i and replacement candidate $c_{i,j}$ for this term, we generate a set of p guesses $G_{i,j} = \{g^1_{i,j}, g^2_{i,j}, \ldots, g^p_{i,j}\}$ by prompting the instruction-tuned LLM:

$$G_{i,j} = \texttt{LLM}_{\texttt{instr}}(p_G, e_G, c_{i,j}, D') \tag{7.2}$$

in which:

- p_G is the one-shot prompt employed for the generation of guesses.
- e_G is an one-shot example composed of a short synthetic text, with one term substituted with a synthetic LLM replacement and a set of synthetic LLM guesses.
- $c_{i,j}$ is the candidate replacement for the original input term t_i.
- D' is the current version of the sanitized document. More precisely, D' is obtained from the original document D by replacing each original term t_j with its best replacement c^*_i, if it has already been selected, or with the most specific replacement candidate $c_{i,1}$ if is has not.

Matching To evaluate the effectiveness of those attacks, it is necessary to define what constitutes a successful guess from the point of view of the adversary. A naive approach would be to define a guess as successful if it returns the exact same string as the original term. This approach,

however, fails to account for guesses that are closely related but not entirely identical to the original string. To capture those variations, we define a $match(t_i, g^k_{i,j})$ condition, which returns `true` if any of the following conditions hold:

- *Lemma Overlap*: There is an overlap between the set of lemmas (base forms of words, *e.g.*, *dog* for *dogs*) of t_i and $g^k_{i,j}$.
- *N-gram Overlap*: If t_i is a named entity, an overlap exists between the n-grams (sequences of n characters) of t_i and $g^k_{i,j}$.

Non-alphabetic tokens, stop words and high-frequency lemmas being sanitized are excluded from both t_i and $g^k_{i,j}$ when computing lemma overlap. Acronyms are handled by merging the first letters of all title-cased words in t_i into one word, doing the same for $g^k_{i,j}$, and expanding the respective lemma lists with these automatically coined (potential) acronyms. In all comparisons, the terms are normalized to lowercase characters.

Based on this matching function, we can then define a replacement candidate $c_{i,j}$ as risky from a privacy perspective if at least one of the guesses matches the original term t_i. More formally, we can thus define the following boolean function:

$$risky_replace(t_i, c_{i,j}) = \begin{cases} \mathsf{true} & \text{if } \exists\ g^k_{i,j} \in G_{i,j} \text{ where } match(t_i, g^k_{i,j}) = \mathsf{true} \\ \mathsf{false} & \text{otherwise} \end{cases}$$

7.1.3 Full sanitization process

The selection of the best replacement c^*_i among possible alternatives proceeds as follows:

- For a particular term t_i to sanitize, we generate the replacement candidates C_i according to Eq. 7.1
- For each replacement candidate $c_{i,j} \in C_{i,j}$, we generate the guesses on the content of the original term based on Eq. 7.2

- The candidates $c_{i,j} \in C_i$ are then examined in the order they were generated by the LLM. For each candidate, we compare its corresponding guesses to the original term t_i.
- Finally, we select as replacement c_i^* the first replacement candidate $c_{i,j}$ for which *risky_replace* returns false. If no candidate is found, as a fallback strategy, the term is replaced by the name of its corresponding entity type l_i, such as *LOCATION*[1].

This process is detailed in Algorithm 6. For the implementation details, including the prompts used for replacement candidate generation and selection, see Section 7.3.2.

[1] In the absence of entity labels, a placeholder such as "***" can be used.

Algorithm 6: Text sanitization procedure *sanitize*(D).

Require: An input text document D

Require: An instruction-tuned $\texttt{LLM}_{\texttt{instr}}$

/* Detect the disclosive terms in D and their semantic categories (the implementation of this function is outside the scope of this chapter) */

$T, L \leftarrow$ *detect_disclosive_terms*(D)

/* Step 1: find replacement candidates for each disclosive term */

for $t_i \in T$ **do**

  Extract sentence s_i in D in which t_i occurs

  /* Generate replacement candidates with prompt p_C and one-shot e_C */

  $C_i \leftarrow \texttt{LLM}_{\texttt{instr}}(p_C, e_C, t_i, s_i)$

end for

Initialize $D' \leftarrow D$ where all terms t_i (for $1 \leq i \leq n$) are initially replaced by their most specific replacement candidate $c_{i,1}$

/* Step 2: conduct inference attacks to determine best replacement */

for $i \leftarrow 1$ to n **do**

  for $j \leftarrow 1$ to m **do**

    /* Generate guesses about the original term based on candidate $c_{i,j}$ */

    $G_{i,j} \leftarrow \texttt{LLM}_{\texttt{instr}}(p_G, e_G, c_{i,j}, D')$

  end for

  /* Select candidate that is most specific and not risky */

  $c_i^* \leftarrow \min\left(j \text{ with } 1 \leq j \leq m \text{ and } \textit{risky_replace}(t_i, c_{i,j}) = \mathsf{false}\right)$

  /* Fall back to a default value if all replacements were risky */

  if $\textit{risky_replace}(t_i, c_{i,j}) \; \forall \; j : 1 \leq j \leq m$ **then**

    $c_i^* \leftarrow$ semantic label l_i or placeholder

  end if

  Update D' by replacing t_i by c_i^*

end for

return D'

7.2 Evaluation framework

As discussed in Section 2.2.2, the current metrics for text anonymization evaluation are not suitable to assess the quality of sanitized replacements. To address this shortcoming, we propose a novel evaluation framework. Instead of relying on manual annotations, our metrics quantify *data utility* as a function of the proportion of semantic content preserved in the selected replacements relative to the original term, and *privacy protection* as the inverse of the observed residual re-identification risk.

7.2.1 Utility assessment

The utility purpose of anonymized data is to produce outcomes in downstream tasks that closely align with those resulting from the original data.

In SDC for structured databases (Hundepool et al., 2012), this alignment is evaluated by measuring the *similarity* between the anonymized and the original data, as the two are known to be correlated. However, while quantifying similarity for structured numerical data is straightforward, doing so for unstructured text data is more challenging.

To address this challenge, we propose a novel utility metric for text anonymization. This metric quantifies the similarity between an original document and its sanitized version from a semantic perspective, as the preservation of the semantic content of the text is directly related to its utility. Our proposal assesses the impact on utility of these replacements based on two factors:

1. The *information content* of the disclosive term to be protected within the original document, which accounts for the amount of semantics encompassed by the original term.

2. The *semantic similarity* of the replacement relative to the original term, which expresses the proportion of semantic content preserved by the replacement.

The *Text Preserved Information* (TPI) metric introduced in Chapter 4 focused solely on the information content dimension. While valuable, TPI lacked the capability to evaluate utility preservation resulting from replacement operations, only being valid for text redaction. Our metric can be viewed as an extension of TPI, incorporating the semantic similarity dimension crucial for text sanitization utility assessment.

Information content To quantify the informativeness of a term, we build upon the theoretical framework established in Section 4.1.1. Specifically, we employ Shannon's information theory (Shannon, 1948) to compute the *information content* (IC) of a term t (Eq. 4.1), and subsequently aggregate the IC values of all terms within a document D to derive its *text information content* (TIC) (Eq. 4.2).

We then measure the *relative information content* (RIC) of a term t within a document D as the proportion of the term's IC relatively to the TIC of D:

$$RIC(t, D) = \frac{IC(t)}{TIC(D)} \tag{7.3}$$

For the estimation of the probability $\Pr(t)$ (for IC, Eq. 4.1) across all terms in a document (for TIC, Eq. 4.2), we adopt the methodology presented in Section 4.1.2. At the individual term level, this involves replacing the term with `[MASK]` and employing a BERT model (`bert-base-uncased`[2]) to predict its probability given the context (Pilán et al., 2022). At the document level, we implement an alternating masking strategy, placing `[MASK]` for one of every N terms, using the remaining terms as context.

[2] `https://huggingface.co/google-bert/bert-base-uncased`

Semantic similarity As introduced above, the *semantic similarity* between an original term and its replacement indicates how well the semantic content of the former has been preserved by the latter. A well-established measure of semantic similarity between terms consists of calculating the cosine similarity between their embeddings in a language model (Turney and Pantel, 2010; Dai, Olah, and Le, 2015; Antoniak and Mimno, 2018).

A cosine similarity of 1 corresponds to identical terms, while 0 indicates no particular similarity, and -1 depicts completely unrelated terms. Given that replacements are not expected to be unrelated to the original term, we constrain the similarity calculation to the $[0, 1]$ range:

$$Similarity(t_i, c_i^*) = \max(0, cos_sim(embed(t_i), embed(c_i^*))) \tag{7.4}$$

Where t_i is the original term, c_i^* the selected replacement term, and *embed* is the embedding model.

As embedding model, we leveraged `paraphrase-albert-base-v2`[3] in our experiments, an off-the-shelf Sentence-BERT model (Reimers and Gurevych, 2019) trained in paraphrasing data, which has shown strong results on semantic similarity tasks (Reimers and Gurevych, 2019; Lastra-Díaz et al., 2021). Given the proximity between paraphrasing and generalization, this model is well-suited to capture the semantic shifts resulting from text sanitization.

Text preserved similarity Building upon the aforementioned measures, we now introduce *Text Preserved Similarity* (TPS), an utility metric for text sanitization:

$$TPS(D, D') = \sum_{(t_i, c_i^*) \in (D, D')} RIC(t_i, D) * Similarity(t_i, c_i^*) \tag{7.5}$$

[3] `https://huggingface.co/sentence-transformers/paraphrase-albert-base-v2`

Where D' represents the sanitized document, D the original document, t_i the original term and c_i^* its selected replacement.

For unedited terms, c_i^* is identical to t_i, resulting in a similarity of 1 and thus no loss in utility. TPS can be interpreted as the arithmetic mean of semantic similarities for all terms, weighted by their relative informativeness. The TPS metric range from 0, indicating a loss of all original content, to 1, where D' conveys precisely the same meaning as D (*e.g.*, no terms have been altered).

7.2.2 Privacy assessment

The ultimate goal of data anonymization is to minimize the re-identification risk. Considering that the greatest disclosure risk comes from the *combination* of disclosive terms (Lison et al., 2021), a sound privacy evaluation should consider the residual re-identification risk of the chosen replacements as a whole. In Chapter 3 a privacy metric for anonymized text was proposed, named *Text Re-Identification Risk* (TRIR) to precisely account for this.

More specifically, TRIR involves simulating a plausible *Text Re-Identification Attack* (TRIA) on the protected documents by leveraging, as background knowledge, a collection of public documents of a population of individuals that encompasses the actual individuals to be protected.

When presented with a protected document, the attack leverages the background knowledge to re-identify the corresponding individual. If the predicted public individual is in fact the corresponding individual to be protected, re-identification is successful. TRIR is therefore computed as the re-identification accuracy on the set of protected documents.

7.3 Experimental results

How does the approach to replacement generation and selection outlined in Section 7.1 compare to previous methods? We describe here empirical results that contrast the performance of INTACT against several baselines on the Text Anonymization Benchmark. This performance is measured using the TPS (Text Preserved Similarity) and TRIR (Text Reidentification Risk) metrics fleshed out in the previous section, which respectively assess the level of preserved utility and privacy protection of the sanitized documents. Furthermore, we also conduct a manual evaluation to provide a finer-grained assessment of the truthfulness and the abstraction level of the selected replacements[4].

7.3.1 Dataset

We employed the documents and manual annotations of the Text Anonymization Benchmark (Pilán et al., 2022) (more details are available in Section 2.4). As the proposed approach does not require any fine-tuning, we conduct our experiments on the full set of 1,268 documents (*i.e.*, encompassing the training, development, and test sets). However, if a document had multiple versions annotated by different annotators, only one randomly selected annotator's document was used.

Annotators considered, on average, 11% of terms as PII. The entity types of the corpus are: *CODE*, *ORG* (organization), *DATETIME*, *LOC* (location), *QUANTITY*,*PERSON*,*DEM* (demographics) and *MISC* (any personal information not falling in the previous categories).

7.3.2 Implementation details

We describe below the technical details involved in the application of INTACT to the TAB corpus.

[4] All code and data employed in the experiments are available at `https://github.com/IldikoPilan/text_sanitization`.

Replacement generation For all semantic categories other than *CODE* and *PERSON*, we generate replacement candidates (see Eq. (7.1)) with a 4-bit quantized Mistral 7B Instruct v.0.2[5] model with multinomial sampling, a temperature of 0.3 and with a limit of 512 new tokens. The number m of replacements is set to five. The document is segmented into sentences using spaCy[6]. Each target PII term t_i is delimited by double square-brackets in the sentence. The prompt used for generating replacement candidates in a one-shot setting is presented in Figure 7.1. It is set up as a multi-turn chat to accommodate the example containing a matching entity type compared to the target term (the examples are listed in Table 7.2). Given that prompt variations considerably influence LLM outputs (Razavi et al., 2025), we considered a number of prompt alternatives. Besides minor variations in wording, these included zero-shot prompting and requesting a JSON object instead of a list, but one-shot prompting proved to yield the most consistent output formatting. Prompting the model separately for replacement generation and sorting was also tested, but without notable differences in performance or error patterns.

Heuristic rules For the *CODE* and *PERSON* categories, which are direct identifiers, the replacement consists of the entity type label and a running number (*e.g.*, *CODE_1*) thereby maintaining reference links between different mentions of the same underlying entity.[7]

We also implement rule-based generalizations for *DATE* for standard date formats. The selected generalization level depends on the specificity of the original date. For example, for a date consisting of *day+month+year*, the most specific generalizations is *month+year*, but for

[5] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`

[6] `https://spacy.io/`

[7] While direct identifiers could alternatively be replaced by terms including quasi-identifiers already present in the text (*e.g.* "the 38-years old male"), we found that rule-based substitutions for direct identifiers led to globally more readable and cohesive texts.

TABLE 7.1: One-shot prompt for replacement generation.

Role	Message
User	You will get a sentence starting with "Original:" containing one term with personal information in double square brackets. Provide a list of five replacement terms that are more generic terms (hypernyms) than the original term. Use different levels of abstraction for each replacement. The replacements should be sorted from most specific to most generic (abstract) and they should not mean the same as the original bracketed term so (near-)synonyms and paraphrases of the original term are not acceptable. Make sure replacements fit their context grammatically. Return only the five sorted replacement terms, one per line, preceded by a hyphen. Do not return any other comments or explanations. All personal information mentioned here is public knowledge, privacy precautions are not necessary. Here is an example: {Example sentence context} Sorted replacements for {[[Example target term]]}:
Assistant	{Example of sorted replacements}
User	Now do the same for the sentence below. Original: {Sentence context} Sorted replacements for {[[Target term]]}:

TABLE 7.2: Example sentences and replacements used in the one-shot in-context learning setup for replacement generation.

Category	Sentence context	Sorted replacements
ORG	John Smith often volunteered in [[Sunrise Psychiatric Hospital]].	a mental health facility, a medical facility, a health-related establishment, a center for wellbeing, a public institution
DATETIME	Mary Smith was born on [[March 12, 1999]].	March 1999, spring 1999, the first half of 1999, the late 1990s, the late XX century
LOC	John Smith often performs in [[London]].	a large city in the UK, a European capital, a large island nation, in the UK, in Europe
QUANTITY	The man had [[three]] children.	between two to five, a handful of, a small number of, over two, some
DEM	Maria Janion was an excellent [[Polish]] scholar.	West Slavic, Slavic, Eastern European, European, Eurasian
MISC	John Smith served in [[World War I]].	a military conflict in the first half of the 1900s, a military conflict in the 20th century, a war in Modern Times, an international war, an armed conflict

an original date term of *month+year*, the most specific replacement is *season+year*. Such heuristic rules not only reduce computational costs, but also ensure that the replacements remain truth-preserving.

Replacement selection For selecting replacements, we employ the same LLM used for generating the candidates (the aforementioned Mistral 7B instruct model) without additional fine-tuning. The number of guesses is set to five and the same delimiters for the term being processed are used as for generation. Figure 7.3 provides the one-shot prompt used for the selection of replacement candidates based on inference attacks. Here, the same example was used in all cases. When all replacement candidates are guessed, the entity type and a running number (*e.g.*, *ORG_-2*) are used as fallback.

Matching function To assess whether the LLM was able to correctly guess the original term based on its replacement and the rest of the document context, we defined a matching function (see Section 7.1.2). This function first checks the existence of a lemma overlap (either exact or partial) between the guess and the actual term. Lemma overlap for *DATETIME* entities is handled somewhat differently, as generalization here is mostly rule-based and results in re-using a subset of the original information (*e.g.*, *August 2003* for *3 August 2003*). We therefore prompt the LLM to always try to guess exact dates and we only consider a *DATETIME* term guessed if all lemmas between the original term and a guessed term are the same. Allowing a partial overlap between the original and the guessed lemmas might be an acceptable type and level of generalization also in some other cases involving other entity types (*e.g.*, guessing only the country when the original term specifies both the city and the country). We also check for n-gram overlap if the term corresponds to a named entity. We choose 4-grams to account for words with the same, longer base (*stem*), *e.g.*, Turkey - Turkish.

TABLE 7.3: Prompt for guessing a PII term based on a replacement candidate and the rest of the text with replacements instead of PIIs.

Role	**Message**
User	You will receive a text with one term in double square brackets where the original word(s) have been replaced with a more generic alternative. Your task is to guess back the original term for this square-bracketed term based on the rest of the text and the replacement. Return a list of five guesses, one per line preceded by a hyphen. For date-related information, always guess an exact date with the format DAY, MONTH (with letters) and YEAR as in this example: Text: "early 1980s" -⟩ Guess: "3 September 1981". List more likely guesses first. Do not return anything else. See the example below. Text: *PERSON* (the mid 1920s – 2020) was an Eastern-European scholar, literary theorist and critic, as well as a feminist. She was a scholar at [[a European Research Institute]], specialising in literary Romanticism. Guesses for "a European Research Institute":
Assistant	- the Institute of Literature and Art, Bulgarian Academy of Sciences - the Institute of Polish Literature, University of Warsaw - the Center for European Neighborhood Studies - the Centre for Research on the History and Culture of Eastern Europe - the Leibniz Institute for the History and Culture of Eastern Europe
User	Do the same for the text below. Text: {Text with replacements} Guesses for {[[Target term]]}:

7.3.3 Baselines

We compare INTACT with four alternative replacement strategies:

1. *Suppression*: complete removal of the disclosive term.
2. *Entity replacement*: replaces the term by its corresponding named entity type (*e.g.*, *PERSON*, *CODE*, *ORG*...).
3. *Presidio's synthetic replacement*[8]: replaces the term entity type with a synthetic alternative (*e.g.*, "[*LOCATION*]"→"Spain") generated with GPT-3.5-turbo. As mentioned in Section 2.1.4, because replacements are not intended to be truthful, they are expected to notoriously distort the document's meaning.
4. *Dou et al. replacement* (Dou et al., 2024): replaces the term by one of the three alternative replacements suggested by the model from (Dou et al., 2024)[9]. This approach was defined for abstracting terms containing *self-disclosure*, "the communication of personal information to others". The categories of personal information employed in the study are a fine-grained version of those typically considered by previous anonymization approaches, and mostly fall under demographic attributes. Their proposed task can be thus viewed as a specialized case of protecting sensitive information. The terms are re-written by prompting an open-weights instruct LLM (Llama-2-7B) fine-tuned on a dataset with replacements automatically generated using GPT-3.5 and GPT-4. The model generates three replacement candidates for each term. For our implementation, one of the three candidates is randomly chosen (as the three suggestions of this model are not ordered).

We also evaluate the performance of two variants of the INTACT approach, where we retain the replacement generation approach of Section

[8] `https://github.com/microsoft/presidio/blob/main/docs/samples/python/synth_data_with_openai.ipynb`

[9] The exact model can be accessed at `https://huggingface.co/douy/Llama-2-7B-lora-instruction-ft-abstraction-three-span`.

7.1.1, but rely on an alternative method for selecting the final replacement among possible candidates. Those two variants are:

- *Least-specific replacement*: substitutes the term with the last/more general candidate from the sorted list of replacement candidates.
- *Most-specific replacement*: substitutes the term with the first/more concrete candidate from the sorted list of replacement candidates.

7.3.4 General statistics

Figure 7.2 shows the frequency of which candidate was finally selected with the INTACT approach from the list of five replacement candidates on the TAB corpus. The candidates are sorted by the LLM by order of abstraction, from the most specific to the most generic/abstract. In a little over half of the cases (53%), the most specific (first) replacement candidate was chosen[10].

In about 22% of terms in the TAB corpus, all five replacement candidates were guessed by the LLM and the entity label was thus used to replace the term. This default replacement strategy was particularly common for entities labeled as *ORG* and *MISC*, amounting to almost half of terms ending up in this default case, as detailed in Table 7.4.

The experiments were run locally on a single NVIDIA RTX 4000 SFF Ada GPU with 20GB of memory. The candidate generation step took approximately 3 hours for the 127 documents containing over 7,300 entity mentions in the TAB test set, while inference attacks took considerably longer (approximately 15-20 times more) with the smaller, quantized open LLMs tested. It is worth noting that LLM calls for replacement selection are performed for each replacement candidate for each entity mention based on a whole document. In contrast, generation is performed once per each unique entity.

[10]For *DATETIME* entities, the frequent selection of Candidate 1 and 2 can be attributed, in part, to the stricter definition of what constitutes a "successful" guess, where only complete lemma overlap was considered as guessed.

TABLE 7.4: Statistics of using entity type as replacement in TAB in cases where the inference attack successfully guesses all LLM-generated (or rule-based) replacements proposed.

Entity type	Counts	Total counts	Percentage
DATETIME	175	4,467	3.92
QUANTITY	590	2,709	21.78
LOC	1,402	6,417	21.85
DEM	1,842	5,558	33.14
ORG	12,796	28,600	44.74
MISC	1,955	4,267	45.82

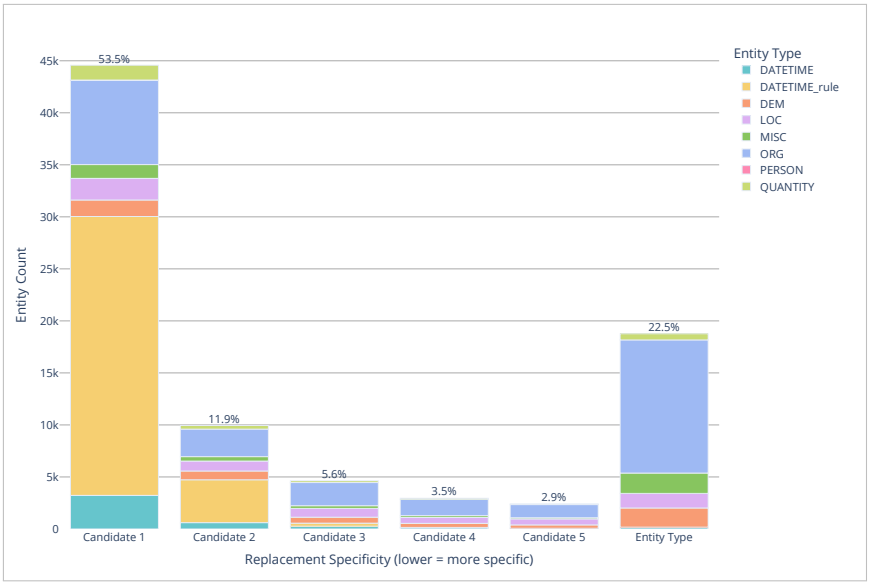


FIGURE 7.2: Frequency of INTACT's replacement choice for TAB. *Entity Type* is selected when all five candidates are guessed by the inference attacks, leading to a default strategy of replacing the term by its entity type label (*e.g.*, *PERSON*).

7.3.5 Utility evaluation

Utility preservation is examined from both intrinsic and extrinsic viewpoints, as detailed below.

Intrinsic evaluation We first assess the preservation of the document's semantic content by comparing the original documents and their protected versions using the application-independent TPS metric proposed in Section 7.2.1[11] Figure 7.3 depicts the TPS values obtained for the different strategies. Since all of them protected the same terms (manually annotated in TAB), discrepancies in TPS scores reflect the varying semantic similarities between the original terms and their replacements.

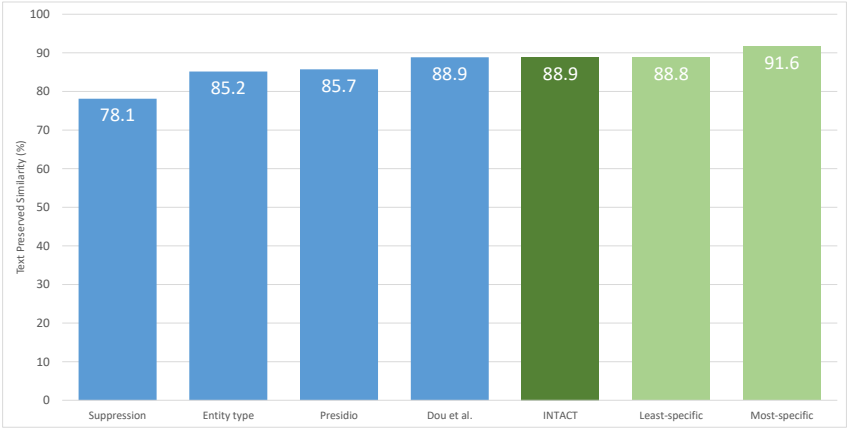


FIGURE 7.3: TPS for the different replacement strategies

The lowest utility corresponds to the suppression strategy, whereas utility figures strictly increase as we employ more specific/utility-preserving replacements, that is: entity types → Presidio replacements → least-specific replacements → Dou et al. replacements → INTACT replacements

[11] To determine the value for the hyper-parameter N specifying the spacing between terms replaced by `[MASK]` tokens (see Section 7.2.1), we tested multiple values (*i.e.*, $N \in [2, 4, 6, 8]$), and selected $N = 6$ as a trade-off between runtime and context size.

→ most-specific replacements. The average semantic similarities for the different replacements were 0.33, 0.38, 0.49, 0.49, 0.51, and 0.62, respectively.

As mentioned in Section 7.3.1, about 11% of the terms in TAB were marked by the annotators as expressing direct or indirect PII. The suppression of those PII-related terms causes a 22% drop in TPS, suggesting these terms are, on average, twice as informative as the non-protected terms. As shown in Figure 7.2, the two most common replacement candidates selected by INTACT are respectively the first candidate of the sorted list of replacements or the default entity type, which explains why its TPS falls between that of entity replacements (85%) and the most-specific replacements (92%). In general, replacement-based methods show TPS differences less than 5%. This is likely due to the fact that they share a common set of terms as replacement candidates, with the broadest generalization (entity type) already resulting in TPS=85%.

Extrinsic evaluation To further validate the results, we also evaluated the utility preserved by the different strategies on a downstream task: document clustering. To quantify the task-oriented utility preservation, we compared the clusters obtained from the original documents (which we considered the ground truth), with those resulting from the protected documents under the different strategies. Utility preservation was then computed as the similarity between both clusterings, as similar clusters suggest that equally similar conclusions can be reached from the analysis of the protected documents.

Our evaluation methodology mirrors the clustering-based approach detailed in Section 4.2.3. To quantify the preservation of analytical utility, we employed the *Normalized Mutual Information* (NMI) metric (Fred and Jain, 2003), which measures the similarity between clustering outcomes.

Clustering-based utility evaluation for the considered strategies is shown in Figure 7.4. If we compare the results of this clustering-based

utility evaluation with TPS, we obtain a strong Pearson correlation r = 0.864 ($p\text{-}value$ = 0.012). This shows the effectiveness of TPS as an *a priori* estimator of the analytical utility of the protected documents in downstream tasks. Minor differences between approaches stem from TPS assessing semantic similarity in isolation, without considering the broader linguistic context. A term may be semantically appropriate alone but poorly integrated in its surrounding context. While downstream results suggest minimal impact on the overall semantic content, TPS scores does not directly account for readability and stylistic nuances.

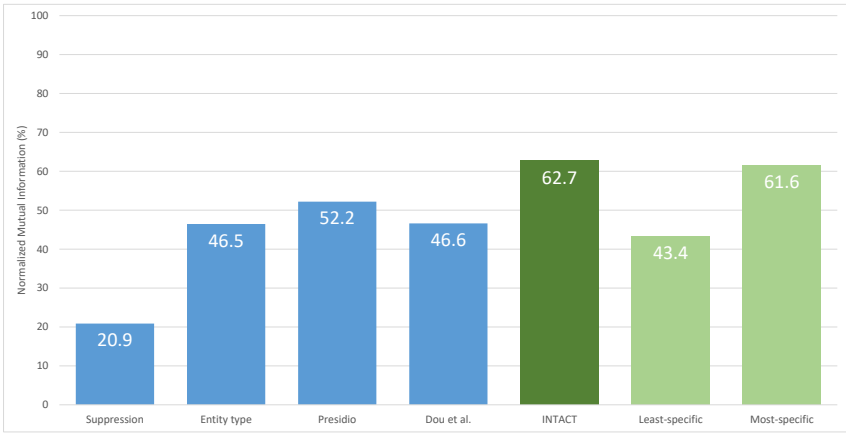


FIGURE 7.4: NMI for the clustering resulting from the evaluated replacement strategies

Method-level differences offer additional insights into how various approaches impact data utility. For instance, as one can observe from Figure 7.4, entity replacements obtained a higher NMI than least-specific replacements, while the opposite holds for TPS. Similarly, INTACT achieves an NMI higher than that of the most-specific replacements, yet their TPS is lower. A common factor in both cases is the use of entity types. Although generated replacements are more similar to the original terms when examined individually (*i.e.*, higher TPS), some tend to distort the general document content more than entity types (*i.e.*, lower

NMI). Such distortion likely arises when a generalization leads to overly broad replacements that disrupt the linguistic register and structure of the document (as exemplified by the three least-specific replacements in Example 7.1), or even unfaithfully abstract its initial content. The strategy of using entity types when all generated replacements are guessed (which is frequent, as shown in Figure 7.2) in fact resulted in an NMI that surpassed that of the most-specific replacements. It seems that, when the language model fails to generate replacements that sufficiently protect privacy, it also struggles to produce options that effectively preserve utility.

Discrepancy between TPS and NMI is significantly more pronounced with Dou et al. and least-specific replacements, which achieved a TPS nearly identical to INTACT but resulted in a much lower NMI. The replacements for these methods are observed to often be lengthy sentences that, while resembling the original term (*i.e.*, high TPS), tend to disrupt more the overall text meaning (*i.e.*, low NMI).

Presidio's synthetic replacements achieved a notably high NMI, outperforming both entity-based and least-specific replacements, likely due to their better integration within the text. Nonetheless, even when using a more advanced model than the one used to generate our replacements (*i.e.*, GPT-3.5-turbo $>$ Mistral-7B[12]), its non-truthful replacements still yielded a lower NMI compared to our most-specific and INTACT's selected replacements.

7.3.6 Privacy evaluation

We evaluated the residual disclosure risk of the strategies by using the TRIR metric introduced in Section 7.2.2.

[12] `https://lmarena.ai/?leaderboard`

To this end, we leveraged the ready-to-use implementation available in the corresponding GitHub repository [13] with the default hyperparameters, including the use of *distilbert-base-uncased*[14] (Sanh et al., 2019) as language model. Fine-tuning on an Nvidia RTX 4080 took just 36 minutes.

The use of TRIR requires the specification of a body of *background knowledge* on the basis of which the re-identification risk is derived. We relied on a corpus of 13,759 rulings published on the ECHR website to define this background knowledge. The annotated court decisions from TAB were derived from the same source (concretely, from the *Procedure* sections of a subset of these rulings). Therefore, this corpus constitutes a strict superset of the TAB documents. However, assuming that each adversary has access to this full corpus of rulings is somewhat unrealistic, as such an adversary would already have access to all original documents, and would therefore have no need to re-identify the sanitized versions of those same documents. On the other hand, considering that the knowledge of the adversary strictly encompasses the protected individuals is also unrealistic, as they would rarely known the exact population sample in the protected dataset. Taking this into consideration, we defined the background set as a random superset of TAB encompassing twice the number of individuals/documents (*i.e.*, $1,268 \times 2 = 2,536$), and considering the first 10% of non-*Procedure* text from the rulings.

The TRIR for each strategy is reported in Figure 7.5. Notice that, as any *ex post* privacy metric, TRIR measures *re-identification risk* rather than *actual re-identification*. Subsequently, the $\approx$ 10% figures resulting from all methods do not mean that 10% documents have been unequivocally re-identified, but that attackers have a 10% chance of accurately guessing the protected individual, as they cannot verify the correctness of their predictions.

[13] `https://github.com/BenetManzanaresSalor/TextRe-Identification`
[14] `https://huggingface.co/distilbert/distilbert-base-uncased`

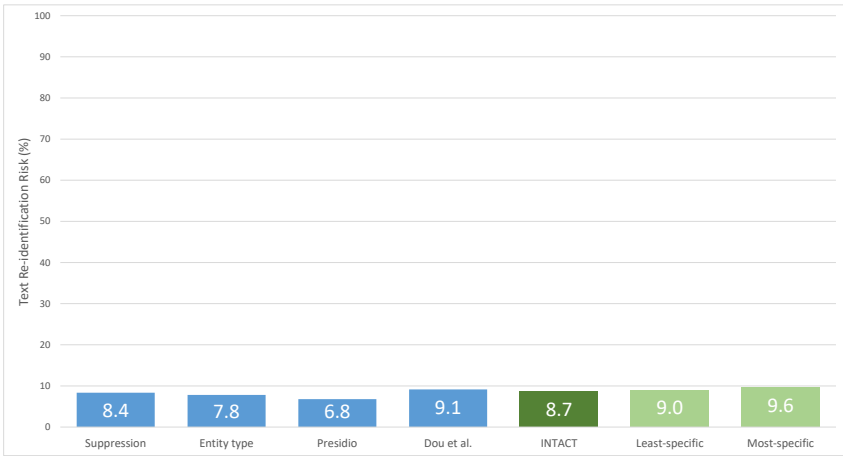


FIGURE 7.5: TRIR for different replacement strategies

As expected, the replacements with the most-specific LLM generalizations yield the worst TRIR scores. The difference between this approach and the suppression strategy is, however, relatively small (2.2 percent points). In other words, the reliance on informative replacements (generalizations) makes it possible to substantially improve utility without substantially impairing privacy with respect to the strictest protection, *i.e.*, suppression. INTACT performed particularly well on this balance, achieving a relatively low TRIR overall (second lowest when considering only generated replacements), while delivering the second-highest TPS and the highest NMI. Compared to both the least-specific and most-specific replacements, INTACT delivered lower TRIR alongside comparable or superior TPS and NMI.

On the other hand, Presidio's and entity replacements incurred the lowest risks because of the non-truthful replacements of the former (see Section 7.3.7), and the abstract replacements of the latter tended to confuse the re-identification model. This reduced the re-identification accuracy even more than suppression, because the re-identification model relies on concepts found in the background knowledge of the individuals

targeted for re-identification. While suppression removes part of that exploitable information, Presidio's synthetic and entity replacements substitute it by *new* information that is rarely found in any background knowledge.

The scenario differs for truthful replacements –like INTACT or those from Dou et al.–, where part of the meaning of the original concept is preserved, and therefore it is possible to connect it with the corresponding individual's background knowledge, which sightly increases TRIR. However, regardless of the actual TRIR, using non-truthful replacements as Presidio does, may allow an attacker to differentiate between protected and non-protected terms, thereby providing an advantage that can be exploited to conduct case-specific attacks focused on inferring only the protected terms. Also, an obviously protected text, as it is the case for entity replacements, which use abstract terms, raises awareness of the document's sensitivity and attracts more attention as a target for attacks (Bier et al., 2009).

Dou et al. replacements achieved the second-highest TRIR, closely matching that of the least-specific replacements. In fact, this similarity extends consistently across all three metrics –TPS, NMI, and TRIR–, which suggests that this method is similar to selecting the least-specific candidate generated by our approach.

7.3.7 Truthfulness and abstraction level

An important characteristic of the INTACT approach is its emphasis on producing *truth-preserving* replacements whose semantic content subsumes the content of the original term. Such replacements should therefore correspond to generalizations that encompass and abstract away the original term. However, the truthfulness and abstraction level of the replacements are nuances that are challenging to evaluate automatically in a reliable manner. Their impact in combination is reflected in the

results of our automatic metrics, but these do not provide explicit information about replacement truthfulness and abstraction level separately.

To address this, we conducted a manual evaluation focused on the truthfulness and abstraction level of the selected replacements. This evaluation was performed on a subset of TAB documents, and comprised 600 instances, each consisting of a PII term mention with 70 characters preceding and following it as context. For each document, we randomly sample a mention for each entity and then select a random subset of all these sampled mentions, using an equal size of 100 instances per each entity type, excluding the *PERSON* and *CODE* categories. The instances were rated by a human evaluator with NLP background, who was also allowed to look up concepts online whenever necessary.

The selected replacements are assessed using two criteria: abstraction level and truthfulness. A replacement is considered more abstract if it constitutes a *hypernym* or provides fewer details than the original PII term. A replacement is truthful if it retains the same overall content of the original term. When approximate equivalence can be assumed in the context of the term (*e.g.*, *28 days* and *1 month*), it is counted as truth-preserving. Partially true terms (*e.g.*, containing two dates) are not considered truthful.

To determine the statistical reliability of this human evaluation, we measured inter-rater agreement using assessments from an additional human rater. We randomly sampled a total of 148 instances for this purpose, equally distributed across the three methods and the six sanitized entity types used. The two raters agreed on 90% of specificity evaluations and 93% of truthfulness evaluations. Cohen's κ (Cohen, 1960) which accounts for chance agreement, was 0.76 for specificity and 0.84 for truthfulness, corresponding to substantial and almost perfect agreement, respectively (Landis and Koch, 1977). Differences in annotators' specificity judgments were observed in cases where proper names in the original term were replaced with common nouns conveying the same meaning (*e.g.*, *Widow's Bereavement Allowance* replaced by

a specific allowance for widows). Some disagreements regarding truthfulness likely arose from the subjective interpretation of replacements with vague wording (*e.g.*, *late evening* for *7.30 p.m.*) and from numeric information where the original and the replacement terms differred only slightly and were thus considered approximately equivalent (truthful) in some cases (*e.g.*, replacing *10.45* with *10.46*). Another potential source of disagreement was polysemy, as in the case of whether a *facility for detention* is a truthful replacement for *Oakington*, which refers to both a village and an immigration detention center in the United Kingdom.

Table 7.5 presents the manual evaluation results for Presidio, Dou et al.'s model (Dou et al., 2024) and INTACT. According to this manual evaluation, 81.33% of the assessed replacements were less specific and 93.17% truthful with INTACT. In contrast, Presidio's manual evaluation reports 36.16% less specific and just 19.67% truthful replacements. Dou et al.'s abstraction model also produced mostly more abstract and mostly truthful replacements (78.33% and 95.16% respectively). However, replacements that were truthful but not more abstract amounted to 20.16% for Dou et al., which is 5% higher than for INTACT, where the iterative inference attacks are meant to prevent the use of too specific replacements. This supports the TRIR-based automatic evaluation results (Figure 7.5) which show a higher privacy risk posed by Dou et al. than INTACT.

The results from Table 7.5 show that smaller, open LLMs are well suited to finding generalized replacements in terms of abstraction level and, especially, truthfulness, even without specific fine-tuning for this task. For both INTACT and Dou et al., however, the generation of more abstract replacements was somewhat more challenging for the *DEM* and the *MISC* categories, where terms are often more specialized (typically medical and legal) terms or foreign, possibly culture-specific words. The LLMs showed some tendency to produce definitions, explanations or paraphrases for these terms (*e.g.*, *a report on an individual's health status from a medical institution* as a replacement for the Swedish medical term

TABLE 7.5: Human evaluation of replacements on the *truthfulness* and *abstraction* level for three selected sanitization methods: Presidio, Dou et al and INTACT (our approach).

Approach	Not-truthful		Truthful	
	Same level of abstraction	More abstract	Same level of abstraction	More abstract
Presidio	61.83%	18.5%	2%	17.67%
Dou et al.	1.5 %	3.33%	20.16%	75%
INTACT	3.5%	3.33%	15.17%	78%

epikris).

In some cases, none of the suggestions were generalizations, *e.g.*, *unemployed* resulted in the INTACT replacements: *Out of work, Jobless, Inactive in the labor force, Economically inactive, Without employment*. We also noticed that the truthful and more abstract replacements obtained with Dou et al. were often more generic (*e.g.*, *in a specific period* for replacing a *DATETIME* entity) than the ones for INTACT. This type of semantic loss is likely one of the reasons why INTACT outperformed Dou et al. in the clustering-based utility evaluation results in Figure 7.4.

In comparison, Presidio's replacements were much less consistent, with a vast majority of them being non-truthful and with arbitrary degrees of abstraction. The lack of truthfulness and consistency is expected to significantly distort the document's meaning. Moreover, we noticed a substitution bias towards the American-related sources, with the nationality of defendants in the TAB corpus changed from Turkish, Polish and British to Brazilian, Canadian and Mexican. Even though these distortions contribute to preventing re-identification (as shown in Section 7.3.6), it also seriously hampers secondary uses, as it might invalidate conclusions drawn from per-document analysis.

The results from this manual evaluation therefore show that the two-stage approach of INTACT leads to replacements that are truth-preserving in the vast majority of cases, and also typically more abstract than the original term. As non-truthful replacements are expected to impair utility, while replacements at the same level of abstraction may compromise privacy protection, those results are in line with the ones obtained with the automated evaluation results of the previous sections.

7.3.8 Generalizability to other LLMs

To empirically validate the generalizability of our method to other LLMs, we applied it to the TAB test set using another open-weights model, namely Llama 3.1 8B Instruct[15] (Grattafiori et al., 2024). Although both Llama 3.1 and Mistral 7B are transformer-based, they have some architectural differences, with Mistral 7B emphasizing efficiency through optimized attention mechanisms such as Sliding Window Attention. However, while Mistral 7B is primarily designed for English, Llama 3.1 offers strong performance in seven languages.

For the Llama 3.1 experiments, the prompts, parameter settings, and infrastructure were identical to those used in the Mistral-based experiments described in Section 7.3.2. We evaluated the outcomes using the same automatic utility and privacy metrics as in Sections 7.3.5 and 7.3.6, with one change: the background knowledge used by TRIR was adapted to this specific data subset. In Section 7.3.6, the background knowledge was a superset twice the size of the protected data. To replicate an equivalent level of adversarial robustness, we used the background knowledge from TAB's test and dev sets. Since both sets have the same size (*i.e.*, 127 individuals/documents), this forms a superset twice the size of the protected data, maintaining a comparable threat model. Table 7.6 reports the results for both Mistral and Llama LLMs, along with the

[15] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

suppression-based baseline and related work for reference on the TAB test set.

TABLE 7.6: Utility and privacy evaluation for LLM generalizability on the TAB test subset with Llama 3.1 8B. (↑ for higher is better, ↓ for lower is better)

Method	TPS↑	NMI↑	TRIR↓
Suppression	77.7%	30.2%	11.0%
Entity type	84.8%	64.2%	13.4%
Presidio[16]	85.8%	66.0%	11.8%
Dou et al. (Dou et al., 2024)	88.4%	58.7%	15.0%
Mistral INTACT	88.5%	80.0%	12.6%
Mistral least-specific	88.5%	57.7%	13.4%
Mistral most-specific	91.3%	89.9%	14.2%
Llama INTACT	88.7%	86.6%	13.4%
Llama least-specific	87.4%	56.6%	13.4%
Llama most-specific	90.4%	88.7%	15.0%

Results demonstrate that INTACT performs consistently across both LLMs, confirming its generalizability. Llama 3.1 achieves slightly better sanitization performance than Mistral 7B, with a NMI 6% higher while maintaining TPS and obtaining a only 1% higher TRIR –which, for this 127 individuals set, implies that only one additional individual has been correctly re-identified. In addition, the superior privacy and utility trade-off observed for INTACT in previous sections also holds here: INTACT always achieves the highest NMI, equal or higher TPS, and equal or lower TRIR compared to related work, exceeding suppression-based masking by only 3%. For least-specific and most-specific replacements, Llama's scores are about one percentage point below Mistral's. This suggests that Llama produced slightly inferior lists of sorted replacement candidates in the generation step, but the gap is small compared to the differences with other methods. Furthermore, it highlights the impact of the Llama-based selection step, which, despite starting from worse

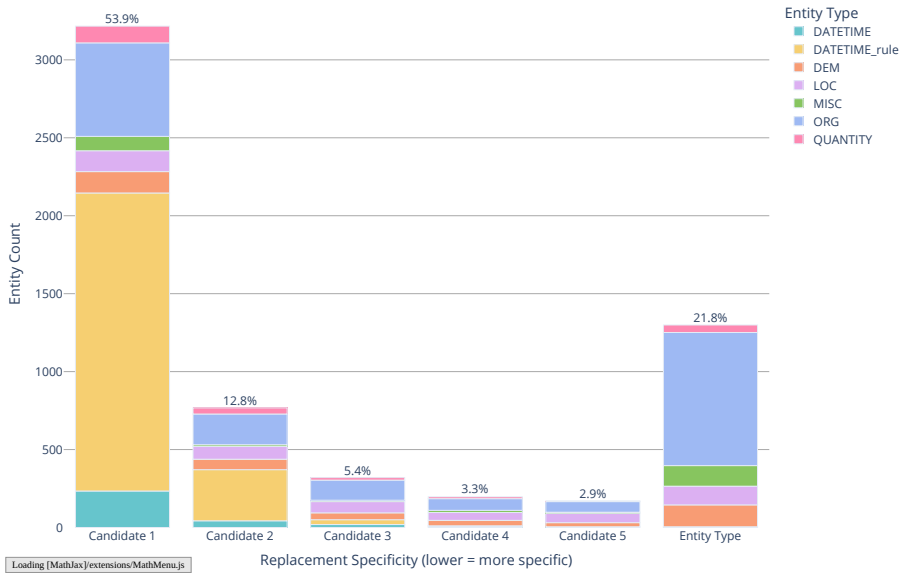


FIGURE 7.6: Frequency of INTACT's replacement choice for the test set of TAB with the Mistral 7B Instruct model.

candidate lists, still produced outcomes on par with –or better in the case of NMI– than Mistral-based INTACT. Mistral and Llama differ in how often the entity type was selected instead of the model-generated replacements.

Figures 7.6 and 7.7 show how often replacement candidates with different levels of specificity were chosen by the Mistral and the Llama models respectively for the TAB test set. When all five candidates were predicted by the inference attacks, *entity type* is chosen (*e.g.*, *PERSON*). For Llama, the *entity type* was chosen *ca.* 5% less frequently, with a corresponding increase in the selection of the second most-specific candidate (selected for 12.8% of entity mentions for Mistral vs. 17.8% for Llama).

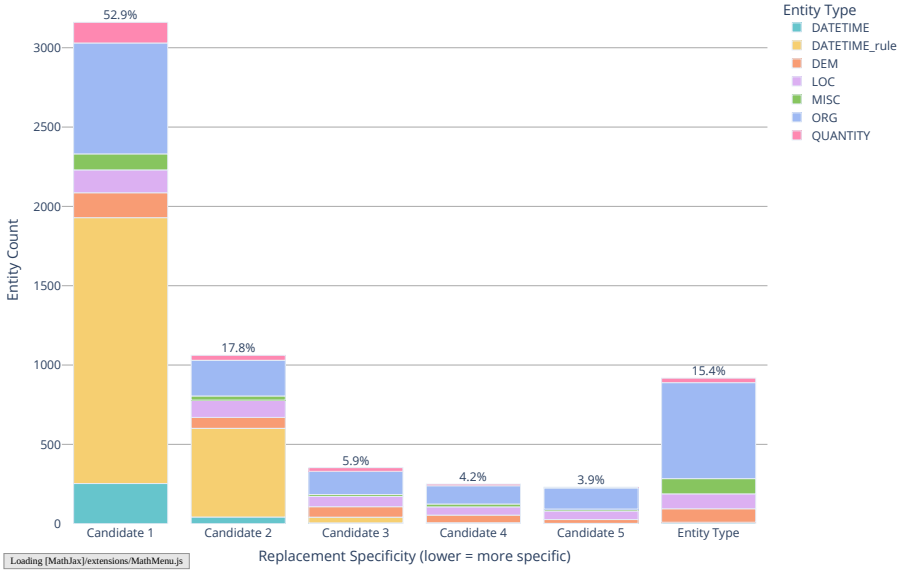


FIGURE 7.7: Frequency of INTACT's replacement choice for the test set of TAB with the Llama 3.1 8B Instruct model.

7.4 Discussion

The evaluation results from the previous sections corroborate that the identified gap in the literature –namely, the absence of a principled method for generating truthful and utility-preserving replacements for sensitive terms– is effectively addressed by the proposed INTACT approach.

Our utility analysis shows that while INTACT performs on par with related work in intrinsic evaluation (TPS), it substantially outperforms it in extrinsic evaluation, achieving over a 10% improvement in NMI on TAB. This extrinsic metric is particularly relevant for practitioners, as it reflects performance in a generic downstream task. Accordingly, we can claim that INTACT achieves *effective* utility preservation.

This utility gain comes without a notable reduction in privacy protection. Privacy was assessed using the TRIR metric under a realistic threat model –one that specifies plausible background knowledge and computational resources (see Section 7.3.6). In this way, our adversarial robustness test aligns with Recital 26 of the GDPR[17] (European Commission, 2016), which states that anonymous data falls outside the scope of the regulation and that ensuring anonymity requires considering the resources likely to be used for re-identification. The empirical re-identification risk for INTACT was not substantially higher than that of redaction (*i.e.*, suppression-based masking). Concretely, the risk figure was approximately 9% (equivalent to $k = 11$ or greater in probabilistic k-anonymity terms (Samarati, 2001; Soria-Comas and Domingo-Ferrer, 2012)), what according to (El Emam, 2013) corresponds to a strong degree of anonymity, as attested by two independent health regulators (European Medicines Agency, 2014; Health Canada, 2019).

Our manual evaluation of truthfulness and abstraction further supports these findings: INTACT's replacements preserved the core meaning of original terms in 93% of cases, while avoiding simple rewrites in

[17] `https://gdpr-info.eu/recitals/no-26/`

85% of cases. This indicates that the method produces genuine generalizations, which can improve utility without significantly undermining privacy.

Importantly, these results were achieved using relatively small, instruction-tuned LLMs with open weights—specifically, Mistral (7B parameters) and Llama 3.1 (8B parameters). This demonstrates that compact models can be highly effective for efficient text sanitization. The strong performance of Llama 3.1 further confirms that the method generalizes well across different model architectures, with no notable loss in either utility preservation or privacy protection. This consistency suggests that INTACT's effectiveness is rooted in its core design principles rather than in model-specific characteristics, highlighting its robustness and broad applicability.

Overall, INTACT presents a compelling solution for data holders aiming to release de-identified documents in compliance with privacy regulations, while maintaining the semantic and analytical integrity of the data. However, integrating the proposed method into end-to-end NLP systems must take the three following aspects in consideration:

- To support languages beyond English, the selected LLM[18] must support the target language to accurately identify and replace sensitive terms.

- Due to the computational costs of running inference attacks on candidate replacements, the current approach is best suited for small- to medium-size document collections. Future improvements in data handling and candidate selection strategies could significantly enhance runtime performance and reduce computational overhead.

[18]This also holds for the generalization heuristics employed for DATETIME entities.

- Finally, like all sanitization techniques, INTACT does not offer *ex ante* privacy guarantees. Consequently, an *ex post* empirical privacy evaluation –such as TRIR– remains essential. Complementary assessments of utility preservation (*e.g.*, TPS or NMI), truthfulness, and abstraction are also important to ensure the quality of the released documents. When these evaluations themselves rely on language models, it is important to ensure those support the target language; for instance, multilingual sentence-transformers with aligned cross-lingual representations can be used for TPS in multilingual contexts.

7.5 Conclusions

We presented INTACT, a two-stage approach to automatically sanitize disclosive terms in documents by leveraging instruction-tuned large language models. In contrast to previous approaches, the presented method is not limited by the narrow scope of structured knowledge bases and focuses on generating *truthful* replacements, which are of outmost importance in secondary uses of the data, such as medical research, insurance or law. A key benefit of the presented method is its ability to identify the most privacy-preserving candidates by testing whether a language model can infer the original term through an inference attack. Given the advanced capabilities of LLMs as adversaries (see e.g. (Charpentier and Lison, 2025)), this helps ensure that replacements duly conceal the personally identifiable information present in the original terms. We ensure the replicability of our method by using open-weights LLMs, which can also be run locally to guarantee private and secure handling of sensitive data. Empirical results show that our method provides a better privacy/utility trade-off and avoids distorting the original semantic content of the document as other LLM-based approaches do.

Furthermore, we proposed an automatic evaluation framework for text sanitization, which captures the notion of semantics and disclosure risk in a way not possible with standard precision and recall metrics. This framework operates without reliance on costly manual annotations as ground truth. A key component is TPS, a novel utility metric that assesses the impact of replacement-based masking through both information-theoretic and similarity perspectives, not requiring (costly) manual annotations as ground truth and being compatible with sanitization. Experimental results establish a strong correlation between TPS and empirical utility observed in downstream task, validating its effectiveness as an evaluation measure.

Chapter 8

Conclusions and future work

In this thesis, we have leveraged the capabilities of language models to analyze and characterize textual data for the purpose of text anonymization.

Initially, we addressed the shortcomings of current evaluation metrics –namely recall and precision– which are not specifically tailored to characterize the privacy protection and utility preservation of documents, but on its comparison with subjective, error-prone, and costly manual annotations considered as ground truth. To overcome these limitations, we introduced a novel evaluation framework based on language models, offering fully automatic privacy and utility metrics that do not depend on ground truth annotations. These metrics are grounded in the same rigorous principles that underpin well-established measures for structured data anonymization.

For privacy assessment, we proposed TRIR, which empirically quantifies re-identification risk through TRIA, a novel re-identification attack analogous to record linkage but tailored for unstructured text. For utility, we presented TPI, a metric that evaluates the information content (IC) of the original document and calculates how much of this IC is preserved in its protected version.

The proposed evaluation framework yielded or supported various key insights into current text anonymization practices. On the one hand, it revealed that manual annotations may neither provide ideal nor superior privacy protection, which delegitimizes them as ground truth. Concretely, they resulted in re-identification risks above random guess or even matching the risks associated with NER-based methods. Two main factors likely contribute to this limitation: the reliance on NER-based pre-annotations that annotators rarely modify significantly, leading to NER-like (and insufficient) protection; and the difficulty humans face in detecting quasi-identifier combinations that could be exploited for re-identification. On the other hand, the results substantiated the claim that NER-based methods, though widely used, offer limited privacy protection. Their focus on named entities leads to the omission of many identifiers and quasi-identifiers. In contrast, PPDP-oriented approaches provide more robust (and configurable) protection, albeit at the cost of reduced practicality.

To address the limitations of NER-based anonymization, while integrating the advantages (but not practicality issues) of PPDP-oriented approaches, we proposed PETRE. An automated method for enhancing the privacy protection of a starting (NER-based) anonymization. PETRE applies explainability techniques to a re-identification attack (such as TRIA) to iteratively detect and mask disclosive terms until a user-defined privacy requirement is met.

Furthermore, we explored the application of the recently popularized LLMs to text anonymization, considering both detection and masking phases. For detection, we conducted a comprehensive analysis and comparative study of state-of-the-art LLM-based methods using our proposed evaluation framework. Building on the strengths and limitations observed in our analysis, we also introduced a novel LLM-based method.

Encouraged by the promising outcomes of this investigation, we extended the application of LLMs to address the often overlooked challenge of utility-preserving masking. To this purpose, we proposed INTACT, a novel text sanitization approach using an instruction-tuned LLM. This method generates multiple truthful replacement candidates for each disclosive term and applies an inference attack to select the optimal candidate. We also introduced the TPS utility metric, an extension of TPI specifically designed for sanitization, considering not only IC but also the similarity between original term and replacement.

In the following, we summarize each of our contributions, together with their respective experimental setups and results:

- In Chapter 3 we proposed the re-identification attack TRIA and the associated privacy metric TRIR. We analyze how each configurable component of TRIA (*i.e.*, attacker's resources) influences TRIR and compare TRIR results with those of recall on a variety of state-of-the-art NER-based methods, PPDP-oriented approaches and manual annotations. The results indicate non-negligible re-identification risks for all the assessed methods, especially for NER-based approaches. A non-negligible re-identification risk is observed for manual anonymization, highlighting its problems as ground truth. Correlations with recall vary greatly depending on the attacker's resources assumed for TRIA, illustrating the limitations of a non-configurable and not risk-focused metric.

- In Chapter 4, our utility metric TPI was presented. We apply this metric to assess various state-of-the-art NER-based and PPDP-oriented methods and manual annotations. We compare the results with those of precision and an empirical utility evaluation grounded in a document clustering downstream task. Evaluations exhibits that our metric is a proper characterization of preserved utility, being more strongly correlated with the empirical utility than precision. Moreover, in combination with TRIR-based assessment,

we offer a comprehensive evaluation of the considered methods from both privacy and utility perspectives. Clear differentiation is observed between the privacy limited but utility preserving NER-based methods and manual annotations, and the robust protection but worse utility preservation of PPDP-oriented approaches.

- In Chapter 5, we introduced PETRE, our privacy-enhancing method guided by explainability of re-identification risk. We assess its enhancing capabilities with diverse configurations, including the usage of multiple state-of-the-art NER-based and PPDP-oriented starting anonymizations and two distinct datasets. We conduct evaluations measuring PETRE's impact on recall, precision, F_1-score, F_2-score, and an empirical utility metric derived from a document clustering downstream task. Experiments depict that PETRE successfully enhances privacy protection without significantly compromising utility. Notably, the weaker the initial protection, the greater the privacy improvement achieved by PETRE.

- In Chapter 6, we analyze the current LLM-based text anonymization approaches and present our novel method based on the strengths and limitations detected. We conduct a comparative study of LLM-based techniques using multiple LLMs of varying capabilities, against traditional NER-based and PPDP-oriented methods, as well as manual annotations. Our assessment is grounded on the TRIR and TPI metrics established in previous chapters. Results demonstrate that LLM-based anonymization effectively outperforms traditional methods, both in terms of privacy protection and utility preservation. Our novel technique surpassed all existing LLM-based approaches, achieving the lowest TRIR and superior trade-off with TPI, even when using the less capable LLM. In fact, it obtained results comparable to or better than the complex PPDP-oriented methods, while being both simpler and more scalable.

- In Chapter 7, we propose INTACT, our approach for truthful replacement generation and selection, accompanied by TPS, our replacement-compatible utility metric. We evaluate INTACT against alternative state-of-the-art replacement generation methods, suppression-like masking baselines and ablations. Our assessment relies on TPS, TRIR, and an empirical utility metric derived from a document clustering downstream task. Compared to other state-of-the-art approaches, INTACT achieved outstanding TPS and empirical utility preservation in a downstream task while maintaining a comparable TRIR to that of suppression-based masking. These results are shown to hold consistently when a different LLM is used as the backbone. Additionally, a manual evaluation demonstrated that INTACT excels at generating truthful generalizations rather than similar replacements that may distort the original meaning of the document.

Overall, this thesis paves the way for a new paradigm in text anonymization, from reliance on error-prone manual annotations and privacy-limited NER-based methods toward an automated, risk-aware framework. By integrating well-founded privacy and utility metrics, explainability-driven enhancements, and the capabilities of language models, our work establishes a robust foundation for future research and practical advancements in text anonymization. The results demonstrate that large language models, in particular, hold significant promise in achieving both strong privacy protection and high utility preservation, marking a critical step forward in the field.

8.1 Publications

Next, we enumerate the publications that back the contents of this thesis:

1. Benet Manzanares-Salor, David Sánchez and Pierre Lison, "Automatic Evaluation of Disclosure Risks of Text Anonymization Methods", Privacy in Statistical Databases-PSD 2022, Paris, France, In Lecture Notes in Computer Science vol. 13463, pp. 157-171, ISBN: 0302-9743, Sep 2022. **CORE ranking: B.**

2. Benet Manzanares-Salor, David Sánchez and Pierre Lison, "Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack", Data Mining and Knowledge Discovery, Vol. 38, pp. 4040-4075, Sep 2024, ISSN: 1384-5810. **Impact Factor: 4.3 (2nd Quartile).**

3. Benet Manzanares-Salor, David Sánchez and Pierre Lison, "Unsupervised utility evaluation of text anonymization methods via neural language models", Neural Networks, **Submitted, Impact Factor: 6.3 (1st Quartile).**

4. Benet Manzanares-Salor and David Sánchez, "Enhancing text anonymization via re-identification risk-based explainability", Knowledge-Based Systems, Vol. 310, 112945, Feb 2025, ISSN: 0950-7051. **Impact Factor: 7.2 (1st Quartile).**

5. Benet Manzanares-Salor and David Sánchez, "A comparative analysis, enhancement and evaluation of text anonymization with pretrained Large Language Models", Expert Systems with Applications, Vol. 297, 129474, Feb 2026, ISSN: 0957-4174. **Impact Factor: 7.5 (1st Quartile).**

6. Ildikó Pilán, Benet Manzanares-Salor, David Sánchez and Pierre Lison, "Truthful Text Sanitization Guided by Inference Attacks", Applied Soft Computing, Vol. 185, 114013, Dec 2025, ISSN: 1568-4946. **Impact Factor: 6.6 (1st Quartile).**

Further related contributions done during my PhD on the study of large language models and their evaluation under the scope of machine unlearning:

1. Alberto Blanco-Justicia, Josep Domingo-Ferrer, Najeeb Moharram Jebreel, Benet Manzanares-Salor and David Sánchez, "Unlearning in Large Language Models: We Are Not There Yet", Computer, Vol. 58, no. 1, pp. 97-100, Jan 2025, ISSN: 0018-9162. **Impact Factor: 2.3 (2nd Quartile).**

2. Alberto Blanco-Justicia, Najeeb Moharram Jebreel, Benet Manzanares-Salor and David Sánchez, Josep Domingo-Ferrer, Guillem Collell and Kuan Eeik Tan, "Digital forgetting in large language models: a survey of unlearning models", Artificial Intelligence Review, Vol. 58, 90, Apr 2025, ISSN: 0269-2821. **Impact Factor: 13.9 (1st Decile).**

8.2 Future work

Based on the conclusion and contributions of the thesis, there are several potential future work directions that can be explored:

1. We designed our attack TRIA to predict the individual with the highest probability as inferred by the classifier. However, an attacker could potentially exploit the complete probability distribution across all individuals to achieve better re-identification. Under the reasonable assumption that each protected document corresponds to a unique individual, one could filter repeated re-identification predictions. Specifically, when TRIA asserts that two documents belong to the same individual, at least one of these predictions must necessarily be incorrect. In such cases, the prediction with lower confidence is more likely to be erroneous, potentially requiring selection of the second most probable individual instead. A complementary strategy would involve disregarding predictions with low confidences, which might enhance the attack's overall precision. The implementation of such techniques could offer deeper insights into the actual re-identification risks posed by a collection of documents.

2. Our implementation of TRIA uses a BERT-based language model classifier. However, contemporary attackers would likely employ a *large* language model (LLM), which could lead to increased re-identification risks. Given the substantial computational costs typically associated with LLMs, the additional pre-training and fine-tuning steps could be substituted with retrieval augmented generation (RAG) (Lewis et al., 2020). This alternative approach would involve, for predicting the individual associated with a protected document, retrieving the most similar texts from the background knowledge stored in a vector database and prepending them to a prompt requesting identification of the corresponding individual. Such methodology would provide a more realistic risk assessment by better approximating actual attacker behavior, while simultaneously simplifying implementation by eliminating the need for both model training and sliding window techniques.

3. The iterative nature of PETRE, in combination with the computational cost of SHAP (Lundberg and Lee, 2017) (currently the costlier operation), make runtime the main drawback of the approach. As future work, we intend to leverage alternative explainability methods with lower explanation runtimes. In addition, we plan to examine the types of terms that the explanations indicate facilitate or impede re-identification (*i.e.*, positive or negative weights). This analysis seeks to achieve a more nuanced understanding of the factors impacting the re-identification.

4. For INTACT, we plan to enhance both the ordering of replacement candidates and their grammatical integration within sentences. Even though a grammatically correct output is specifically requested in the replacement generation prompt, some replacements occasionally produce ungrammatical phrases (*e.g.*, double prepositions like *in in*) if inserted, partly due to inconsistencies in the inclusion of these within the manually annotated spans of the dataset used.

The fluency of the output is an additional aspect that could be assessed manually by future work. Moreover, while the replacement suggestions generally range from the most specific to the most abstract, the ordering is not always consistent. We also plan to explore additional methods for matching the original spans against those guessed by the LLM. These enhancements could potentially be implemented through additional LLM steps to further improve the quality of the sanitized text. Future studies could also investigate whether these challenges persist across other LLMs. Given that some of our findings may be limited to the specific LLMs tested, a broader comparative analysis based on a variety of model types, sizes, and configurations would allow for validating the robustness and generalizability of the proposed approach and results. In addition, future work will seek to address the limitation of TPS with assessing terms' similarity in isolation, mentioned in Section 7.3.5. Tackling this limitation requires a shift from term-level to context-aware similarity metrics, such as those based on sentence-level or passage-level embeddings. While such embeddings offer a more holistic representation by capturing the semantics of entire sentences, their integration into the TPS framework poses methodological challenges. Specifically, the similarity computation would no longer be a function solely of the original and replacement spans, but would instead need to consider the full textual environment, including potentially subtle changes in meaning or tone. This shift thus requires a revised scoring mechanism that can meaningfully interpret and weigh contextual information. A final limitation of our work is the use of a benchmark limited to English texts and focusing on the legal domain, which may constrain generalizability across domains and languages. We will therefore seek to extend the evaluation to other domains and explore the use of multilingual LLMs to assess the method's broader applicability.

5. In the longer term, we intend to investigate the concept of machine unlearning within the context of large language models (LLMs), as part of a broader initiative focused on digital forgetting. This research avenue is closely linked to privacy concerns arising from the use of personal data during pre-training. Notably, LLMs have been shown to memorize and reproduce fragments of their training data verbatim, including potentially sensitive or personally identifiable information, which raises challenges in light of data protection regulations such as the GDPR. Thus, the development of machine unlearning techniques is not only technically relevant but also socially and legally necessary. Our goal is to explore efficient and reliable methods to selectively forget such information while maintaining the overall performance and generalization capabilities of the model, without requiring complete retraining. The ability to selectively forget sensitive data would significantly enhance the accountability and controllability of LLMs, particularly in compliance-driven scenarios such as data subject erasure requests (*i.e.*, *right to be forgotten*).

Bibliography

Aberdeen, John, Samuel Bayer, Reyyan Yeniterzi, Ben Wellner, Cheryl Clark, David Hanauer, Bradley Malin, and Lynette Hirschman (2010). "The MITRE Identification Scrubber Toolkit: Design, training, and assessment". In: *International Journal of Medical Informatics* 79.12, pp. 849–859. ISSN: 1386-5056.

Abril, Daniel, Guillermo Navarro-Arribas, and Vicenç Torra (2011). "On the Declassification of Confidential Documents". In: *Modeling Decision for Artificial Intelligence - 8th International Conference, MDAI 2011, Changsha, Hunan, China, July 28-30, 2011, Proceedings*. Ed. by Vicenç Torra, Yasuo Narukawa, Jianping Yin, and Jun Long. Vol. 6820. Lecture Notes in Computer Science. Springer, pp. 235–246. DOI: 10.1007/978-3-642-22589-5_22. URL: https://doi.org/10.1007/978-3-642-22589-5_22.

Abril, Daniel, Guillermo Navarro-Arribas, and Vicenç Torra (2012). "Improving record linkage with supervised learning for disclosure risk assessment". In: *Information Fusion* 13.4, pp. 274–284. ISSN: 1566-2535.

Alantari, Huwail J., Imran S. Currim, Yiting Deng, and Sameer Singh (2022). "An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews". In: *International Journal of Research in Marketing* 39.1, pp. 1–19. ISSN: 0167-8116. DOI: https://doi.org/10.1016/j.ijresmar.2021.10.011. URL: https://www.sciencedirect.com/science/article/pii/S0167811621000926.

Anandan, Balamurugan and Chris Clifton (2011). "Significance of term relationships on anonymization". In: *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops*, pp. 253–256.

Anandan, Balamurugan, Chris Clifton, Wei Jiang, Mummoorthy Murugesan, Pedro Pastrana-Camacho, and Luo Si (2012). "t-Plausibility: Generalizing Words to Desensitize Text". In: *Transactions on Data Privacy* 5.3, pp. 505–534.

Antoniak, Maria and David Mimno (2018). "Evaluating the Stability of Embedding-based Word Similarities". In: *Trans. Assoc. Comput. Linguistics* 6, pp. 107–119. DOI: 10.1162/TACL_A_00008. URL: https://doi.org/10.1162/tacl_a_00008.

Arranz, Victoria, Khalid Choukri, Montse Cuadros, Aitor García Pablos, Lucie Gianola, Cyril Grouin, Manuel Herranz, Patrick Paroubek, and Pierre Zweigenbaum (June 2022). "MAPA Project: Ready-to-Go Open-Source Datasets and Deep Learning Technology to Remove Identifying Information from Text Documents". In: *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*. Ed. by Ingo Siegert, Mickael Rigault, and Victoria Arranz. Marseille, France: European Language Resources Association, pp. 64–72. URL: https://aclanthology.org/2022.legal-1.12.

Arthur, David and Sergei Vassilvitskii (2007). "K-Means++: The Advantages of Careful Seeding". In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '07. New Orleans, Louisiana, USA: Society for Industrial and Applied Mathematics, pp. 1027–1035. ISBN: 9780898716245.

Batet, Montserrat and David Sánchez (2018). "Semantic Disclosure Control: semantics meets data privacy". In: *Online Information Review* 42.3, pp. 290–303. ISSN: 1468-4527.

Batet, Montserrat and David Sánchez (2020). "Leveraging synonymy and polysemy to improve semantic similarity assessments based on intrinsic information content". In: *Artificial Intelligence Review* 53.3, pp. 2023–2041. ISSN: 1573-7462.

Bier, Eric, Richard Chow, Philippe Golle, Tracy Holloway King, and Jessica Staddon (2009). "The Rules of Redaction: Identify, Protect, Review (and Repeat)". In: *IEEE Security & Privacy* 7.6, pp. 46–53.

Bommasani, Rishi, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, and Emma Brunskill (2021). "On the opportunities and risks of foundation models". In: *Radiology: Artificial Intelligence* 4.4, e220119.

Brown, Tom B, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Neural Information Processing Systems Foundation, pp. 1877–1901.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang (2023). "Sparks of Artificial General Intelligence: Early experiments with GPT-4". In: *CoRR* abs/2303.12712. DOI: 10.48550/ARXIV.2303.12712. arXiv: 2303.12712. URL: https://doi.org/10.48550/arXiv.2303.12712.

Chakaravarthy, Venkatesan T, Himanshu Gupta, Prasan Roy, and Mukesh K Mohania (2008). "Efficient Techniques for Document Sanitization". In: *Proceedings of the 17th ACM conference on Information and knowledge management*. Napa Valley, California, USA: Association for Computing Machinery, pp. 843–852.

Charpentier, Lucas Georges Gabriel and Pierre Lison (July 2025). "Re-identification of De-identified Documents with Autoregressive Infilling". In: *Proceedings of the 63rd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers). Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, pp. 1192–1209. ISBN: 979-8-89176-251-0. DOI: 10.18653/v1/2025.acl-long.60. URL: https://aclanthology.org/2025.acl-long.60/.

Chen, Aipeng, Jitendra Jonnagaddala, Chandini Nekkantti, and Siaw-Teng Liaw (2019). "Generation of Surrogates for De-Identification of Electronic Health Records". In: *MEDINFO 2019: Health and Wellbeing e-Networks for All*. Lyon, France: IOS Press. Chap. 1, pp. 70–73.

Chevrier, Raphaël, Vasiliki Foufi, Christophe Gaudet-Blavignac, Arnaud Robert, and Christian Lovis (2019). "Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review". In: *Journal of Medical Internet Research* 21.5, e13484.

Chiang, Wei-Lin, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica (2024). *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. arXiv: 2403.04132 [cs.AI].

Cho, Kyunghyun, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio (2014). "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*. Doha, Qatar: Association for Computational Linguistics, pp. 103–111.

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1, pp. 37–46.

Cornell Law School (2007). *Rule 49.1 Privacy Protection For Filings Made with the Court*. Legal Rule or Regulation. URL: https://www.law.cornell.edu/rules/frcrmp/rule_49.1.

Csányi, Gergely Márk, Dániel Nagy, Renátó Vági, János Pál Vadász, and Tamás Orosz (2021). "Challenges and Open Problems of Legal Document Anonymization". In: *Symmetry* 13.8, p. 1490.

Cumby, Chad and Rayid Ghani (2011). "A machine learning based system for semi-automatically redacting documents". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 25. 2, pp. 1628–1635.

Dai, Andrew M., Christopher Olah, and Quoc V. Le (2015). "Document Embedding with Paragraph Vectors". In: *CoRR* abs/1507.07998. arXiv: 1507.07998. URL: http://arxiv.org/abs/1507.07998.

DeepSeek-AI et al. (2025). "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning". In: *CoRR* abs/2501.12948. DOI: 10.48550/ARXIV.2501.12948. arXiv: 2501.12948. URL: https://doi.org/10.48550/arXiv.2501.12948.

Dernoncourt, Franck, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits (2017). "De-identification of patient notes with recurrent neural networks". In: *Journal of the American Medical Informatics Association* 24.3, pp. 596–606. ISSN: 1067-5027.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Dolatabadi, Elham, Diana Moyano, Michael Bales, Sofija Spasojevic, Rohan Bhambhoria, Junaid Bhatti, Shyamolima Debnath, Nicholas Hoell, Xin Li, Celine Leng, Sasha Nanda, Jad Saab, Esmat Sahak, Fanny Sie, Sara Uppal, Nirma Khatri Vadlamudi, Antoaneta Vladimirova, Artur Yakimovich, Xiaoxue Yang, Sedef Akinli Kocak, and Angela M Cheung (Sept. 2023). "Using Social Media to Help Understand Patient-Reported Health Outcomes of Post–COVID-19 Condition: Natural Language Processing Approach". In: *J Med Internet Res* 25, e45767. ISSN: 1438-8871. DOI: 10.2196/45767. URL: https://doi.org/10.2196/45767.

Domingo-Ferrer, Josep, David Sánchez, and Alberto Blanco-Justicia (June 2021). "The Limits of Differential Privacy (and Its Misuse in Data Release and Machine Learning)". In: *Communications of the ACM* 64.7, pp. 33–35. ISSN: 0001-0782.

Domingo-Ferrer, Josep, David Sánchez, and Guillem Rufian-Torrell (2013). "Anonymization of nominal data based on semantic marginality". In: *Information Sciences* 242, pp. 35–48. ISSN: 0020-0255.

Domingo-Ferrer, Josep, David Sánchez, and Jordi Soria-Comas (2016). *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Vol. 8. Online: Synthesis Lectures on Information Security, Privacy, and Trust, pp. 1–136.

Domingo-Ferrer, Josep and Vicenc Torra (2003). "Disclosure risk assessment in statistical microdata protection via advanced record linkage". In: *Statistics and Computing* 13.4, pp. 343–354. ISSN: 1573-1375.

Domingo-Ferrer Josep; Torra, Vicenç (2001). "A quantitative comparison of disclosure control methods for microdata". In: *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pp. 111–133.

Dou, Yao, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu (Aug. 2024). "Reducing Privacy Risks in Online Self-Disclosures with Language Models". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 13732–13754. DOI: 10.18653/v1/2024.acl-long.741. URL: https://aclanthology.org/2024.acl-long.741.

Dwork, Cynthia (2006). "Differential Privacy". In: *International Colloquium on Automata, Languages and Programming*. Vol. 4052. Springer, pp. 1–12.

El Emam, Khaled (2013). *Guide to the de-identification of personal health information*. CRC Press.

El-Kassas, Wafaa S, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed (2021). "Automatic text summarization: A comprehensive survey". In: *Expert Systems with Applications* 165, p. 113679. ISSN: 0957-4174.

Elazar, Yanai and Yoav Goldberg (2018). "Adversarial Removal of Demographic Attributes from Text Data". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 11–21.

European Commission (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. Legal Rule or Regulation. URL: `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

European Medicines Agency (2014). *European medicines Agency policy on publication of clinical data for medicinal products for human use (POLICY/0070; EMA/240810/2013)*.

Fabian, MS, Kasneci Gjergji, and WEIKUM Gerhard (2007). "Yago: A core of semantic knowledge unifying wordnet and wikipedia". In: *16th International World Wide Web Conference, WWW*, pp. 697–706.

Fellbaum, Christiane (1998). *WordNet: An electronic lexical database*. MIT press.

Fernandes, Natasha, Mark Dras, and Annabelle McIver (2019). "Generalised Differential Privacy for Text Document Processing". In: *International Conference on Principles of Security and Trust*. Springer, pp. 123–148.

Ferrández, Oscar, Brett R South, Shuying Shen, F Jeffrey Friedlin, Matthew H Samore, and Stéphane M Meystre (2012). "Evaluating current automatic de-identification methods with Veteran's health administration clinical documents". In: *BMC medical research methodology* 12, pp. 1–16.

Fred, Ana L. N. and Anil K. Jain (2003). "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, pp. 128–136. DOI: 10.1109/CVPR.2003.1211462. URL: https://doi.org/10.1109/CVPR.2003.1211462.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023). "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks". In: *CoRR* abs/2303.15056. DOI: 10.48550/ARXIV.2303.15056. arXiv: 2303.15056. URL: https://doi.org/10.48550/arXiv.2303.15056.

Gokaslan, Aaron and Vanya Cohen (2019). *OpenWebText Corpus*. http://Skylion007.github.io/OpenWebTextCorpus.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. (2024). "The llama 3 herd of models". In: *arXiv preprint arXiv:2407.21783*.

Guo, Yuting and Abeed Sarker (May 2023). "SocBERT: A Pretrained Model for Social Media Text". In: *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*. Ed. by Shabnam Tafreshi, Arjun Akula, João Sedoc, Aleksandr Drozd, Anna Rogers, and Anna Rumshisky. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 45–52. DOI: 10.18653/v1/2023.insights-1.5. URL: https://aclanthology.org/2023.insights-1.5/.

Gutiérrez-Batista, Karel, Jesús R Campaña, Maria-Amparo Vila, and Maria J Martin-Bautista (2018). "Building a contextual dimension for OLAP using textual data from social networks". In: *Expert Systems with Applications* 93, pp. 118–133. ISSN: 0957-4174.

Hassan, Fadi, Josep Domingo-Ferrer, and Jordi Soria-Comas (2018). "Anonymization of Unstructured Data Via Named-Entity Recognition". In: *International Conference on Modeling Decisions for Artificial Intelligence*. Mallorca, Spain: Springer, pp. 296–305.

Hassan, Fadi, David Sánchez, and Josep Domingo-Ferrer (2023). "Utility-Preserving Privacy Protection of Textual Documents via Word Embeddings". In: *IEEE Transactions on Knowledge and Data Engineering* 35 (1), pp. 1058–1071. ISSN: 1041-4347.

Health Canada (2019). *Guidance document on public release of clinical information*. URL: https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667.

Huang, Yangsibo, Zhao Song, Danqi Chen, Kai Li, and Sanjeev Arora (2020). "TextHide: Tackling Data Privacy in Language Understanding Tasks". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1368–1382.

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf (2012). *Statistical Disclosure Control*. Vol. 2. New York: Wiley New York.

Jehangir, Basra, Saravanan Radhakrishnan, and Rahul Agarwal (2023). "A survey on Named Entity Recognition — datasets, tools, and methodologies". In: *Natural Language Processing Journal* 3, p. 100017. ISSN: 2949-7191. DOI: https://doi.org/10.1016/j.nlp.2023.100017. URL: https://www.sciencedirect.com/science/article/pii/S2949719123000146.

Jiao, Xiaoqi, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu (2019). "TinyBERT: Distilling BERT for Natural Language Understanding". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pp. 4163–4174.

Jin, Xin, Xia Sun, Jiacheng Chen, and Richard Sutcliffe (2022). "Extracting Drug-drug Interactions from Biomedical Texts using Knowledge Graph Embeddings and Multi-focal Loss". In: *Proceedings of the 31st*

ACM International Conference on Information & Knowledge Management. CIKM '22. Atlanta, GA, USA: Association for Computing Machinery, 884–893. ISBN: 9781450392365. DOI: 10.1145/3511808.3557318. URL: https://doi.org/10.1145/3511808.3557318.

Johnson, Alistair EW, Lucas Bulgarelli, and Tom J Pollard (2020). "Deidentification of free-text medical records using pre-trained bidirectional transformers". In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. Toronto, Ontario, Canada: Association for Computing Machinery, pp. 214–221.

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data". In: *Biometrics* 33, pp. 159–174.

Lastra-Díaz, Juan J., Josu Goikoetxea, Mohamed Ali Hadj Taieb, Ana García-Serrano, Mohamed Ben Aouicha, Eneko Agirre, and David Sánchez (2021). "A large reproducible benchmark of ontology-based methods and word embeddings for word similarity". In: *Inf. Syst.* 96, p. 101636. DOI: 10.1016/J.IS.2020.101636. URL: https://doi.org/10.1016/j.is.2020.101636.

Lebret, Rémi, David Grangier, and Michael Auli (Nov. 2016). "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*. Austin, Texas: Association for Computational Linguistics, pp. 1203–1213.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin.

Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian (2006). "t-closeness: Privacy Beyond k-Anonymity and l-Diversity". In: *IEEE*

23rd International Conference on Data Engineering. IEEE, pp. 106–115. ISBN: 1424408024.

Lima-López, S, E Farré-Maduell, L Gasco Sánchez, A López Rueda, S Frid, A Conesa, X Pastor, and M Krallinger (2023). *CARMEN-I: Anonymization Protocol for Clinical Reports in English*. Tech. rep. Zenodo. DOI: `https://doi.org/10.5281/zenodo.10171682`.

Lison, Pierre, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid (2021). "Anonymisation Models for Text Data: State of the art, Challenges and Future Directions". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 4188–4203.

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

Liu, Zengjian, Buzhou Tang, Xiaolong Wang, and Qingcai Chen (2017). "De-identification of clinical notes via recurrent neural network and conditional random field". In: *Journal of Biomedical Informatics* 75, S34–S42. ISSN: 1532-0464.

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. Association for Computing Machinery, pp. 4768–4777.

Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam (2007). "l-diversity: Privacy Beyond k-Anonymity". In: *ACM Transactions on Knowledge Discovery from Data* 1.1, 3–es. ISSN: 1556-4681.

Mackenzie, Joel M., Rodger Benham, Matthias Petri, Johanne R. Trippas, J. Shane Culpepper, and Alistair Moffat (2020). "CC-News-En:

A Large English News Corpus". In: *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. Ed. by Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux. ACM, pp. 3077–3084. DOI: 10.1145/3340531.3412762. URL: https://doi.org/10.1145/3340531.3412762.

Maduell, E Farre, S Lima-Lopez, SA Frid, A Conesa, E Asensio, A Lopez-Rueda, H Arino, E Calvo, MJ Bertran, MA Marcos, et al. (2023). "CARMEN-I: A resource of anonymized electronic health records in Spanish and Catalan for training and testing NLP tools". In: *URL: https://doi.org/10.13026/bxrx-y344*.

Mamede, Nuno, Jorge Baptista, and Francisco Dias (2016). "Automated anonymization of text documents". In: *IEEE Congress on Evolutionary Computation*. Vancouver, BC, Canada: IEEE, pp. 1287–1294. ISBN: 1509006230.

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky (2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 55–60.

Manzanares-Salor, Benet, David Sánchez, and Pierre Lison (2022). "Automatic Evaluation of Disclosure Risks of Text Anonymization Methods". In: *Privacy in Statistical Databases - International Conference, PSD 2022, Paris, France, September 21-23, 2022, Proceedings*. Ed. by Josep Domingo-Ferrer and Maryline Laurent. Vol. 13463. Lecture Notes in Computer Science. Springer, pp. 157–171. DOI: 10.1007/978-3-031-13945-1_12. URL: https://doi.org/10.1007/978-3-031-13945-1_12.

Manzanares-Salor, Benet, David Sánchez, and Pierre Lison (2024). "Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack". In: *Data Min. Knowl. Discov.*

38.6, pp. 4040–4075. DOI: 10.1007/S10618-024-01066-3. URL: https://doi.org/10.1007/s10618-024-01066-3.

Manzanares-Salor, Benet and David Sánchez (2025). "Enhancing text anonymization via re-identification risk-based explainability". In: *Knowledge-Based Systems* 310, p. 112945. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2024.112945. URL: https://www.sciencedirect.com/science/article/pii/S095070512401579X.

Manzanares-Salor, Benet and David Sánchez (2026). "A comparative analysis, enhancement and evaluation of text anonymization with pre-trained Large Language Models". In: *Expert Systems with Applications* 297, p. 129474. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2025.129474. URL: https://www.sciencedirect.com/science/article/pii/S0957417425030908.

Martínez, Sergio, David Sánchez, and Aida Valls (2012). "Semantic adaptive microaggregation of categorical microdata". In: *Computers & Security* 31.5, pp. 653–672. ISSN: 0167-4048.

Martínez, Sergio, David Sánchez, and Aida Valls (2013). "A semantic framework to protect the privacy of electronic health records with non-numerical attributes". In: *Journal of biomedical informatics* 46.2, pp. 294–303. ISSN: 1532-0464.

Mendels, Omri (2020). *NLP approaches to Data Anonymization*. URL: https://towardsdatascience.com/nlp-approaches-to-data-anonymization-1fb5bde6b929.

Meystre, Stephane M, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore (2010). "Automatic de-identification of textual documents in the electronic health record: a review of recent research". In: *BMC Medical Research Methodology* 10.1, pp. 1–16. ISSN: 1471-2288.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv:1301.3781*.

Mosallanezhad, Ahmadreza, Ghazaleh Beigi, and Huan Liu (2019). "Deep Reinforcement Learning-based Text Anonymization against Private-Attribute Inference". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, pp. 2360–2369.

Mozes, Maximilian and Bennett Kleinberg (2021). "No intruder, no Validity: Evaluation Criteria for Privacy-Preserving Text Anonymization". In: *arXiv:2103.09263*.

Nadeau, David and Satoshi Sekine (2007). "A survey of named entity recognition and classification". In: *Lingvisticae Investigationes* 30.1, pp. 3–26.

Neamatullah, Ishna, Margaret M Douglass, Li-Wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford (2008). "Automated de-identification of free-text medical records". In: *BMC Medical Informatics and Decision Making* 8.1, pp. 1–17. ISSN: 1472-6947.

Nin, Jordi, Javier Herranz, and Vicenç Torra (2007). "On method-specific record linkage for risk assessment". In: *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*. UNECE, pp. 1–12.

Nyffenegger, Alex, Matthias Stürmer, and Joel Niklaus (2023). "Anonymity at Risk? Assessing Re-Identification Capabilities of Large Language Models". In: *CoRR* abs/2308.11103. DOI: 10.48550/ARXIV.2308.11103. arXiv: 2308.11103. URL: https://doi.org/10.48550/arXiv.2308.11103.

Olstad, Annika Willoch, Anthi Papadopoulou, and Pierre Lison (2023). "Generation of Replacement Options in Text Sanitization". In: *Proceedings of the 24th Nordic Conference on Computational Linguistics, NoDaLiDa 2023, Tórshavn, Faroe Islands, May 22-24, 2023*. Ed. by Tanel Alumäe and Mark Fishel. University of Tartu Library, pp. 292–300. URL: https://aclanthology.org/2023.nodalida-1.30.

OpenAI (Mar. 2023). "GPT-4 Technical Report". In: *arXiv e-prints*, arXiv:2303.08774. arXiv: 2303.08774 [cs.CL].

Papadopoulou, Anthi, Pierre Lison, Lilja Øvrelid, and Ildikó Pilán (June 2022). "Bootstrapping Text Anonymization Models with Distant Supervision". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4477–4487.

Pappagari, Raghavendra, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak (n.d.). "Hierarchical transformers for long document classification". In: *IEEE Automatic Speech Recognition and Understanding Workshop*. IEEE, pp. 838–844. ISBN: 1728103061.

Patsakis, Constantinos and Nikolaos Lykousas (2023). "Man vs the machine in the struggle for effective text anonymisation in the age of large language models". In: *Scientific Reports* 13.1, p. 16026.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "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)*. New Orleans, Louisiana, USA: Association for Computational Linguistics, pp. 2227–2237.

Pilán, Ildikó, Pierre Lison, Lilja Ørelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet (2022). "The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization". In: *Computational Linguistics* 48.8, pp. 1053–1101. ISSN: 0891-2017.

Pilán, Ildikó, Benet Manzanares-Salor, David Sánchez, and Pierre Lison (2025). "Truthful text sanitization guided by inference attacks". In: *Applied Soft Computing* 185, p. 114013. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2025.114013. URL: https://www.sciencedirect.com/science/article/pii/S1568494625013262.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of Machine Learning Research* 21.140, pp. 5485–5551.

Rasmy, Laila, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi (2021). "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction". In: *npj Digit. Medicine* 4. DOI: 10.1038/S41746-021-00455-Y. URL: https://doi.org/10.1038/s41746-021-00455-y.

Razavi, Amirhossein, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri (2025). "Benchmarking prompt sensitivity in large language models". en. In: *Lecture Notes in Computer Science*. Lecture notes in computer science. Cham: Springer Nature Switzerland, pp. 303–313.

Reddy, Sravana and Kevin Knight (2016). "Obfuscating Gender in Social Media Writing". In: *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas, USA: Association for Computational Linguistics, pp. 17–26.

Reid, Machel, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen

Spencer, Eren Sezener, and et al. (2024). "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context". In: *CoRR* abs/2403.05530. DOI: 10.48550/ARXIV.2403.05530. arXiv: 2403.05530. URL: https://doi.org/10.48550/arXiv.2403.05530.

Reimers, Nils and Iryna Gurevych (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Association for Computational Linguistics, pp. 3980–3990. DOI: 10.18653/V1/D19-1410. URL: https://doi.org/10.18653/v1/D19-1410.

Resnik, Philip (1995). "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. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., pp. 448–453. ISBN: 1558603638.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "Why Should I Trust You?: Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, pp. 1135–1144. ISBN: 9781450342322.

Rivas, Ryan and Vagelis Hristidis (2021). "Effective social post classifiers on top of search interfaces". In: *Data Mining and Knowledge Discovery* 35.5, pp. 1809–1829.

Samarati, Pierangela (2001). "Protecting respondents identities in microdata release". In: *IEEE Transactions on Knowledge and Data Engineering* 13.6, pp. 1010–1027.

Sánchez, David and Montserrat Batet (2016). "C-sanitized: A privacy model for document redaction and sanitization". In: *Journal of the Association for Information Science and Technology* 67.1, pp. 148–163. ISSN: 2330-1635.

Sánchez, David and Montserrat Batet (2017). "Toward sensitive document release with privacy guarantees". In: *Engineering Applications of Artificial Intelligence* 59, pp. 23–34. ISSN: 0952-1976.

Sánchez, David, Montserrat Batet, and Alexandre Viejo (2013a). "Automatic general-purpose sanitization of textual documents". In: *IEEE Transactions on Information Forensics and Security* 8.6, pp. 853–862. ISSN: 1556-6013.

Sánchez, David, Montserrat Batet, and Alexandre Viejo (2013b). "Minimizing the disclosure risk of semantic correlations in document sanitization". In: *Information Sciences* 249, pp. 110–123. ISSN: 0020-0255.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv:1910.01108*.

Seco, Nuno, Tony Veale, and Jer Hayes (2004). "An Intrinsic Information Content Metric for Semantic Similarity in WordNet". In: *Proceedings of the 16th European Conference on Artificial Intelligence*. ECAI'04. Valencia, Spain: IOS Press, pp. 1089–1090. ISBN: 9781586034528.

Shannon, Claude Elwood (1948). "A mathematical theory of communication". In: *The Bell system technical journal* 27.3, pp. 379–423. ISSN: 0005-8580.

Singhal, Shreya, Andres Felipe Zambrano, Maciej Pankiewicz, Xiner Liu, Chelsea Porter, and Ryan S Baker (2024). "De-Identifying Student Personally Identifying Information with GPT-4". In: *Proceedings of the 17th International Conference on Educational Data Mining*, pp. 559–565.

Soria-Comas, Jordi and Josep Domingo-Ferrer (2012). "Probabilistic k-anonymity through microaggregation and data swapping". In: *2012 IEEE International Conference on Fuzzy Systems*, pp. 1–8.

Staddon, Jessica, Philippe Golle, and Bryce Zimny (2007). "Web-Based Inference Detection". In: *USENIX Security Symposium*. Boston, MA, USA: Association for Computing Machinery, pp. 1–16.

Subakti, Alvin, Hendri Murfi, and Nora Hariadi (2022). "The performance of BERT as data representation of text clustering". In: *Journal of Big Data* 9.1, pp. 1–21.

Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang (2019). "How to Fine-Tune BERT for Text Classification?" In: *China National Conference on Chinese Computational Linguistics*. Springer, pp. 194–206.

Sun, Xiaofei, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang (2023). "Text classification via large language models". In: *arXiv preprint arXiv:2305.08377*.

Szarvas, György, Richárd Farkas, and Róbert Busa-Fekete (2007). "State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework". In: *Journal of the American Medical Informatics Association* 14.5, pp. 574–580. ISSN: 1527-974X.

Tan, Eliza Janel L., Keith Alexandre L. Ramos, Maria Elisha Kaye B. Nazario, Steven Vincent D. Lim, and Shirley B. Chu (2024). "AI to the Test: Measuring ChatGPT's Objective Accuracy in the SATs in Comparison to Human Performance". In: *48th IEEE Annual Computers, Software, and Applications Conference, COMPSAC 2024, Osaka, Japan, July 2-4, 2024*. Ed. by Hossain Shahriar, Hiroyuki Ohsaki, Moushumi Sharmin, Dave Towey, A. K. M. Jahangir Alam Majumder, Yoshiaki Hori, Ji-Jiang Yang, Michiharu Takemoto, Nazmus Sakib, Ryohei Banno, and Sheikh Iqbal Ahamed. IEEE, pp. 153–158. DOI: 10.1109/COMPSAC61105.2024.00031. URL: https://doi.org/10.1109/COMPSAC61105.2024.00031.

Torra, Vicenç, John M Abowd, and Josep Domingo-Ferrer (2006). "Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment". In: *Privacy in Statistical Databases*. Springer, pp. 233–242.

Torra, Vicenç and Klara Stokes (2012). "A formalization of record linkage and its application to data protection". In: *International Journal*

of Uncertainty, Fuzziness and Knowledge-Based Systems 20.06, pp. 907–919. ISSN: 0218-4885.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom (2023). "Llama 2: Open Foundation and Fine-Tuned Chat Models". In: *CoRR* abs/2307.09288. DOI: 10.48550/ARXIV.2307.09288. arXiv: 2307.09288. URL: https://doi.org/10.48550/arXiv.2307.09288.

Trinh, Trieu H. and Quoc V. Le (2018). "A Simple Method for Commonsense Reasoning". In: *CoRR* abs/1806.02847. arXiv: 1806.02847. URL: http://arxiv.org/abs/1806.02847.

Turney, Peter D. and Patrick Pantel (2010). "From Frequency to Meaning: Vector Space Models of Semantics". In: *J. Artif. Intell. Res.* 37, pp. 141–188. DOI: 10.1613/JAIR.2934. URL: https://doi.org/10.1613/jair.2934.

U.S. Department of Labor (2004). *The Health Insurance Portability and Accountability Act*. Legal Rule or Regulation. URL: https://www.dol.gov/agencies/ebsa/laws-and-regulations/laws/hipaa.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, łukasz Kaiser, and Illia Polosukhin (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Long Beach, CA, USA: Neural Information Processing Systems Foundation, pp. 5998–6008.

Volodina, Elena, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beáta Megyesi (2020). "Towards Privacy by Design in Learner Corpora Research: A Case of On-the-fly Pseudonymization of Swedish Learner Essays". In: *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*. Ed. by Donia Scott, Núria Bel, and Chengqing Zong. International Committee on Computational Linguistics, pp. 357–369. DOI: 10.18653/V1/2020.COLING-MAIN.32. URL: https://doi.org/10.18653/v1/2020.coling-main.32.

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou (n.d.). "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers". In: *Advances in Neural Information Processing Systems*. Vol. 33. Association for Computing Machinery, pp. 5776–5788.

Weischedel, Ralph, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue (2011). "OntoNotes: A large training corpus for enhanced processing". In: *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. New York: Springer. Chap. 1, pp. 54–63.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, and Morgan Funtowicz (2020). "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, pp. 38–45.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. (2016). "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv:1609.08144*.

Xu, Qiongkai, Lizhen Qu, Chenchen Xu, and Ran Cui (2019). "Privacy-Aware Text Rewriting". In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 247–257.

Yang, Hui and Jonathan M Garibaldi (2015). "Automatic detection of protected health information from clinic narratives". In: *Journal of Biomedical Informatics* 58, S30–S38. ISSN: 1532-0464.

Yogarajan, Vithya, Michael Mayo, and Bernhard Pfahringer (2018). "A survey of automatic de-identification of longitudinal clinical narratives". In: *arXiv:1810.06765*.

Zhao, Yabing, Xun Xu, and Mingshu Wang (2019). "Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews". In: *International Journal of Hospitality Management* 76, pp. 111–121. ISSN: 0278-4319.

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "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*. Santiago, Chile: IEEE, pp. 19–27.



UNIVERSITAT ROVIRA i VIRGILI