Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 257 (2025) 888–895

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Emerging Data and Industry (EDI40)
April 22-24, 2025, Patras, Greece

Striking the Balance: Generalization vs. Memorization in Anonymization and De-anonymization through LLMs

Méliane Angèle Kokmel[a,*], Antragama Ewa Abbas[b], Igor Tchappi[b]

[a]DCS, FSTM, University of Luxembourg, 2 Av. de l'Universite, 4365, Esch-sur-Alzette, Luxembourg
[b]FINATRAX, SnT, University of Luxembourg, 2 Av. de l'Universite, 4365, Esch-sur-Alzette, Luxembourg

Abstract

Text anonymization aims to enable the secure sharing of information between parties. One of the main challenges in data anonymization is achieving a balance between ensuring data privacy and maintaining data utility. To address these challenges, recent studies have explored the use of Large Language Models (LLMs), which have shown improved performance on datasets from Europe. Based on these findings, this paper aims to create a dataset from less explored parts of the world, specifically Africa, to assess the relevance of LLMs on diverse datasets and to discuss the generalization of the results. Additionally, this paper proposes an evaluation framework for assessing various anonymization techniques, including those utilizing LLMs. The performance of these techniques is assessed using several metrics, such as BERTScore for semantic evaluation and Information Loss for utility preservation.

Keywords: Anonymization, de-anonymization, LLMs, Generalization, Diversity in AI

1. Introduction

The volume and availability of data through big data has led to a rapid increase in technological advancements bringing benefits to society in different sectors. However, the massive exploitation of generated data can pose a serious privacy issue. Therefore, data sharing must be approached with great caution to prevent the disclosure of sensitive contents [3]. To this end, GDPR (General Data Protection Regulation) was proposed. GDPR provides a framework for data sharing and recommends anonymization prior to data sharing [5]. Various anonymization methods have been proposed in the literature, and most of them mainly focus on structured data [15]. Regarding textual data, the main methods available in the literature are highly based on Named Entity Recognition (NER) [22]. These methods are limited to identifying predefined named entities, leaving other potentially identifiable elements such as nouns

* Corresponding author.
E-mail address: angelemelagne@gmail.com

1877-0509
10.1016/j.procs.2025.03.114

and phrases unprotected, which can result in inadequately anonymized outputs [22]. Another significant challenge is balancing privacy and utility in text documents [7]. In recent years, significant advances in Generative AI (GenAI) have led to the development of large language models (LLMs) [26]. LLMs are capable of generating text, images, audio, etc. and can perform a wide range of tasks via prompts. LLMs are also capable of performing tasks that they were not explicitly trained to perform, which leads to their widespread adoption in various domains [6, 25]. For example, LLMs are used in applications such as time series forecasting, sentiment analysis, and even in data anonymization and de-anonymization [2]. In data anonymization, de-anonymization problem, the contextual understanding of LLMs represents a significant advantage leading to more effective and precise anonymization of textual data [2]. In fact, Patsakis and Lykousas [18] evaluated several state-of-the-art anonymization models such as Textwash [14], and LLM and compared them with combination of Textwash and GPT [18], highlighting the improved performance of the latter. Moreover, regarding de-anonymization, LLMs also showed a better performance, even compared to humans, in revealing the identity of individuals from anonymized texts of celebrities Patsakis and Lykousas [18]. To evaluate their work, Patsakis and Lykousas [18] used a dataset of celebrities from the UK [14]. One open question is therefore to investigate the generalization of their results. In fact, as LLMs are mainly trained on Western-centric data (due to insufficient data from other sources), it may be unclear whether the performance of LLMs in de-anonymization of celebrities comes from the ability of LLM to retain and recall the specific data they were trained on, or from their capacity to generalize and de-anonymize. To this end, and inspired by their work, this paper aims to extend the existing body of knowledge by evaluating the performance of existing anonymization models, as well as several LLM-based anonymization techniques on diverse dataset i.e., a dataset coming from less explored regions such as Africa and a dataset from UK. The motivation stems from recognizing that most available data are western-centric, potentially limiting the generalizability of findings. In addition, diversity and inclusivity are particularly pertinent given the multicultural and global nature of data and the need to protect sensitive information in diverse regions of the world. This paper makes the following three key contributions:

- We evaluate existing anonymization techniques using a dataset from a less explored region, specifically Africa.
- We analyze the effectiveness of various primarily open-source Large Language Models (LLMs) in the context of data anonymization.
- We investigate the anonymization strength achieved by combining LLMs with other anonymization methods.

The remainder of the paper is organized as follows: Section 2 presents the related works, Section 3 outlines the methodology and the dataset used. Section 4 describes the experiment and the evaluation, and finally, Section 5 draws the conclusion.

2. Related Works

Real-world data from various domains is generated every day. They may be associated with individuals or entity information giving them a sensitive character for which sharing must be done with caution in order to protect individual's privacy. Ensuring secure data sharing has been the cornerstone of GDPR since its creation [27]. As stated above, according to GDPR, anonymization is one of the best approaches for secure sharing of data. Once data have passed the test of anonymization, it aims to do not be traced back to an individual. It may be shared with a third party. Usually, the process of anonymization consists of two main steps: (i) identifying disclosive pieces of data, and (ii) masking them. Over the years, several privacy-preserving data anonymization techniques for structured data have been proposed in the literature. Majeed and Lee [15] present a detailed survey of anonymization methods for social network and relational data. The techniques are categorized, along with the metrics used to assess their effectiveness. Additionally, the author discusses potential attacks on sanitized published data and outlines the various actors involved in the anonymization process. Regarding unstructured data, there are also several approaches presented in literature. Most of the techniques for textual data are based on machine learning and dictionary based technique Perez-Lainez et al. [20]. Scrub was introduced by Sweeney [24] it relies on dictionary and pattern-based to anonymize entities. It was designed to anonymize patients medical records. Hassan et al. [9] also presented a method using word embedding for detecting sensitive information in order to achieve automatic text anonymization. All entities in the document are represented as word vectors. By this, an entity is protected by removing all the other entities (person or organization) in the document having the same similarity vector to that of the entity in a document. Kleinberg et al. [13]

developed NETANOS, an open-source system that uses machine-learning to perform NER. It was designed to foster researchers in their raw data publishing. It achieves anonymization by pinpointing contextual information (e.g. Luxembourg) and replacing them with context-preserving category label (e.g.: Location_1). Perez-Lainez et al. [19] as well presented a system called ANONYMITEXT, aiming to anonymize medical records. His method relies on dictionary induction technique to achieve anonymization. After anonymization, a business expert provides a review by rejecting or approving the output text result and leaving a final feedback to the system. Textwash is an automated open-source text anonymizer introduced by Kleinberg et al. [14]. It was introduced to address the different limitations of NER-based techniques. Based on a machine learning-based methods from fine-tuned BERT model, it uses linguistic patterns extracted from contextual information to predict whether individual words and phrases contain sensitive information. Textwash uses more than 11 pre-defined tags to replace identifiable tokens. There are: PERSON_FIRSTNAME, PERSON_LASTNAME, OCCUPATION, LOCATION, TIME, ORGANIZATION, DATE, ADDRESS, PHONE_NUMBER, EMAIL_ADDRESS, OTHER_IDENTIFYING_ATTRIBUTE, NONE. It is capable to preserve the semantic properties of anonymize texts while removing Personally Sensitive Information (PSI). For example, given a sentence: "Marie weintz is a 20 years old student at the University of Luxembourg"., Textwash anonymization output will be "FIRSTNAME_1 LASTNAME_1 is a NUMERIC OCCUPATION at the LOCATION". Several other techniques based on NER were also presented [21] [12] [8]. Additionally, several software packages are available for performing NER tasks, including spaCy [10] and Stanford NER[16].

3. Anonymization - De-anonymization Framework

This section presents the application of anonymization techniques to two distinct datasets. It outlines the steps involved in anonymization and de-anonymization and describes the datasets used. The use of two datasets is needed to evaluate the capabilities of the models and the behavior of LLMs in different cultural contexts.

3.1. Anonymization - De-anonymization workflow

This section outlines the process of anonymizing and de-anonymizing datasets, highlighting each critical step involved. To this end, two anonymization workflows will be used, namely, classical anonymization and proposed anonymization. First, the classical anonymization process is a single-step procedure. The original dataset is anonymized using either an algorithm such as Texwash, spaCy or a LLM, and presented in Figure 1. Second, the proposed anonymization workflow consisting of three major steps illustrated in Figure 3 is detailed below:

1. First, algorithms such as Textwash or spaCy are used to anonymize the original dataset
2. Next, all tags/labels inserted by the algorithms in the first stage are removed from the anonymized text.
3. Finally, LLMs such as GPT or LlaMA are prompted to further anonymize the texts from the second stage.

The anonymization performed by the LLMs is based on zero-shot learning. Regarding the de-anonymization, the motivated intruder test is used to assess the privacy effectiveness of an anonymization model or technique. It involves conducting a controlled attack seeking to reverse the anonymization process by revealing hidden information in the anonymized data. Motivated intruder test is therefore used as an assessment to evaluate how well anonymization techniques protect sensitive data from re-identification. In this work, the motivated intruder test is conducted always using GPT-3 as shown on Figure 2. Following the attack on the dataset, the texts are classified into two groups: correctly de-anonymized where the attack is successful and incorrectly de-anonymized texts where it fails.

3.2. Dataset

The paper relies on two datasets for the evaluation of the methodology. The first dataset is drawn from the literature, specifically sourced from the study of Kleinberg et al. [14]. It includes 1080 texts descriptions on 20 well-known public figures from the UK. Each description possesses both original and anonymized texts by Textwash. This dataset was compiled by recruiting 200 individuals in the UK to write about a randomly selected set of celebrities. It serves as a benchmark for the comparison in this paper. The second dataset, centered on prominent West African figures, was

Fig. 1: Classical Anonymization Workflow

Fig. 2: De-anonymization Workflow

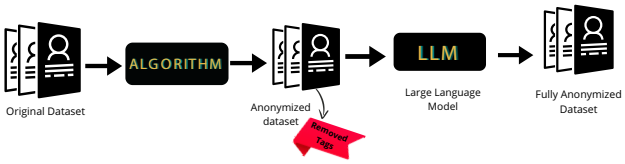

Fig. 3: Proposed Anonymization Workflow

developed as the first contribution of this work. Its creation followed the methodological framework of Kleinberg et al. [14], ensuring consistency and comparability with existing approaches. The dataset consists of 752 text descriptions of 20 celebrities from Ivory Coast and Nigeria. It was built through an online survey in which 104 participants described 20 national celebrities.

4. Experiments and Evaluation

This section presents the different experiments, their evaluations and results. This experiment examines two LLMs: GPT-3.5, a fine-tuned version of GPT-3 introduced by OpenAI in 2020, and LlaMA, an open-source LLM by Meta AI known for its advanced capabilities and accessibility.

4.1. STUDY-1: Anonymization of African and UK dataset

This study aims to evaluate the performance of three anonymization models namely spaCy, Textwash, and "Textwash+GPT" [18] on both African and UK datasets. The anonymization of spaCy and Textwash was conducted according to the workflow presented in Figure 1 while the anonymization of the combination model "Textwash+GPT" was conducted according to Figure 3.

4.1.1. Motivated Intruder Test

GPT was responsible for performing the motivated intruder test (de-anonymization) as presented in Figure 2. The performance of each anonymization method across both datasets is visually represented in Figure 4, offering a comparative analysis of their effectiveness in each case. The spaCy model performed the least effectively on the African dataset, with GPT successfully de-anonymizing 180 texts out of 752 (23%) against 798 texts (73,88%) on the western dataset. Following spaCy, the Textwash anonymizer had 26 texts (3,45%) successfully de-anonymized by GPT in the African dataset against 668 texts (61,85%) in the UK dataset. Finally, the "Textwash+GPT" method demonstrated the best performance, with only 1.46% of African texts successfully de-anonymized, compared to 49.35% for the UK texts (533 out of 1,080).

4.1.2. Discussion related to the performance of GPT in de-anonymization

After the de-anonymization of the datasets, all the correctly de-anonymized texts were collected. GPT was then prompted to return from the anonymized version of the text, all the sensitive tokens that could lead to the identification of the person being described. A detailed analysis of the de-anonymized texts identified specific tokens and contextual information that led to breaches in anonymization. This analysis will focus mainly on the African dataset. The following are a few examples selected from the many identified cases.

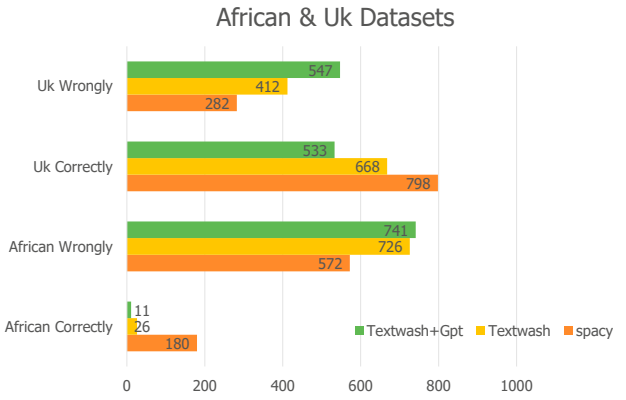

Fig. 4: Motivated intruder test result per method on the African and Uk Datasets

spaCy. : The failure to capture sensitive tokens such as names and titles were the primary cause of de-anonymization. In many instances, spaCy did not anonymize names of well-known individuals, leading to easy re-identification. For example, in the case of "Atiku Abubakar", spaCy failed to anonymize tokens like "politician", "Vice President" and "Presidential ambitions" which GPT used to re-identify the individual.

Textwash. : While more effective in anonymizing names, Textwash often missed key terms and roles that GPT used for re-identification. For example, when anonymizing "Funke Akindele", Textwash anonymized the name but missed terms like "actress", "producer", and "Nollywood", which were sufficient for re-identification. This is inline with the results of Kleinberg et al. [14] saying that Textwash has a poor performance on concealing job title.

"Textwash+GPT". : The combination of method generally replaced specific names with generic terms but occasionally failed to hide detailed contextual information, leading to de-anonymization. For instance, while anonymizing "Godwin Emefiele", the Textwash+GPT method replaced the name with "an economist and banker" but did not hide detailed contextual clues like "governor of a countrys central bank" and "implemented economic policies" leading to the successful de-anonymization by GPT. In general, the previous highlights the difficulty in recognizing certain sensitive African entities. In fact, The African dataset may contain cultural and linguistic nuances that spaCy and textwash components are not well-equipped to handle, leading to a higher failure rate in accurately anonymizing sensitive information [23]. Moreover, the contexts in which sensitive information appears in the African dataset might differ from those in the UK dataset. Regarding the names, if the description is about Atiku Abubakar, the de-anonymization might incorrectly yield a name like "Joe Biden". Another example is the famous Nigerian international football player "Victor Osimhen". In this case, the de-anonymization might incorrectly return names like "David Beckham" or "Cristiano Ronaldo" other well-known international football players rather than African football celebrities names such as "Samuel Eto'o" or "Didier Drogba", who were also great players. This argument is supported by the analysis of names generated during the motivated intruder test on spaCy-anonymized texts. Only 25% of the names returned were of African origin, including those that were incorrectly de-anonymized (successful anonymization). The majority of the names were from non-African regions, predominantly Europe and America, with a smaller proportion from Asia. The previous could be close to the challenge exposed by Ahidi Elisante Lukwaro et al. [1] when applying lemmatization to a text with some swahili language contents. According to their result, the traditional models struggle with African datasets which is illustrated by its failure to anonymize terms and contexts specific to African culture.

4.1.3. Performance Comparison of De-Anonymization Techniques

Performance comparison between the African and UK datasets reveals notable disparities. The de-anonymization process demonstrated a significantly higher level of success when applied to the UK dataset, compared to the African dataset by a considerable margin, as presented in Figure 4. This disparity in performance could suggest that GPT, in this case, might be drawing more effectively from its training data when dealing with the UK dataset, potentially due to its familiarity with the data sources or cultural context associated with the UK. In fact [4], LLMs are trained on

datasets derived from a wide array of online content, including books, academic articles, websites, and other publicly available text sources [4] dominated by western countries. Therefore, it is possible that the model's ability to correctly de-anonymize is more influenced by the prominence or availability of information related to the UK celebrities in the training corpus, rather than reflecting its true capacity for general de-anonymization. These findings suggest a potential limitation in GPT's ability to generalize for de-anonymization, as it performs well on data closely aligned with its training set but struggles with more diverse or underrepresented datasets. These findings may point to a potential limitation of GPT's ability to generalize for de-anonymization, as it performs well on data closely aligned with its training set but struggles with more diverse or underrepresented datasets like the African one. Further investigation is needed to assess GPT's ability to generalize across different cultural contexts and datasets.

4.2. STUDY-2

This study evaluates the performance of various large language models (LLMs) in text anonymization across datasets from diverse origins. The African and Western datasets are anonymized sequentially with GPT-3 and LlaMA-3 to analyze the LLMs's behavior as presented in Figure 3. The result from the motivated intruder test by GPT is presented in Graph 5. The anonymization of the African and UK datasets using GPT and LlaMA shows varying

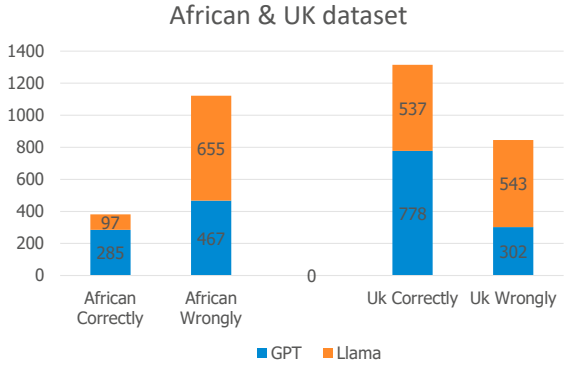

Fig. 5: African and UK dataset anonymized with GPT and LlaMA

levels of performance. A lower de-anonymization rate indicates a stronger anonymization capability of the method employed. The first bar in Figure 5 represents the correctly de-anonymized texts from the African dataset, highlighting that both LlaMA and GPT have difficulty to correctly de-anonymize African dataset. For example, the correctly de-anonymized African texts total 285 (37.89%) and 97 (12.89%) for GPT and LlaMA, respectively, compared to 778 (72.03%) and 537 (49%) for the UK dataset. Similar to Study-1, analyzing the texts incorrectly de-anonymized by both LLMs is crucial to better understand the disparity in results. During the anonymization process, GPT and LlaMA replaced celebrity names with "the individual" and substituted pronouns like "he/she" with "they/them" often used by non-binary individuals. Moreover, the anonymized texts lack tags or labels, resulting in a more readable and fluid output while concealing PIIs. However, most of the incorrectly de-anonymized texts from the African dataset do not result from strongly concealed personally identifiable information (PIIs). Analysis reveals that many PIIs and sensitive details that could lead to the re-identification of individuals were not adequately hidden by the LLMs. The LLMs struggled significantly with anonymization, leaving many named entities (NEs), such as dates, locations, and job positions, untouched and unanonymized in the texts. Despite the leakage of sensitive information during the anonymization process, GPT still failed to identify the individuals described in the African dataset. It was unable to reveal the identities of the anonymized celebrities. In contrast, GPT successfully identified individuals in the UK dataset, even though some of the same named entities existed in the anonymized texts.

4.3. STUDY-3

This study evaluates the performance of the proposed method in text anonymization. The objective is to identify the most effective models for achieving a superior balance between data privacy and utility. By systematically testing

different combination of LLMs associated with GPT and LlaMA, the study aims to determine which provide the most robust anonymization for text data. Due to the unsatisfactory performance of LLMs on African datasets, the experiment is conducted solely on Western dataset to achieve more accurate results while testing different approaches for a stronger anonymization model. The new anonymization technique assess the combination of "Textwash+LlaMA" represented as "T+LlaMA", "Textwash+GPT" as "T+GPT", "Spacy+LlaMA" as "S+LlaMA", and "Spacy+GPT" as "S+GPT". All these methods will undergo privacy testing via de-anonymization test and then the utility testing, which will assess information loss, the proportion of anonymized text, semantic similarity, and readability. Table 1 presents the GPT motivated intruder test results, and Table 2 shows the normalized scores for each method.

Table 1: Aggregated results for the different methods

Motivated Intruder Test						
OUTPUT	GPT	LlaMA	T+LlaMA	T+GPT	S+LlaMA	S+GPT
Performance Rate	302 (27.66%)	543 (50.27%)	609 (56.38%)	547 (50.64%)	475 (43.98%)	455 (42.12%)

Table 2: Normalized Score per Method

Calculated Normalized Score per Method				
Methods	**spaCy + GPT**	**spaCy + LLaMA**	**Textwash + GPT**	**Textwash + LLaMA**
Information Loss Scores	1.0	0.0	1.0	0.0
Proportion Anonymized Text	0.96526	0.0	1.0	0.77502
Semantic Similarity	0.09433	1.0	0.08962	0.0
Readability	0.00265	1.0	0.66896	0.03813
Final Score	0.51556	0.5	0.68964	0.20328

T

The combination of traditional anonymization methods with LLMs presented better results. For example, the "spaCy+LlaMA" combination performed best, preserving semantic content with the highest F1 score, indicating a balanced precision and recall. "Textwash+GPT" also showed strong performance, particularly in precision. The evaluation of the proposed models against the different metrics [11] reveal that "Textwash+GPT" provide the most robust anonymization. By holding the highest composite score, it proves to be the best method in balancing text utility and privacy preserving in UK dataset. This is inline with literature as Wang et al. [28] and Papernot et al. [17] have showed that, combining machine learning techniques (ML) with traditional methods can lead to significant improvement in data privacy. They have proved that combining traditional and modern techniques can achieve greater performance. However, the generalization of the previous needs further investigation.

5. Conclusion

In conclusion, this paper developed a method for anonymizing and de-anonymizing textual data from diverse sources and assessed the generalizability of GenAI. The paper founds that traditional methods like spaCy and Textwash struggled with non-Western dataset, especially African text. In the same vein, the paper found that GPT had difficulty re-identifying local celebrities from African dataset compared to UK dataset. Future work will address biases in data collection and refine anonymization methods to better handle diverse linguistic and cultural contexts.

6. Acknowledgments

This research was funded in whole by the Luxembourg National Research Fund (FNR) and PayPal, PEARL grant reference 13342933/Gilbert Fridgen. For the purpose of open access, and in fulfillment of the obligations arising from the grant agreement, the author has applied a Creative Commons Attribution 4.0 International (CC BY 4.0) license to any Author Accepted Manuscript version arising from this submission.

References

[1] Elia Ahidi Elisante Lukwaro, Khamisi Kalegele, and Devotha G Nyambo. A review on nlp techniques and associated challenges in extracting features from education data. *International Journal of Computing and Digital Systems*, 16(1):961–979, 2024.

[2] Dimitris Asimopoulos, Ilias Siniosoglou, Vasileios Argyriou, Sotirios K Goudos, Konstantinos E Psannis, Nikoleta Karditsioti, Theocharis Saoulidis, and Panagiotis Sarigiannidis. Evaluating the efficacy of ai techniques in textual anonymization: A comparative study. *arXiv preprint arXiv:2405.06709*, 2024.

[3] United Nations. General Assembly. *Universal declaration of human rights*, volume 3381. Department of State, United States of America, 1949.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Josep Domingo-Ferrer. Personal big data, gdpr and anonymization. In *Flexible Query Answering Systems: 13th International Conference, FQAS 2019, Amantea, Italy, July 2–5, 2019, Proceedings 13*, pages 7–10. Springer, 2019.

[6] Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. A bibliometric review of large language models research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology*, 15(5):1–25, 2024.

[7] Andrea Gadotti, Luc Rocher, Florimond Houssiau, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*, 10(29):eadn7053, 2024.

[8] Fadi Hassan, Josep Domingo-Ferrer, and Jordi Soria-Comas. Anonymization of unstructured data via named-entity recognition. In *Modeling Decisions for Artificial Intelligence: 15th Int. Conf., MDAI 2018, Mallorca, Spain, October 15–18, 2018, Proc. 15*, pages 296–305. Springer, 2018.

[9] Fadi Hassan, David Sánchez, Jordi Soria-Comas, and Josep Domingo-Ferrer. Automatic anonymization of textual documents: detecting sensitive information via word embeddings. In *2019 18th IEEE Int. Conf. On Trust, Security And Privacy In Computing And Communications/13th IEEE Int. Conf. On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 358–365. IEEE, 2019.

[10] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. *online*, 2020.

[11] Joris Hulstijn, Igor Tchappi, Amro Najjar, and Reyhan Aydoğan. Metrics for evaluating explainable recommender systems. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 212–230. Springer, 2023.

[12] Bennett Kleinberg and Maximilian Mozes. Web-based text anonymization with node. js: Introducing netanos (named entity-based text anonymization for open science). *Journal of Open Source Software*, 2(14):293, 2017.

[13] Bennett Kleinberg, Maximilian Mozes, Yaloe van der Toolen, et al. Netanos-named entity-based text anonymization for open science. *journal*, 2017.

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

[15] Abdul Majeed and Sungchang Lee. Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE access*, 9:8512–8545, 2020.

[16] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[17] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. Sok: Security and privacy in machine learning. In *2018 IEEE European symposium on security and privacy (EuroS&P)*, pages 399–414. IEEE, 2018.

[18] Constantinos Patsakis and Nikolaos Lykousas. Man vs the machine in the struggle for effective text anonymisation in the age of large language models. *Scientific Reports*, 13(1):16026, 2023.

[19] Rebeca Perez-Lainez, Ana Iglesias, and César de Pablo-Sánchez. Anonimytext: Anonimization of unstructured documents. In *International Conference on Knowledge Discovery and Information Retrieval*, 2009. URL https://api.semanticscholar.org/CorpusID:33934217.

[20] Rebeca Perez-Lainez, Ana Iglesias, and Cesar de Pablo-Sanchez. Anonimytext: Anonimization of unstructured documents. In *International Conference on Knowledge Discovery and Information Retrieval*, volume 2, pages 284–287. SCITEPRESS, 2009.

[21] Rahul Sharnagat. Named entity recognition: A literature survey. *Center For Indian Language Technology*, pages 1–27, 2014.

[22] Peng Sun, Xuezhen Yang, Xiaobing Zhao, and Zhijuan Wang. An overview of named entity recognition. In *2018 International Conference on Asian Language Processing (IALP)*, pages 273–278. IEEE, 2018.

[23] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, 2021.

[24] Latanya Sweeney. Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association, 1996.

[25] Melissa Tessa, Sarah Abchiche, Yves Claude Ferstler, Igor Tchappi, Karima Benatchba, and Amro Najjar. Enhancing explanaibility in ai: Food recommender system use case. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, pages 395–397, 2023.

[26] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[27] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[28] Yanling Wang, Qian Wang, Lingchen Zhao, and Cong Wang. Differential privacy in deep learning: Privacy and beyond. *Future Generation Computer Systems*, 148:408–424, 2023.