Beyond the Parameters: Measuring Actual Privacy in Obfuscated Texts

Francesco Luigi, De Faveri[1], Guglielmo, Faggioli[1] and Nicola, Ferro[1]

[1]*Department of Information Engineering, University of Padova, Padova, Italy*

Abstract

Evaluating privacy provided by obfuscation mechanisms remains an open problem in the research community. Especially for textual data, in Natural Language Processing (NLP) and Information Retrieval (IR) tasks, privacy guarantees are measured by analyzing the hyper-parameters of a mechanism, e.g., the privacy budget ε in Differential Privacy (DP), and the impact of these on the performances. However, considering only the privacy parameters is not enough to understand the actual level of privacy achieved by a mechanism from a real user perspective. We analyse the requirements and the features needed to actually evaluate the privacy of obfuscated texts beyond the formal privacy provided by the analysis of the mechanisms' parameters, and suggest some research directions to devise new evaluation measures for this purpose.

Keywords

Privacy-Preserving Information Retrieval, Evaluation Measures, Differential Privacy, Information Security

1. Introduction

Natural Language Processing (NLP) and Information Retrieval (IR) systems are developed using extensive textual datasets, including queries, documents, reviews, and online posts, which frequently contain sensitive and personal user information. The presence in such texts of personal information, e.g., the user profile and personal opinions, poses a serious matter for users interacting with the systems. Such privacy concerns, if not properly mitigated, can endanger the users' safety after text analysis has been pursued. For instance, imagine a scenario, where a user expresses on a social network his disagreement with a specific political view in an illiberal country [1, 2]. Moreover, from the browser search history, it is possible to infer and disclose sensitive information, such as his salary or medical conditions, by analyzing the search queries and documents retrieved [3, 4]. To address such privacy concerns in a formal manner, ε-Differential Privacy (DP) [5], k-anonymity [6], ℓ-diversity [7], t-closeness [8] represent the Gold Standard definitions of providing privacy. Formally, the privacy parameters, e.g., ε and k, regulate the amount of obfuscation provided to the original data. Thus, these values serve as indicators of the amount of formal privacy provided by the obfuscation mechanisms.

IIR 2024: 14th Italian Information Retrieval Workshop, September 5-6, 2024, Udine, Italy

francescoluigi.defaveri@phd.unipd.it (F. L. De Faveri); faggioli@dei.unipd.it (G. Faggioli); ferro@dei.unipd.it (N. Ferro)

https://www.dei.unipd.it/~defaverifr/ (F. L. De Faveri); https://www.dei.unipd.it/~faggioli/ (G. Faggioli); https://www.dei.unipd.it/~ferro/ (N. Ferro)

0009-0005-8968-9485 (F. L. De Faveri); 0000-0002-5070-2049 (G. Faggioli); 0000-0001-9219-6239 (N. Ferro)

However, the actual amount of real privacy achieved is often overlooked. Indeed, just analyzing these values is not sufficient for assessing the real level of privacy of the obfuscated texts produced. For instance, a ε-DP mechanism might replace the original text "melanoma breast cancer" with "disease breast tumour". Although the formal privacy level is ensured by the ε value, the actual privacy is effectively zero because an external human auditor can quickly infer the user information need. As confirmed by previous studies [9, 10], taking into account only the value of privacy parameters is not enough to measure the privacy and new measures of evaluation are needed to inspect the similarities of the original and obfuscated text [11, 12, 13].

In this paper, we discuss several key aspects to be considered when measuring privacy in textual obfuscation. We present the theoretical background on the mechanisms used for textual obfuscation in IR tasks. Once an obfuscation mechanism has produced the private texts, we stress the need for a privacy analysis beyond the only study of the mechanisms' privacy parameters. Furthermore, we discuss the challenges of evaluating privacy, highlighting the distinctions between the formal and practical value of privacy in the obfuscated texts.

The paper is organized as follows: Section 2 explains how privacy is achieved in text obfuscation; Section 3 addresses the challenges in measuring privacy and proposes new directions on how to assess a concrete level of privacy beyond the obfuscation parameters.

2. Preserving Privacy in texts

2.1. Parameter-Based Obfuscation

k-anonymity and ℓ-diversity [6, 7] are privacy metrics based on the pre-image cardinality of an obfuscation mechanism. By generalizing and suppressing the original data, k-anonymity and ℓ-diversity offer the needed amount of obfuscation determined by the parameters k and ℓ.

On the other hand, t-closeness [8] is a metric based on the conditional distribution between original and obfuscated data. t-closeness estimates the divergence between the obfuscated data released which is controlled by the threshold t. Akin to k-anonymity and ℓ-diversity the data are generalized and suppressed until the required level t is verified. However, privacy is only measured by the privacy budget set, without considering the effective obfuscation of the data.

2.2. Differential Privacy Obfuscation

The principal framework to formally define privacy is ε-DP, introduced by Dwork et al. [5]. Essentially, DP introduces carefully calibrated noise levels during output computation using a privacy budget denoted as ε, which controls the equilibrium between data privacy and utility. The definition of DP is based on the notion of neighbouring datasets, i.e., datasets that diverge by at most one record. The formal definition of DP states that a randomized mechanism $\mathcal{M}$, i.e., a mechanism that takes as input the original data and produces a noisy output, is ε-DP if for any pair of neighbouring datasets D and D' and a privacy budget $\varepsilon \in \mathbb{R}^+$, Equation 1 holds.

$$\Pr\left[\mathcal{M}(D) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\mathcal{M}(D') \in S\right] \quad \forall S \subset \mathrm{Im}(\mathcal{M}) \tag{1}$$

When a randomized mechanism respects ε-DP, it ensures that the likelihood of observing any output remains nearly equal for any neighbouring datasets. Consequently, the mechanism keeps

user privacy by introducing uncertainty regarding the original data. As per the definition, the lower values of ε correspond to elevated privacy levels. Specifically, if $\varepsilon = 0$, the output of the mechanism becomes independent of the input, as indicated by the equation $\Pr[\mathcal{M}(D) \in S] = \Pr[\mathcal{M}(D') \in S] \quad \forall S \subset \mathrm{Im}(\mathcal{M})$. Thus, the important information-theoretic property that DP provides the user is the "Plausible Deniability", i.e., the statistical indistinguishability: thus, an adversary cannot definitively link the original data to the obfuscated output.

Different obfuscation methods have been proposed to protect sensitive information in texts for the NLP domain. Once the word embedding vector $\phi(w)$ has been computed, a DP mechanism takes as input $\phi(w)$ randomizes its original value. Specifically, to achieve DP, two strategies can be employed: after adding statistical noise to the original word embedding, the nearest value is selected as the candidate obfuscated word. On the other hand, the mechanism accounts for the distances from all words in a vocabulary to create a ranked list of potential candidates, then, the new term is chosen from the list based on a probability distribution, parametrized by ε.

2.3. Heuristic-Based Obfuscation

Obfuscating texts using DP is not the only strategy to achieve the privacy desired. Previously proposed methods relied on non-formal privacy methods. A study by Arampatzis et al. [14] explored an obfuscation technique leveraging WordNet [15]. This method involves extracting synonyms, hypernyms, and holonyms from WordNet for each term in a sentence. Therefore, the approach considers sets of terms two steps away on the WordNet hierarchy and computes the obfuscation candidates as the Cartesian product between the sets. To avoid exposure of the initial terms, a similarity function is employed to filter out words too similar to the original. Fröbe et al. [16] develop an extension of the Arampatzis et al. [17, 18] method based on word statistics. This approach involves using a local corpus to select and filter candidate obfuscations. To enhance privacy, all candidates chosen from the possible combinations among the top-k documents are discarded if they are synonyms, hypernyms, or hyponyms. On one hand, these approaches effectively mask the actual user information needs by altering words through a lexical approach, thereby providing tangible privacy that limits the inference of the true meaning of the texts. On the other hand, no formal privacy guarantees are provided, as the mechanisms operate independently of ε or any other formal parameter.

3. General Challenges of Measuring Privacy

3.1. Contextual Privacy Evaluation

While individual terms lack inherent sensitivity, their impact is contingent on the context in which they are used. The sensitivity of a term is generated by the interactions between words used in the sentence where the term is placed. Thus, the context of a sentence provides the original meaning that the user needs to communicate by using the initial word. Introducing a metric to assess the context's sensitivity before text obfuscation can help moderate the difference between formal and actual privacy. By evaluating the context prior to the obfuscation, privacy can be measured in a broader scenario. Such a context is independent of the obfuscation parameters and needs to be assessed during an initial privacy analysis of texts, as also motivated by prior

studies [19, 20, 21]. The absence of context in the obfuscation may result in the mechanisms failing to capture the intended meaning, tone, or sentiment of the text, consequently generating irrelevant or inaccurate obfuscations. Therefore, utilizing a NLP model to comprehend context is essential for effectively managing the variability and richness of natural language, taking into account both the lexical and semantic structures of sentences [22, 23].

3.2. Limits of Traditional Privacy Measures

Traditional privacy measures are based on the information-theoretic concept of entropy [24]. Specifically, the metrics used to evaluate the failure rates of DP an obfuscation mechanism $\mathcal{M}$ are derived from an extension of classical entropy, namely the entropy introduced by Rényi [25]. This entropy is employed to assess the probability of a failure during the obfuscation, denoted as $N_w = \Pr[\mathcal{M}(w) = w]$, and the cardinality of the smallest output set defined as $S_w = \min|\{S \subseteq X : \Pr[\mathcal{M}(w) \notin S] \leq \eta\}|$. These measures are typically calculated by simulating the obfuscation of a randomly sampled word w over a specified number of iterations T, e.g., in [26, 27, 28] the authors agree to use $T = 100$. The frequency with which the mechanism returns w represents the approximation of N_w, and the ratio of unique words generated over the T simulations provides the estimate of S_w. Intuitively, a proper setting of ε in a DP mechanism should ensure that N_w is small and S_w is (almost) all words $w \in X$.

However, these uncertainty statistics retain certain limitations. Consider a scenario where a mechanism appends a special symbol to the last character of the obfuscated word regardless of the original term, or it changes a word with a synonym like [14, 16]. As a result, the probability that the mechanism obfuscates a word by mapping it to itself would always be equal to zero. Consequently, $N_w = 0$ for all possible values of the obfuscation parameter. On the other hand, if we let the mechanism change the special symbol appended to the obfuscated word, the set of possible obfuscated candidates will never be zero. Therefore, conducting such an assessment on its own is not enough to truly comprehend the concrete privacy provided by the mechanisms. Furthermore, widely used metrics for evaluating the quality of machine-generated texts, such as BLEU [29], ROUGE [30], and METEOR [31], are not traditionally used to evaluate the lexical and semantic obfuscation levels provided by the DP mechanisms.

These metrics can be helpful for estimating the proportions of n-grams, longest common subsequences, stemming, and synonyms in the obfuscated texts; thus ensuring a deeper analysis of the privacy provided. However, such measures suffer from a lack of semantic understanding, not catching the true semantic relationships between terms in the text, thus obfuscating the original terms with synonyms and hyponyms. To address this limitation, Transformer-based metrics, e.g., BERT-Scores [32], can effectively encode the entire sentence and capture the true similarities between original and obfuscated phrases. Nonetheless, BERT-Scores deeply depend on the pre-trained models used (RoBERTa, ALBERT, etc.) and yield different results, introducing biases in the privacy evaluation from models trained on specific domains. As a preliminary study, Faggioli and Ferro [33] proposed to adopt also the Jaccard Index and the sentence embedding computed by a Language Model to explore the lexical and semantic similarity between original and obfuscated texts. However, the former does not consider synonyms and hyponyms of terms, and the Transformer method can yield different results depending on the model used. We recap the benefits and the disadvantages of adopting the aforementioned measures in Table 1.

Table 1
Summary table of measure to evaluate privacy beyond the privacy parameters.

Measure	PROs	CONs
BLEU [29]	- Effective short n-grams comparison - Computationally efficient	- No semantic similarity - Long distance dependences
ROUGE [30]	- n-gram comparison - Computationally efficient	- No semantic similarity - Long distance dependences
METEOR [31]	- Synonyms and stemming - Paraphrasing	- Heuristic-based - Computational expensive
Jaccard Index	- Lexical similarity - Computationally efficient	- No semantic similarity - No synonyms or paraphrasing
N_w and S_w [25]	- Lexical similarity - Measure of mechanism failure	- No semantic similarity - Easy to deceive
BERTScore [32]	- Semantic similarity - Long distance dependences	- Pre-trained model dependent - Computational expensive
Transformers Sentence Embeddings Similarity	- Semantic similarity - Long distance dependences	- Pre-trained model dependent - Computational expensive

3.3. Adversarial Capacity of Breaking Privacy

Privacy is a security objective that guarantees that the information is used without revealing personal information to external entities. Attackers seek to decode obfuscated sentences and uncover the original content with the purpose of gaining insights into user data for malicious activities. Different strategies are used to evaluate the probability of a successful attack [12]. The attack scenarios chosen are based on the attacker's characteristics, i.e., available resources and prior knowledge. Additionally, the principal classes used to evaluate the success probability of an attack include Membership and Attribute Inference Attacks [34, 35], whose probability of success is strictly evaluated depending on the privacy budget parameter. The former estimates the likelihood that a sampled text belongs to a specific dataset; the latter quantifies the probability that an adversary can infer sensitive attributes about individuals from the obfuscated texts.

However, no longitudinal analysis of textual obfuscation has been conducted to investigate the adversarial risks associated with disclosing personal information in the output generated by the obfuscation mechanisms. To address this gap, a parameter-independent metric for assessing adversarial privacy risks over time could be the key to mitigating this issue effectively. Such metrics should take into account the evolving distributions of text obfuscation and the advancing capabilities of the adversaries. This approach facilitates the identification of potential vulnerabilities that may emerge as attackers adapt and refine their strategies. Moreover, it enables the continuous improvement of obfuscation mechanisms to neutralise these evolving threats, thereby improving the overall robustness of the privacy-preserving methods.

4. Conclusion and Future Work

In this study, we framed the problem of assessing privacy for obfuscated texts beyond the privacy parameters, e.g., ε, k. Furthermore, we described the theoretical context of text obfuscation and the open challenges of measuring privacy, proposing different aspects to take into account when assessing the privacy provided to the original texts. In future directions, we plan to evaluate Privacy-Preserving Information Retrieval pipelines to empirically address such measures for estimating the concrete level of privacy achieved by the obfuscation mechanisms.

References

[1] H. Le, R. Maragh, B. Ekdale, A. High, T. Havens, Z. Shafiq, Measuring political personalization of google news search, in: The World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2957–2963. URL: https://doi.org/10.1145/3308558.3313682. doi:10.1145/3308558.3313682.

[2] E. Mustafaraj, E. Lurie, C. Devine, The case for voter-centered audits of search engines during political elections, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 559–569. URL: https://doi.org/10.1145/3351095.3372835. doi:10.1145/3351095.3372835.

[3] M. Barbaro, T. Zeller, A Face Is Exposed For AoL Searcher No. 4417749, New York Times (2006).

[4] S. Bavadekar, A. M. Dai, J. Davis, D. Desfontaines, I. Eckstein, K. Everett, A. Fabrikant, G. Flores, E. Gabrilovich, K. Gadepalli, S. Glass, R. Huang, C. Kamath, D. Kraft, A. Kumok, H. Marfatia, Y. Mayer, B. Miller, A. Pearce, I. M. Perera, V. Ramachandran, K. Raman, T. Roessler, I. Shafran, T. Shekel, C. Stanton, J. Stimes, M. Sun, G. Wellenius, M. Zoghi, Google COVID-19 search trends symptoms dataset: Anonymization process description (version 1.0), CoRR abs/2009.01265 (2020). URL: https://arxiv.org/abs/2009.01265. arXiv:2009.01265.

[5] C. Dwork, F. McSherry, K. Nissim, A. D. Smith, Calibrating noise to sensitivity in private data analysis, in: S. Halevi, T. Rabin (Eds.), Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings, volume 3876 of *Lecture Notes in Computer Science*, Springer, 2006, pp. 265–284. URL: https://doi.org/10.1007/11681878_14. doi:10.1007/11681878_14.

[6] L. Sweeney, k-anonymity: A model for protecting privacy, Int. J. Uncertain. Fuzziness Knowl. Based Syst. 10 (2002) 557–570. URL: https://doi.org/10.1142/S0218488502001648. doi:10.1142/S0218488502001648.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, l-diversity: Privacy beyond k-anonymity, in: L. Liu, A. Reuter, K. Whang, J. Zhang (Eds.), Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA, IEEE Computer Society, 2006, p. 24. URL: https://doi.org/10.1109/ICDE.2006.1. doi:10.1109/ICDE.2006.1.

[8] N. Li, T. Li, S. Venkatasubramanian, t-closeness: Privacy beyond k-anonymity and l-diversity, in: R. Chirkova, A. Dogac, M. T. Özsu, T. K. Sellis (Eds.), Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007, IEEE Computer Society, 2007, pp. 106–115. URL: https://doi.org/10.1109/ICDE.2007.367856. doi:10.1109/ICDE.2007.367856.

[9] J. Domingo-Ferrer, D. Sánchez, A. Blanco-Justicia, The limits of differential privacy (and its misuse in data release and machine learning), Commun. ACM 64 (2021) 33–35. URL: https://doi.org/10.1145/3433638. doi:10.1145/3433638.

[10] J. Mattern, B. Weggenmann, F. Kerschbaum, The limits of word level differential privacy, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics,

Seattle, United States, 2022, pp. 867–881. URL: https://aclanthology.org/2022.findings-naacl.65. doi:10.18653/v1/2022.findings-naacl.65.

[11] J. Parra-Arnau, D. Rebollo-Monedero, J. Forné, Measuring the privacy of user profiles in personalized information systems, Future Gener. Comput. Syst. 33 (2014) 53–63. URL: https://doi.org/10.1016/j.future.2013.01.001. doi:10.1016/J.FUTURE.2013.01.001.

[12] I. Wagner, D. Eckhoff, Technical privacy metrics: A systematic survey, ACM Comput. Surv. 51 (2018). URL: https://doi.org/10.1145/3168389. doi:10.1145/3168389.

[13] M. Mozes, B. Kleinberg, No intruder, no validity: Evaluation criteria for privacy-preserving text anonymization, CoRR abs/2103.09263 (2021). URL: https://arxiv.org/abs/2103.09263. arXiv:2103.09263.

[14] A. Arampatzis, P. S. Efraimidis, G. Drosatos, A query scrambler for search privacy on the internet, Inf. Retr. 16 (2013) 657–679. URL: https://doi.org/10.1007/s10791-012-9212-1. doi:10.1007/S10791-012-9212-1.

[15] G. A. Miller, Wordnet: a lexical database for english, Commun. ACM 38 (1995) 39–41. URL: https://doi.org/10.1145/219717.219748. doi:10.1145/219717.219748.

[16] M. Fröbe, E. O. Schmidt, M. Hagen, Efficient query obfuscation with keyqueries, in: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT '21, Association for Computing Machinery, New York, NY, USA, 2022, p. 154–161. URL: https://doi.org/10.1145/3486622.3493950. doi:10.1145/3486622.3493950.

[17] A. Arampatzis, G. Drosatos, P. S. Efraimidis, A versatile tool for privacy-enhanced web search, in: P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. M. Rüger, E. Agichtein, I. Segalovich, E. Yilmaz (Eds.), Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings, volume 7814 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 368–379. URL: https://doi.org/10.1007/978-3-642-36973-5_31. doi:10.1007/978-3-642-36973-5_31.

[18] A. Arampatzis, G. Drosatos, P. S. Efraimidis, Versatile query scrambling for private web search, Inf. Retr. 18 (2015) 331–358. URL: https://doi.org/10.1007/s10791-015-9256-0. doi:10.1007/s10791-015-9256-0.

[19] H. Nissenbaum, Respecting context to protect privacy: Why meaning matters, Sci. Eng. Ethics 24 (2018) 831–852. URL: https://doi.org/10.1007/s11948-015-9674-9. doi:10.1007/S11948-015-9674-9.

[20] S. Sousa, R. Kern, How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing, Artif. Intell. Rev. 56 (2023) 1427–1492. URL: https://doi.org/10.1007/s10462-022-10204-6. doi:10.1007/S10462-022-10204-6.

[21] O. Klymenko, S. Meisenbacher, F. Matthes, Differential privacy in natural language processing the story so far, in: O. Feyisetan, S. Ghanavati, P. Thaine, I. Habernal, F. Mireshghallah (Eds.), Proceedings of the Fourth Workshop on Privacy in Natural Language Processing, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1–11. URL: https://aclanthology.org/2022.privatenlp-1.1. doi:10.18653/v1/2022.privatenlp-1.1.

[22] S. Meisenbacher, N. Nandakumar, A. Klymenko, F. Matthes, A comparative analysis of word-level metric differential privacy: Benchmarking the privacy-utility trade-off, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and

Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 174–185. URL: https://aclanthology.org/2024.lrec-main.16.

[23] F. L. De Faveri, G. Faggioli, N. Ferro, py-PANTERA: A Python PAckage for Natural language obfuscaTion Enforcing pRivacy & Anonymization, in: Proceedings of the 33rd ACM Interna- tional Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA., Springer, 2024, p. 6. URL: https://doi.org/10.1145/3627673.3679173. doi:10.1145/3627673.3679173.

[24] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 623–656. URL: https://doi.org/10.1002/j.1538-7305.1948.tb00917.x. doi:10.1002/J.1538-7305.1948.TB00917.X.

[25] A. Rényi, On measures of entropy and information, 1961. URL: https://api.semanticscholar.org/CorpusID:123056571.

[26] O. Feyisetan, B. Balle, T. Drake, T. Diethe, Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations, in: Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 178–186. URL: https://doi.org/10.1145/3336191.3371856. doi:10.1145/3336191.3371856.

[27] Z. Xu, A. Aggarwal, O. Feyisetan, N. Teissier, A differentially private text perturbation method using regularized mahalanobis metric, in: O. Feyisetan, S. Ghanavati, S. Malmasi, P. Thaine (Eds.), Proceedings of the Second Workshop on Privacy in NLP, Association for Computational Linguistics, Online, 2020, pp. 7–17. URL: https://aclanthology.org/2020.privatenlp-1.2.pdf. doi:10.18653/v1/2020.privatenlp-1.2.

[28] X. Yue, M. Du, T. Wang, Y. Li, H. Sun, S. S. M. Chow, Differential privacy for text analytics via natural text sanitization, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 3853–3866. URL: https://aclanthology.org/2021.findings-acl.337. doi:10.18653/v1/2021.findings-acl.337.

[29] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA, ACL, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.

[30] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[31] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909.

[32] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[33] G. Faggioli, N. Ferro, Query obfuscation for information retrieval through differential

privacy, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 278–294.

[34] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017, IEEE Computer Society, 2017, pp. 3–18. URL: https://doi.org/10.1109/SP.2017.41. doi:10.1109/SP.2017.41.

[35] J. Mattern, F. Mireshghallah, Z. Jin, B. Schölkopf, M. Sachan, T. Berg-Kirkpatrick, Membership inference attacks against language models via neighbourhood comparison, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 11330–11343. URL: https://doi.org/10.18653/v1/2023.findings-acl.719. doi:10.18653/V1/2023.FINDINGS-ACL.719.