DP-COMET: A Differential Privacy Contextual Obfuscation MEchanism for Texts in Natural Language Processing

Francesco Luigi De Faveri*
University of Padova
Padova, Italy
francescoluigi.defaveri@phd.unipd.it

Guglielmo Faggioli
University of Padova
Padova, Italy
guglielmo.faggioli@unipd.it

Nicola Ferro
University of Padova
Padova, Italy
nicola.ferro@unipd.it

Abstract

Protecting sensitive information within textual data strongly depends on the context in which the data is presented. However, current privacy-preserving obfuscation mechanisms based on ε-Differential Privacy (DP) produce an obfuscated private text changing the original phrase term-by-term without considering the context in which such a term is placed. This paper introduces DP-COMET, an ε-DP obfuscation mechanism that evaluates a text's context before producing its private version. The mechanism defines a representation of the original text that considers the entire context within the text, producing an obfuscated version after adding noise to this representation and depending on the privacy parameter ε. We test DP-COMET on different Natural Language Processing (NLP) and Information Retrieval (IR) downstream tasks, and our findings show that our obfuscation mechanism not only achieves comparable performance results to traditional term-by-term mechanisms but also produces obfuscated texts less similar to the originals. To promote the reproducibility of DP-COMET, we make the code publicly available at https://github.com/Kekkodf/DP-COMET.

CCS Concepts

• **Security and privacy** → **Privacy-preserving protocols**; **Privacy protections**; **Usability in security and privacy**.

Keywords

Differential Privacy, Text Obfuscation, Information Security, Information Retrieval, Natural Language Processing

ACM Reference Format:
Francesco Luigi De Faveri, Guglielmo Faggioli, and Nicola Ferro. 2025. DP-COMET: A Differential Privacy Contextual Obfuscation MEchanism for Texts in Natural Language Processing. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3746252.3760888

1 Introduction

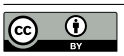
Approximately 80% of the web's content is in unstructured form [15], i.e., textual data, like social network and blog posts, queries submitted to search engines, and other textual shared content as primary examples. This data often contains sensitive and private details of users, which can be inadvertently ingested and processed if the data is used to train a machine learning model, such as those used for Natural Language Processing (NLP) and Information Retrieval (IR) tasks. Therefore, when handling and processing such data, users' privacy must be properly preserved [19, 22, 32].

The ε-Differential Privacy (DP) [10] is among the most popular frameworks used to formally quantify privacy and obfuscate the data by adding noise. However, when obfuscating textual data, DP state-of-the-art mechanisms work by obfuscating texts in a term-by-term style [6, 12, 21, 23, 35, 36], i.e., tokenising the input text and changing each term independently from the context they appear in. Failing to consider the context of a word might result in inadequate privacy protection [1, 7, 8, 26]: for example, the token "312" carries different sensitivity levels in the sentences "The Eiffel Tower is 312 meters high" and "My Card CVV is 312".

In this paper, we introduce the ε-DP **C**ontextual **O**bfuscation **ME**chanism for **T**extual data (DP-COMET), a novel text obfuscation strategy for NLP and IR tasks. Rather than relying on precomputed word embeddings and independently concatenating each term, DP-COMET employs an encoder to project the entire sentence into a latent contextualised embedding. Such an embedding is then perturbed with statistical noise, whose distribution and magnitude are derived from the DP framework as a function of the privacy budget ε. The noisy contextual vectors are then used to retrieve a suitable obfuscation piece of text from a publicly available corpus. Finally, the obfuscation text is used instead of the original one to carry out the downstream task. To empirically validate the effectiveness of DP-COMET, we consider two downstream tasks: sentiment analysis and query obfuscation protocol. Our findings show that contextual privatised embeddings obtained with DP-COMET present higher utility and stronger privacy guarantees compared to term-by-term privacy mechanisms, exhibiting higher task-specific effectiveness with lower semantic and lexical similarity to the original text. The results indicate that DP-COMET produces private texts that preserve a similarity of only 5% to the original ones on average while achieving the same effectiveness as state-of-the-art mechanisms.

Section 2 introduces the framework of ε-DP, focusing on how DP obfuscation mechanisms for textual data work. Section 3 presents the DP-COMET mechanism and how it provides textual obfuscation, producing private texts. Section 4 shows the experimental comparison between DP-COMET and the traditional state-of-the-art DP mechanisms regarding effectiveness and privacy analysis. Finally, Section 5 reports the conclusions and outlines future work.

*Corresponding Author.



CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3760888

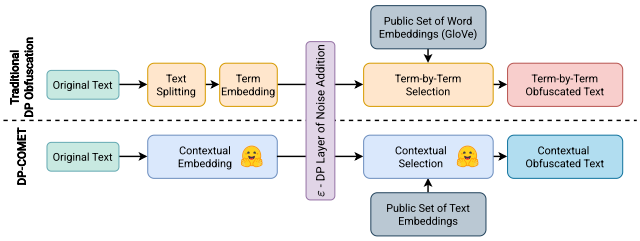


Figure 1: High-level view of ε-DP obfuscation for texts.

2 Preliminaries

2.1 ε-Differential Privacy (DP) framework

According to the ε-DP mathematical definition [10], a mechanism $\mathcal{M} : \{D_i\}_{i\in\mathbb{N}} \to \text{Image}(\mathcal{M})$ is said to be ε-DP if and only if for any two neighbouring datasets[1] D_1, D_2 and any $\varepsilon \in \mathbb{R}^+$, identified with the terms "*privacy budget*", Inequality 1 holds.

$$\Pr[\mathcal{M}(D_1) \in \mathcal{S}] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D_2) \in \mathcal{S}] \quad \forall \mathcal{S} \subseteq \text{Image}(\mathcal{M}) \quad (1)$$

Given two similar inputs, D_1 and D_2, a DP mechanism should provide similar corresponding outputs. The lower the privacy budget ε, the harder it is to identify which input (i.e., either D_1 or D_2) the output corresponds to. Conversely, larger ε values reduce the privacy guarantees, making it easier to identify the input corresponding to a given output. Typically, this is achieved by introducing statistical noise that depends on the privacy parameter ε in the computation.

When dealing with metric spaces, the standard notion of DP (Eq. 1) is often too strict. Hence, Chatzikokolakis et al. [3] relaxed it by introducing the concept of *metric*-DP. Let $\mathcal{X}$ be a vector space. A mechanism $\mathcal{M}$ provides *metric*-DP if, considering a distance function $\delta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and any $\varepsilon \in \mathbb{R}^+$, Inequality 2 holds.

$$\Pr[\mathcal{M}(x) = \tilde{x}] \leq e^{\varepsilon \cdot \delta(x,x')} \cdot \Pr[\mathcal{M}(x') = \tilde{x}] \quad \forall x, x', \tilde{x} \in \mathcal{X} \quad (2)$$

This relaxation allows us to map points close in $\mathcal{X}$ on the same output $\tilde{x}$—with the notion of "close" being regulated by ε—thus increasing privacy. On the other hand, points that are already far apart are more likely to be mapped to different outputs.

2.2 ε-DP Obfuscation Mechanisms for Text

Providing privacy to textual data requires changing the text so that an adversary can not entirely reconstruct the information contained in the original text while maintaining (part of) its utility. ε-DP obfuscation mechanisms for textual data [2, 32, 37] take strings as input and, after a computation process, produce a modified version of the original text [11, 14, 17], thus protecting its original content.

The state-of-the-art text obfuscation ε-DP mechanisms [12, 23, 35, 36] have three phases, depicted in Figure 1. The *preprocessing* step is the initial phase, i.e., the input string is tokenised, and each token is converted into a latent vector representation, using non-contextual embedding approaches, such as GloVe [27]. After the tokenisation, the *distortion* phase changes the encoding by adding statistical noise sampled from a probability distribution—defined by the DP mechanism—whose magnitude often depends on the privacy budget ε. Finally, during the *selection* step, the algorithm replaces the original token by selecting an obfuscation term starting from the noisy embedding built during the previous step. Examples of approaches that follow this paradigm include [12, 35, 36], whose major differences consist of the distribution of the noise and how the obfuscation term is sampled. Recent advances, such as the one proposed by Meisenbacher et al. [23], employ contextualised pre-trained masked language models, e.g., BERT [9]. Despite this, the major limitation of all these approaches lies in the fact that all of them obfuscate the terms individually, disregarding the entire sentence and all possible relations between the different terms.

[1] D_1, D_2 differ only for at most one record, i.e., a single row is cut or added from one.

3 Obfuscation with DP-COMET

To overcome the limitations mentioned above, we propose to obfuscate the representation of the entire string so that its full semantic meaning is protected, rather than just its individual tokens. To this end, we require an encoder $\phi : \mathcal{W}^* \to \mathbb{R}^d$ that maps strings into a d-dimensional latent embedding space. Here, $\mathcal{W}$ denotes a vocabulary of tokens—e.g., the English vocabulary—and $\mathcal{W}^*$ represents a string of arbitrary length. We do not impose any specific constraints on this embedding function; for example, it could be any sentence-level transformer-based encoder [33]. One common approach is to use the `[CLS]` token representation from a BERT model. We define $\mathbf{t} = \phi(t)$, where $t \in \mathcal{W}^*$, as the latent representation of the original sensitive string t. DP-COMET then computes an obfuscated version $\tilde{\mathbf{t}} = \mathbf{t} + \eta$, where $\eta \sim p_\varepsilon$ is noise sampled from a distribution p and regulated by the privacy parameter ε. In light of NLP-oriented downstream tasks, we require a function $\psi : \mathbb{R}^d \to \mathcal{W}^*$ that, given $\tilde{\mathbf{t}}$, maps it back into a human-readable string $\tilde{t} = \psi(\tilde{\mathbf{t}})$, which can be directly used for downstream applications such as training, classification, or retrieval. In practice, both p and ψ must satisfy the metric-DP requirement described in Equation 2. In the following paragraph, we focus on two suitable probability distributions p_ε used for sampling the noise in a DP way: Cumulative Multivariate Perturbation Mechanism (CMP)[12] and Mahalanobis (Mhl)[35]. Finally, we describe the mapping function ψ and discuss the magnitude of the budget ε.

Probability Distribution p_ε. We instantiate the DP-COMET framework by taking inspiration from what is typically done at a token level. In detail, as noise probability distribution p_ε, we employ the distributions used by two well-established state-of-the-art token-level obfuscation mechanisms: CMP [12] and Mhl [35] methods.

CMP [12] samples the perturbation noise from a d-dimensional Laplace whose scale depends on the embedding space dimension and the parameter ε. The sampled noise is later normalised to project it in the unit ball, achieving ε-DP as demonstrated in [34].

Mhl, proposed by Xu et al. [35], is an evolution of CMP that tunes the noise distribution towards densely populated areas of the embedding space, ensuring that the noise $\eta \sim p_\varepsilon$ points towards different points, and increasing the chances of perturbing the vector with a different one every time. To achieve this, the normalised noise is rescaled using the λ-regularised Mahalanobis norm, thus following an elliptical distortion of the vector, stretching the direction of the noisy embedding towards more similar vectors.

The mapping function ψ. Since the considered downstream tasks (sentiment analysis and IR) operate directly on text, we need to

remap the perturbed vector $\tilde{\mathbf{t}}$ to a perturbed string $\tilde{t}$. To obtain a meaningful string, we propose to take it from a set of strings $\mathcal{T}$. In other terms, ψ is such that $\psi(\tilde{\mathbf{t}}) = \tilde{t}$, with $\tilde{t} \in \mathcal{T}$. This has two advantages: first, we ensure that the output string is a well-formed natural language sentence, and second, if the set of strings is publicly available, an adversary observing it would not have more information than what is already publicly available. Hence, the final step is to define how ψ should extract $\tilde{t}$ from $\mathcal{T}$, given $\tilde{\mathbf{t}}$. Following the strategy proposed in [12], we consider $\psi(\tilde{\mathbf{t}}) = \arg\max_{\tilde{t} \in \mathcal{T}}(\tilde{\mathbf{t}} \cdot \phi(\tilde{t}))$. In other terms, ψ selects as obfuscation for the original text t, the string $\tilde{t} \in \mathcal{T}$, whose representation $\phi(\tilde{t})$ has the largest dot product with the obfuscation vector $\tilde{\mathbf{t}}$.

On the magnitude of ε. Feyisetan et al. [12] observed that, for larger text embeddings, larger ε achieve the same privacy-utility trade-off. In other words, different embedding sizes shift the privacy trade-off represented by ε, making it impossible to directly compare it across different embedding sizes: a larger ε does not correspond to lower privacy if the embedding space is larger. Feyisetan et al. [12] and Xu et al. [35] consider at most embeddings with 300 dimensions. Modern contextual encoders have from hundreds to thousands of dimensions.[2] This makes it even more challenging to compare the mechanisms for the same ε. To mitigate this, inspired by the solution adopted by Vaswani et al. [33] for the Transformer attention, we rescale Σ (the covariance of the embedding space matrix, used to sample noise) by $\sqrt{d}$, the square root of the embedding dimension. Importantly, since it does not impact how the noise is sampled, this rescaling maintains the CMP and Mhl mechanisms ε-DP, but makes the trade-off of privacy *vs.* utility dependent only on the ε budget, uncoupling it from the embedding size d.

4 Empirical Experiments

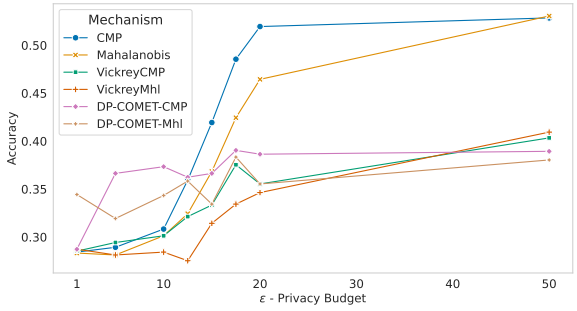


Figure 2: Accuracy results across different privacy budgets ε with the obfuscated tweets on the sentiment dataset [24].

Setup. We validate DP-COMET by testing it on two downstream tasks: the NLP sentiment analysis and the IR query obfuscation protocol. For the first task, we consider two datasets of labelled tweets from the Kaggle platform [24, 31]. The former is our reference collection, while the latter is used as a set $\mathcal{T}$ from which DP-COMET takes the obfuscated pieces of text. We use a BERT model for sentiment analysis [28] to predict the tweets' labels. For the query obfuscation protocol [13], we use two TREC collections – DL'20 [4] (54 queries) and Medline'04 [30] (50 queries) – obfuscating each query 50 times, cf. Section 4.2. As set $\mathcal{T}$, we use the MS-MARCO query set [25] composed by ~809k queries. We employed TAS-B [16] as a retriever and Dragon+ [20] for the reranking.

For both tasks, we employ Contriever [18] as text encoder ϕ. The state-of-the-art mechanisms, i.e., CMP, Mahalanobis (Mhl) and their respective Vickrey's variants [36], are instantiated with GloVe [27] 300-d embeddings in the pyPANTERA privacy framework [5]. To measure the privacy levels, we computed the Jaccard score and cosine similarity with `all-MiniLM-L6-v2` [29] between the original and the obfuscated texts. The source code is publicly available[3].

4.1 NLP: Sentiment Analysis

Effectiveness Analysis. This analysis aims to assess the model's ability to predict the sentiment of obfuscated tweets at different levels of formal privacy defined by the parameter ε. Figure 2 shows the accuracy results of the Tweet-Sentiment classifier [28] across the different privacy budgets. The trends in Figure 2 reveal two distinct patterns in performance across different obfuscation mechanisms: state-of-the-art mechanisms that employ term-by-term obfuscation tend to exhibit reduced accuracy in high-privacy contexts, specifically under low ε privacy budgets. Contrarily, the accuracy achieved with DP-COMET, both using DP-COMET-CMP and DP-COMET-Mhl methodologies, outperforms that of the state-of-the-art at the same ε levels. As the privacy budget value increases, the second pattern emerges with CMP and Mhl mechanisms surpassing the accuracy obtained using the other four strategies. These results are in line with previous ones presented in pyPANTERA [5]. However, the accuracy alone does not allow us to tell whether we have achieved better results, simply sacrificing more privacy: we analyse the accuracy with respect to privacy in the next section.

Privacy Analysis. To assess the privacy guarantees, we compute lexical and semantic similarity between the original and obfuscated texts, using Jaccard similarity at the term level and the cosine similarity between their `all-MiniLM-L6-v2` [29] embeddings, respectively. High similarity indicates low privacy. Table 1 shows the results across the ε values. CMP and Mahalanobis exhibit a significant privacy loss, as highlighted by the increasing cosine and Jaccard scores with an increase in ε. Conversely, Vickrey's mechanisms and the DP-COMET variants are more robust according to both similarity measures, with DP-COMET retaining the lowest similarity, especially at higher ε. If we combine the results reported in Figure 2 and Table 1, we observe that with high levels of formal privacy, i.e., low ε, the DP-COMET mechanisms achieve higher performances without completely destroying the effectiveness on the task. When ε increases, the DP-COMET is by far the most protective strategy, with accuracy comparable to Vickrey's.

4.2 IR: Query Obfuscation Protocol

Effectiveness Analysis. The effectiveness of the document retrieval task is measured by considering the final reranked list of retrieved documents. Figure 3 shows the effectiveness trends obtained considering the Precision@10 (Figures 3(a), 3(b)) and nDCG@10

[2] In our experiments, we used Contriever [18] with an embedding dimension of 768.

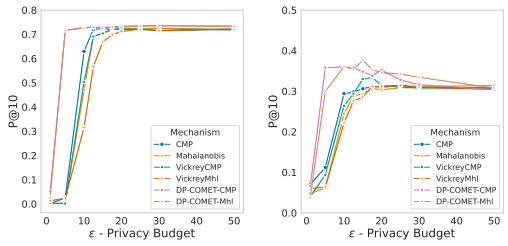
[3] https://github.com/Kekkodf/DP-COMET

Table 1: The Similarity scores between obfuscated and original tweets of [24]. They reflect the degree of resemblance in terms of contextual, i.e., Cosine Similarity, and lexical, i.e., Jaccard Score. The lower the score, the better the privacy.

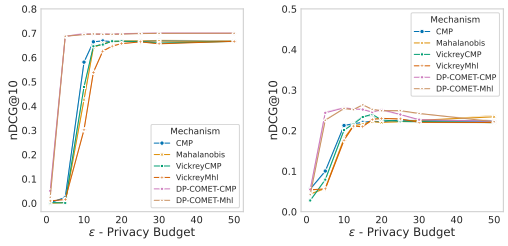
	Cosine Similarity (`all-MiniLM-L6-v2` [29])								*Jaccard Score*							
ε	1.0	5.0	10.0	12.5	15.0	17.5	20.0	50.0	1.0	5.0	10.0	12.5	15.0	17.5	20.0	50.0
CMP	0.064	0.079	0.203	0.373	0.564	0.700	0.771	0.823	0.000	0.002	0.077	0.211	0.402	0.579	0.703	0.841
Mahalanobis	0.066	0.072	0.136	0.229	0.371	0.517	0.640	0.824	0.000	0.003	0.043	0.108	0.220	0.358	0.499	0.840
VickreyCMP	0.064	0.076	0.166	0.242	0.321	0.394	0.429	0.551	0.000	0.001	0.041	0.085	0.123	0.145	0.151	0.178
VickreyMhl	0.059	0.07	0.113	0.169	0.238	0.303	0.360	0.539	0.000	0.003	0.024	0.052	0.087	0.118	0.137	0.175
DP-COMET-CMP	0.111	0.182	0.288	0.313	0.321	0.327	0.342	0.350	0.022	0.041	0.049	0.048	0.054	0.057	0.059	0.057
DP-COMET-Mhl	0.109	0.124	0.197	0.242	0.242	0.313	0.295	0.344	0.017	0.025	0.045	0.042	0.041	0.052	0.051	0.057

Table 2: The Similarity scores between obfuscated and original queries. They reflect the degree of resemblance in terms of contextual, i.e., Cosine Similarity, and lexical, i.e., Jaccard Score. The lower the score, the better the privacy. Due to limited page space, to avoid encumbering, the full similarity results table, also available on DL'20 [4], is accessible in the online repository.

TREC - Medline'04 [30]																
	Cosine Similarity (`all-MiniLM-L6-v2` [29])								*Jaccard Score*							
ε	1.0	5.0	10.0	12.5	15.0	17.5	20.0	50.0	1.0	5.0	10.0	12.5	15.0	17.5	20.0	50.0
CMP	0.043	0.063	0.278	0.488	0.646	0.735	0.772	0.791	0.000	0.003	0.125	0.304	0.505	0.653	0.728	0.781
Mahalanobis	0.042	0.053	0.167	0.314	0.485	0.621	0.710	0.790	0.000	0.003	0.061	0.162	0.314	0.469	0.594	0.781
VickreyCMP	0.046	0.062	0.188	0.276	0.352	0.401	0.437	0.558	0.000	0.003	0.055	0.094	0.128	0.137	0.139	0.169
VickreyMhl	0.044	0.056	0.122	0.196	0.267	0.328	0.382	0.551	0.000	0.002	0.028	0.064	0.094	0.116	0.134	0.171
DP-COMET-CMP	0.026	0.275	0.535	0.557	0.570	0.577	0.584	0.591	0.006	0.029	0.066	0.072	0.073	0.074	0.075	0.076
DP-COMET-Mhl	0.025	0.137	0.453	0.515	0.549	0.562	0.569	0.589	0.006	0.015	0.053	0.067	0.069	0.071	0.075	0.077



(a) P@10 - DL'20 (b) P@10 - Medline'04



(c) nDCG@10 - DL'20 (d) nDCG@10 -Medline'04

Figure 3: Comparison on DL'20 and Medline'04 of the P@10 and nDCG@10 obtained using different obfuscations.

(Figures 3(c), 3(d)) on the DL'20 and Medline'04. In both collections, the DP-COMET mechanisms outperform the state-of-the-art both in terms of P@10 and nDCG@10 across all the epsilon configurations. In particular, at lower values of ε, DP-COMET reveals a clear advantage over the other models operating at the same ε. These results highlight the importance of considering the contextual query phrase, as it significantly improves the quality of search results.

Privacy Analysis. Table 2, reports original and obfuscated strings similarity figures for the Medline'04 collection. To avoid encumbering, the similarity analysis on the DL'20 collection is available in the online repository and shows a similar trend. Regarding the cosine similarity analysis on Medline'04, we observe that the CMP and Mahalanobis mechanisms at high values of ε do not preserve any privacy, producing texts that are similar to the originals with similarity as high as 0.791. Conversely, Vickrey's and DP-COMET exhibit a lower similarity than CMP and Mahalanobis for all the epsilons. The Jaccard similarity confirms such patterns, with DP-COMET achieving a score of 0.006 at $\varepsilon = 1$ and 0.077 with $\varepsilon = 50$. DP-COMET shows robust privacy protection, reducing the proportion of common words between original texts and the respective obfuscations.

5 Conclusion

Privacy in NLP and IR tasks remains an open challenge. Standard obfuscation mechanisms employ ε-DP to provide privacy in a term-by-term fashion, i.e., changing each word independently from its context. In this paper, we introduced DP-COMET, a textual obfuscation mechanism that, within the ε-DP framework, considers the context in which a term is used, thus selecting the obfuscated texts based on the contextual similarity with the original. Our findings indicate that DP-COMET outperforms state-of-the-art methods in NLP and IR tasks, particularly at low privacy budget values, while ensuring strong privacy guarantees. In particular, for sentiment analysis, DP-COMET offers robust contextual and lexical privacy protections for the original text when analysing its obfuscations.

In future work, we intend to investigate new methods for producing obfuscated text using LLMs to dynamically generate new textual information based on contextual representations. Additionally, we plan to explore the impact of other probability noise distributions on DP-COMET performances in the privacy *vs.* utility tradeoff.

GenAI Usage Disclosure

Following ACM's guidelines on the use of generative AI tools (Grammarly Pro), we disclose that generative AI technologies were used solely to assist in code debugging and grammar checking during the preparation of this paper. All research ideas, experiments, analysis, and writing were conducted and critically reviewed by the authors. No part of the scientific content or creative reasoning has been generated or substantially rewritten by generative AI tools.

References

[1] Sebastian Benthall and Rachel Cummings. 2024. Integrating Differential Privacy and Contextual Integrity. In *Proceedings of the Symposium on Computer Science and Law, CSLAW 2024, Boston, MA, USA, March 12-13, 2024*. ACM, 9–15. doi:10.1145/3614407.3643702

[2] Danushka Bollegala, Tomoya Machide, and Ken-ichi Kawarabayashi. 2022. Query Obfuscation by Semantic Decomposition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, 6200–6211. https://aclanthology.org/2022.lrec-1.667

[3] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the Scope of Differential Privacy Using Metrics. In *Privacy Enhancing Technologies - 13th International Symposium, PETS 2013, Bloomington, IN, USA, July 10-12, 2013. Proceedings (Lecture Notes in Computer Science, Vol. 7981)*, Emiliano De Cristofaro and Matthew K. Wright (Eds.). Springer, 82–102. doi:10.1007/978-3-642-39077-7_5

[4] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. *CoRR* abs/2102.07662 (2021). arXiv:2102.07662 https://arxiv.org/abs/2102.07662

[5] Francesco Luigi De Faveri, Guglielmo Faggioli, and Nicola Ferro. 2024. pyPANTERA: A Python PAckage for Natural language obfuscaTion Enforcing pRivacy & Anonymization. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA*. Springer, 6. doi:10.1145/3627673.3679173

[6] Francesco Luigi De Faveri, Guglielmo Faggioli, and Nicola Ferro. 2024. Words Blending Boxes. Obfuscating Queries in Information Retrieval using Differential Privacy. *CoRR* abs/2405.09306 (2024). arXiv:2405.09306 doi:10.48550/ARXIV.2405.09306

[7] Francesco Luigi De Faveri, Guglielmo Faggioli, and Nicola Ferro. 2025. A Comparative Study of Large Language Models and Traditional Privacy Measures to Evaluate Query Obfuscation Approaches. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne (Eds.). ACM, 2711–2716. doi:10.1145/3726302.3730158

[8] Francesco Luigi De Faveri, Guglielmo Faggioli, and Nicola Ferro. 2025. Measuring Actual Privacy of Obfuscated Queries in Information Retrieval. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 15572)*, Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto (Eds.). Springer, 49–66. doi:10.1007/978-3-031-88708-6_4

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. doi:10.18653/V1/N19-1423

[10] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings (Lecture Notes in Computer Science, Vol. 3876)*, Shai Halevi and Tal Rabin (Eds.). Springer, 265–284. doi:10.1007/11681878_14

[11] Guglielmo Faggioli and Nicola Ferro. 2024. Query Obfuscation for Information Retrieval Through Differential Privacy. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 14608)*, Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (Eds.). Springer, 278–294. doi:10.1007/978-3-031-56027-9_17

[12] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy- and Utility-Preserving Textual Analysis via Calibrated Multivariate Perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang (Eds.). ACM, 178–186. doi:10.1145/3336191.3371856

[13] Arthur Gervais, Reza Shokri, Adish Singla, Srdjan Capkun, and Vincent Lenders. 2014. Quantifying Web-Search Privacy. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*, Gail-Joon Ahn, Moti Yung, and Ninghui Li (Eds.). ACM, 966–977. doi:10.1145/2660267.2660367

[14] Ivan Habernal. 2021. When differential privacy meets NLP: The devil is in the detail. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 1522–1528. doi:10.18653/V1/2021.EMNLP-MAIN.114

[15] Khodor Hammoud, Salima Benbernou, Mourad Ouziri, Yücel Saygin, Rafiqul Haque, and Yehia Taher. 2019. Personal Information Privacy: What's Next?. In *Proceedings of the 2nd International Conference on Big Data and Cyber-Security Intelligence, Versailles, France, December 16-17, 2019 (CEUR Workshop Proceedings, Vol. 2622)*, Marie-Rita Hojeij, Yehia Taher, Karine Zeitouni, Béatrice Finance, Rafiqul Haque, and Mohammed Dbouk (Eds.). CEUR-WS.org, 30–37. https://ceur-ws.org/Vol-2622/paper5.pdf

[16] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 113–122. doi:10.1145/3404835.3462891

[17] Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2024. Differentially Private Natural Language Models: Recent Advances and Future Directions. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, 478–499. https://aclanthology.org/2024.findings-eacl.33

[18] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Trans. Mach. Learn. Res.* 2022 (2022). https://openreview.net/forum?id=jKN1pXi7b0

[19] Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential Privacy in Natural Language Processing: The Story So Far. *CoRR* abs/2208.08140 (2022). arXiv:2208.08140 doi:10.48550/ARXIV.2208.08140

[20] Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to Train Your Dragon: Diverse Augmentation Towards Generalizable Dense Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 6385–6400. doi:10.18653/V1/2023.FINDINGS-EMNLP.423

[21] Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2024. A Collocation-based Method for Addressing Challenges in Word-level Metric Differential Privacy. *CoRR* abs/2407.00638 (2024). arXiv:2407.00638 doi:10.48550/ARXIV.2407.00638

[22] Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2025. On the Impact of Noise in Differentially Private Text Rewriting. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, 514–532. https://aclanthology.org/2025.findings-naacl.32/

[23] Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. 2024. DP-MLM: Differentially Private Text Rewriting Using Masked Language Models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 9314–9328. doi:10.18653/V1/2024.FINDINGS-ACL.554

[24] NA. 2025. Twitter Entity Sentiment Analysis. https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis. Accessed: 2025-06-03.

[25] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[26] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Wash. L. Rev.* 79 (2004), 119.

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[28] Juan Manuel Pérez, Juan Carlos Giudici, and Franco M. Luque. 2021. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. *CoRR* abs/2106.09462 (2021). arXiv:2106.09462 https://arxiv.org/abs/2106.09462

[29] Nils Reimers 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*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[30] Patrick Ruch, Christine Chichester, Gilles Cohen, Frédéric Ehrler, Paul Fabry, Johan Marty, Henning Müller, and Antoine Geissbühler. 2004. Report on the TREC 2004 Experiment: Genomics Track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004 (NIST Special Publication, Vol. 500-261)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec13/papers/uhosp-geneva.geo.pdf

[31] Abhishek Shrivastava. 2025. Sentiment Analysis Dataset. https://www.kaggle.com/datasets/abhi8923shriv/sentiment-analysis-dataset. Accessed: 2025-06-03.

[32] Samuel Sousa and Roman Kern. 2023. How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. *Artif. Intell. Rev.* 56, 2 (2023), 1427–1492. doi:10.1007/S10462-022-10204-6

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[34] Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton. 2017. Bolt-on Differential Privacy for Scalable Stochastic Gradient Descent-based Analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data* (Chicago, Illinois, USA) *(SIGMOD '17)*. Association for Computing Machinery, New York, NY, USA, 1307–1322. doi:10.1145/3035918.3064047

[35] Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020. A Differentially Private Text Perturbation Method Using Regularized Mahalanobis Metric. In *Proceedings of the Second Workshop on Privacy in NLP*. Association for Computational Linguistics. doi:10.18653/v1/2020.privatenlp-1.2

[36] Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2021. On a Utilitarian Approach to Privacy Preserving Text Generation. *CoRR* abs/2104.11838 (April 2021). arXiv:2104.11838 [cs.CL] doi:10.48550/ARXIV.2104.11838

[37] Ying Zhao and Jinjun Chen. 2022. A Survey on Differential Privacy for Unstructured Data Content. *ACM Comput. Surv.* 54, 10s (2022), 207:1–207:28. doi:10.1145/3490237