SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Information Systems

Investigating the State-of-the-Art of Differential Privacy in Natural Language Processing

Natalia Milanova

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Information Systems

Investigating the State-of-the-Art of Differential Privacy in Natural Language Processing

Untersuchung des Stands der Technik der Differential Privacy in der natürlichen Sprachverarbeitung

Author:	Natalia Milanova
Supervisor:	Prof. Dr. Florian Matthes
Advisor:	Stephen Meisenbacher, M.S.c
Submission Date:	15.10.2023

I confirm that this bachelor's thesis in information systems is my own work and I have documented all sources and material used.

Munich, 15.10.2023 Natalia Milanova

Acknowledgments

I would like to thank my supervisor, Stephen Meisenbacher, M.S.c, for his valuable guidance throughout this scientific work, and all the participants who participated in the data collection process. Their contribution to this research has provided me with valuable insights and data, which have been vitally important for achieving the objectives of this research. Finally, I would also like to thank my family for their support and love.

Abstract

Natural Language Processing (NLP) has become an essential tool for various applications, including chatbots and sentiment analysis. Despite its rapid growth, this trend towards NLP has also raised concerns about privacy, as the technology often relies on large datasets that contain sensitive user data. To address these concerns, researchers have explored the use of differential privacy, a technique that ensures that individual privacy is not violated when performing data analysis while still providing valuable insights. Nowadays, it is used by researchers, companies and even governments [58]. However, applying differential privacy in NLP comes with its own set of challenges, such as handling textual data and adding noise while preserving the utility and the semantics of the data [108].

Since most academic papers focus on the development of differentially private algorithms, in this paper the research will focus on the appropriate usage of differential privacy when this privacy-preserving technology is mapped to NLP use cases and the arising challenges and benefits of using them together.

Therefore, this paper builds on the literature review of differential privacy and NLP to provide a comprehensive analysis of the properties of differential privacy that are particularly valuable for its practical applications, and provides a foundation for how these properties are applied in NLP. For this purpose, this research paper uses the approach of Gallersdoerfer and Matthes [39] to examine the characteristics of differential privacy from two perspectives. First of all, the current work concentrates on the features that are mandatory from a technical point of view, and secondly, it focuses on the additional properties that make this privacy-preserving technology suitable.

Interviews are then conducted to help expand knowledge of differential privacy and NLP use cases. Thus, our main findings describe the characteristics of differential privacy, that define its practical application from the described perspectives, and explain how these characteristics are applied in NLP use cases, and what are the current challenges and success factors of implementing differential privacy with NLP. Finally, the paper outlines the limitations of this study and presents opportunities for future research that will build on the knowledge gathered to date.

Kurzfassung

Die Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) ist zu einem unverzichtbaren Werkzeug für verschiedene Anwendungen geworden, darunter Chatbots und Stimmungsanalysen. Trotz des rasanten Wachstums hat dieser Trend zu NLP auch Bedenken hinsichtlich des Datenschutzes aufgeworfen, da die Technologie häufig auf großen Datensätzen beruht, die sensible Nutzerdaten enthalten. Um diese Bedenken auszuräumen, haben Forscher die Verwendung der differenziellen Privatsphäre (Differential Privacy) erforscht, eine Technik, die sicherstellt, dass die Privatsphäre des Einzelnen bei der Datenanalyse nicht verletzt wird und dennoch wertvolle Erkenntnisse liefert. Heutzutage wird sie von Forschern, Unternehmen und sogar Regierungen eingesetzt [58]. Die Anwendung des differenziellen Datenschutzes in der NLP bringt jedoch eine Reihe von Herausforderungen mit sich, z. B. den Umgang mit Textdaten und das Hinzufügen von Rauschen bei gleichzeitiger Wahrung des Nutzens und der Semantik der Daten [108].

Da sich die meisten akademischen Arbeiten auf die Entwicklung von Algorithmen zur Wahrung der Privatsphäre konzentrieren, liegt der Schwerpunkt dieser Arbeit auf der angemessenen Nutzung der differentiellen Privatsphäre, wenn diese datenschutzfreundliche Technologie auf NLP-Anwendungsfälle übertragen wird, sowie auf den Herausforderungen und Vorteilen, die sich aus der gemeinsamen Nutzung ergeben.

Daher baut dieses Papier auf der Literaturübersicht über differentielle Privatsphäre und NLP auf, um eine umfassende Analyse der Eigenschaften der differentiellen Privatsphäre zu liefern, die für ihre praktischen Anwendungen besonders wertvoll sind, und eine Grundlage dafür zu schaffen, wie diese Eigenschaften in NLP angewendet werden. Zu diesem Zweck verwendet diese Forschungsarbeit den Ansatz von Gallersdörfer und Matthes [39], um die Eigenschaften der differentiellen Privatsphäre aus zwei Perspektiven zu untersuchen. Zum einen konzentriert sich die vorliegende Arbeit auf die aus technischer Sicht zwingend erforderlichen Merkmale, zum anderen auf die zusätzlichen Eigenschaften, die diese datenschutzfreundliche Technik geeignet machen.

Anschließend werden Interviews durchgeführt, um das Wissen über differenzierten Datenschutz und NLP-Anwendungsfälle zu erweitern. Unsere Hauptergebnisse beschreiben die Merkmale des differenzierten Datenschutzes, die seine praktische Anwendung aus den beschriebenen Perspektiven definieren, und erklären, wie diese Merkmale in NLP-Anwendungsfällen angewendet werden und was die aktuellen Herausforderungen und Erfolgsfaktoren bei der Umsetzung des differenzierten Datenschutzes mit NLP sind. Abschließend werden die Grenzen dieser Studie aufgezeigt und Möglichkeiten für künftige Forschungen vorgestellt, die auf dem bisher gesammelten Wissen aufbauen.

Contents

1. Introduction

This chapter aims to introduce the reader to the problem statement, motivation, and research objectives of this paper. Finally, we will outline the research questions and provide an overview of the paper's structure and content.

1.1. Problem Statement and Motivation

As digitalization progresses, more and more data is circulating. Undeniably, this data exchange provides plenty advantages: A large number of services such as online shopping, online banking and much more are being offered because of the data we, as users, have provided [9]. As the British mathematician Clive Humby stated in 2006, "data is the new oil" [104] now. Raw data, itself, does not represent any knowledge, but when collected, cleaned, processed and analyzed, it can serve decision-makers to make better informed decisions, scientists to infer insights about the population such as health and employment status, educational and demographic statistics, etc. Thus, statistical analysis can nowadays contribute to the overall improvement of society [119].

However, the risk of data breach has also greatly increased. Thus, privacy protection has received a great interest within academic circles. As a main method to protect privacy, organizations and governments used to use de-identification techniques, which removed personally identifiable information (information that can expose the identify of an individual) from the data sets [108].

However, it has been proved data anonymization is not sufficient to preserve privacy [108]: Netflix published large dataset (101 000 000 movie ratings, 500 000 subscribers, time period December 1999-December 2005) with the intention to receive suggestions on how to improve their movie recommendations. Narayanan and Shmatikov demonstrated that it is feasible to identify users with the help of background information (Internet Movie Database) and unleash their personal data [83]. Furthermore, the US Census Bureau performed a database reconstruction attack using the published data from 2010 and found out that they were able to accurately reconstruct private information of more than 40% of the US population. However, not only are structured data sets vulnerable to attacks: Findings demonstrated that an attacker might be able to infer the gender of the text's author with roughly accuracy of 80 out of 100 [64]. Other data breach examples include Facebook, who leaked private information such as phone numbers and user-IDs online in 2019 [124]. In 2020, the private data of more than 6 million Israeli citizens were exposed [124].

Advances in computer power and algorithms, but also the bulk of data we, as users, generate, facilitated the privacy attacks against anonymized data sets. Therefore, scientists all

over the world come up with various privacy-preserving technologies such as k-anonymity, l-divergence, differential privacy, etc. [97, 63]. However, the dominating technology seems to be differential privacy. According to many scientists like Ding and Zhao, "Differential privacy has become a de facto standard"[24, 129].

Due to the fact that we aim to provide more details about differential privacy and its properties in later chapters, here will only provide a short introduction to the topic in order to make it easier for the reader to understand our motivation for this research.

Differential privacy is a privacy-preserving technology, introduced in 2006 by C. Dwork [28]. It provides mathematical guarantees about the privacy of individuals by adding calibrated noise (random values) according to a chosen distribution [117] to a data. This way, it can not be distinguished whether an individual is present or absent in a data set. This uncertainty makes it difficult for an adversary who has access to a result of some statistics performed on a certain data set to infer anything about the individual [117]. Originally, differential privacy is designed for tabular data (table with rows and columns, where a row represents an individual and a column - an attribute) [72]. Due to the nice "promises", differential privacy gives - preserving privacy while maintaining utility, scientists try to adopt this technology in various fields such as networks [105], images [129], machine learning tasks [13, 117], etc. We are also motivated to explore differential privacy with unstructured data, more precisely to investigate the application of differential privacy in NLP, which comes with its own set of challenges, such as handling textual data and adding noise while preserving the utility and the semantics of the data [97]. For instance, according to Shi et al., adding noise to a whole sentence might make the model fail to learn the corresponding language pattern [97]. This challenge is not surprising for the NLP community, where NLP models often rely on vast amounts of text data, including personal and sensitive information, on which the models should be trained, and simultaneously they should generate accurate predictions [33, 86]. This raises concerns about the privacy of individuals whose data is being used. Except for privacy, fairness is also a concern in any machine learning task [86]. We are also motivated by the fact that it is found that more than 3/4 of the bulk of data we publish is actually textual data [46].

Additionally, according to the author's knowledge, there is a prevalent number of scientific papers that focus on differentially private algorithms and diverse differentially private relaxations. However, according to author's knowledge, a few articles and scientific journals concentrate on how differential privacy can be applied in practice, more precisely in the NLP realm because there is a lack of a straightforward connection between differential privacy and natural language processing [63]. Therefore, we are interested in facilitating the answer of this question. Furthermore, a bulk of scientific papers are focused on defining differentially private algorithms and proving them wrong [44]. ADePT and DPText were claimed to be differentially private, then Habernal proved they were not [44, 43]. Although there is a lot of scientific literature about differential privacy available, according to the author's knowledge, there are not any papers that discuss what the requirements are in order to use differential privacy. This is something, we identified as a research gap.

1.2. Objective and Research Questions

This paper intends to fill the research gap, discussed in the previous section, and to provide an overview of the characteristics of differential privacy that are particularly important for its practical applications, then to map these properties to appropriate NLP use cases. We would like to summarize all the needed requirements in order to bridge the missing connection between pure differential privacy and differential privacy within NLP. Additionally, we hope that these requirements might serve as a foundation for checking whether a claimed algorithm is really differentially private one, but they might also summarize which properties from the original ϵ-differential privacy fade out when relaxed versions of differential privacy for NLP settings are used. This is something, we identify as necessary due to the fact that according to Habernal, it is still unclear whether the privacy-preserving technology we are using is actually the intended one [42]. After breaching this gap, we want to dive deeper into the application of differential privacy within NLP and understand what the challenges to adoption and the success factors might be. Our goal is therefore to provide an overview of the state of the art in differential privacy and its applicability to NLP use cases.

Lastly but not least, we hope our work to be an inspiration for more deep research in the NLP sphere with the aim to ensure fast adoption on differential privacy in this realm.

With the intention to fulfil the goals defined in this paper, the following research questions (RQ) have been formulated:

1. RQ1: What are the properties of Differential Privacy that define its practical application settings?
2. RQ2: How can these characteristics be mapped appropriately to Natural Language Processing use cases?
3. RQ3: What are the barriers to adoption for differential privacy in natural language processing, and what success factors have been observed?

1.3. Outlook

The remainder of the paper is structured as follows:

- In chapter 2, we present the corresponding foundations, needed for this paper.
- Chapter 3 will cover the related work used to address the posed research questions.
- In chapter 4, the methodology, used during the research, will be presented.
- In chapters 5, 6, and 7 the findings from the research are discussed.
- Chapter 8 present the limitations of study and provides an extensive discussion on possible future research.
- Finally, a conclusion of the paper is presented in chapter 9.

2. Foundations

In this chapter we provide an introduction to differential privacy and natural language processing, so that it is more easier to the reader to follow this paper.

2.1. Preliminary

2.1.1. Natural Language Processing

In order to improve the reader's understanding, we have decided to give a brief introduction to natural language processing, even though we do not go into its details extensively.

Definition 1.

Natural Language Processing uses methods to translate human language (in both - written and spoken form) into machine language that computers are able to understand it [96].

Within an NLP setting, we usually have a vocabulary V, and a sentence which represents a combination of words, also known as tokens, that belong to this vocabulary V. The vector that results from the mapping of the sentence to a real vector space, is called sentence embedding [87]. This vector representation holds the semantic meaning between the sentence and of its words [87]. The embeddings can be classified as word or sentence embeddings, etc. In a word embedding ($\phi : W \rightarrow \mathbb{R}^n$), a word W is mapped to a vector of real numbers $\mathbb{R}^n$ where the word belongs to some vocabulary V [35]. For more information regarding the embeddings, we advise the reader to have a look on [87].

A bag of words is another way to represent sentences as vectors. The value in each dimension of the vector represents the presence of the corresponding word in the document. However, these representations do not include any information regarding the word order or the semantic meaning [111]. Word2vec representations (word embeddings), however , have a dense representation (a vector where most of its elements are not zero), and if two words have similar context, then their vector representations are pointing to the same direction in the vector space [133].

When it comes to NLP, numerical vector representations are usually used due to the fact that these vector representations of unstructured data can be used for computational purposes more convenient than natural text [63].

2.1.2. Differential Privacy

In order to delve deeper into the application of differential privacy in natural language processing, our research first needs to offer a more comprehensive introduction to this

technology. This section can be skipped by readers who are already familiar with differential privacy.

In 1977, the statistician Tore Dalenius formulated his idea of privacy preservation:

Tore Dalenius: "Nothing about an individual should be learnable from the database that cannot be learned without access to the database" [28]

Although his idea sounds appealing, Dwork proved Dalenius wrong. She demonstrated that due to the auxiliary information an attacker might possess, it is impossible to fulfil the formulated requirement and consequentially she came up with the privacy-preserving technology differential privacy [28]. The aim of differential privacy is concerned about the privacy risk as a result of participating in a database. The risks can also be seen as negative consequences that an individual might experience due to the participation in a database [28]. For instance, if there is a survey revealing whether or not a class of students received good average grades, the result of the survey would not change significantly if the student Tom is participating or not in the survey and differential privacy is applied. Thus, an adversary is unable to conclude what is the average grade of Tom. In the ideal world, the result would not change at all. To sum up, differential privacy guarantees that the associated risk would not increase significantly if a person is within a database. However, anyone is able to disclose information based on the survey [28].

Therefore and as stated earlier, differential privacy allows useful data analysis while providing mathematical guarantees to protect individual private information. That means an adversary is unable to determine whether an individual is present or absent in a database [28]. These guarantees are build upon several definitions:

Neighbouring data sets Two data sets are considered to be neighbouring, if D_2 contains the same records as D_1 (they are identical) but includes or excludes one record A [28].

ϵ-Differential Privacy A randomized operation M satisfies ϵ-differential privacy if the probability of observing an output S after performing this operation on two neighbouring data sets(D_1, D_2) is indistinguishable [28].

The formula for ϵ-differential privacy is provided below [28]:

$$\Pr[\mathcal{M}(D_1) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(D_2) \in S]$$

Differential privacy guarantees that if an individual I is within a database D_1 and after performing some randomized operation M on this database D_1, the probability of observing the output S will be indistinguishable from the probability of observing the output of the randomized operation M performed on the database D_2, where databases D_1 and D_2 differ only in one record A, which contains the data of individual I. This gives the individual plausible deniability - an attacker can not be sure whether individual I is in database D_1, because the attacker can not differentiate between the opt-out and the opt-in scenarios [28]. According to Dwork and Roth, this technology allows us to learn useful information about a population, and simultaneously nothing about a single individual [27].

The Fundamental Law of Information Recovery laid the foundations for the development of differential privacy, so Dwork and Roth, because as it states, privacy information is leaked whenever a query is published and the only way to preserve privacy is to insert random values into results [27] because as Dwork stated: "overly accurate answers to too many questions can destroy any reasonable notion of privacy" [29].

This indistinguishability between the outputs $\Pr[\mathcal{M}(D_1) \in S]$ and $\Pr[\mathcal{M}(D_2) \in S]$ is ensured this way by adding random values in order to obfuscate the contribution of a single record. These random values are called noise and are added to the computation via Laplace,Normal or Exponential distribution, or via randomized response [28]. The noise addition increases privacy guarantees, but harms data utility [28]. The trade-off between privacy and utility is determined by the choice of ϵ value, known in the scientific community as privacy budget [28]. The lower the ϵ value, the lower e^{ϵ} and the lower the difference between $\Pr[\mathcal{M}(D_1) \in S]$ and $\Pr[\mathcal{M}(D_2) \in S]$. Thus, privacy preservation is higher, but utility decreases. Analogically, the conclusions regarding high ϵ value can be made.

Having the privacy budget in its definition, differential privacy requires scientists and practitioners to make a compromise regarding utility: either utility, or privacy needs to be compromised. However, the privacy budget is not fixed and therefore allows to be adjusted accordingly depending on the specific use case, the data and the privacy preferences we have. How much privacy to be preserved and how much utility the analysis will provide, are questions that every practitioner should decide on its own. Questions that might help are: How valuable and needed the utility is? How much privacy do we need to guarantee? It's clear that a data set containing information such as chat history does not need the same privacy preservation as the one containing medical data [116]. Additionally, differential privacy does not only offer adjustments to the privacy budget, but also to the selection of the differentially private mechanism - Laplace distribution, Normal distribution, etc. Scientists all over the world also try to apply differential privacy on structured and on unstructured databases (tabular data sets but also images and graphs) [105, 129], but they also depicted that it is possible with the assistance of a policy function to work with databases which contain both sensitive and insensitive data [97]. Thus, differential privacy ensures at least theoretically that it can be adjusted to any data set and use case, we have. The privacy budget can be chosen by every practitioner depending on their wish [63].

Furthermore, there are other guarantees that differential privacy provides and therefore is considered as a standard privacy-preserving technology. [24, 124, 129] Not only does differential privacy protect individual privacy, but it is also immune to linkage and other unknown future attacks, that anonymization technique is unable to protect from as it was depicted in the chapter Introduction (1) [117]. According to the definition of differential privacy, an adversary may have access to all records except for record A, meaning we assume the worst-case scenario that an attack might have almost complete knowledge of the database [117]. This leads to an inability of the adversary with full, limited or incomplete knowledge to infer private information, unknown before the attack, of any individual [117]. We can conclude that independent of what additional information an attacker might possess even in the future, he/she would not be able to infer private information. Additionally, even if an

adversary is able to add a data record in the database in order to modify the outcome and gain valuable insights, the added record would not have a significant influence on the general result [132]. However it is worth mentioning, a differentially private algorithm does still not protect 100% against inference attacks [63] and still could lead to data leakage [117]. However, the amount of data that is going to be leaked can be controlled through the privacy budget [117].

The privacy budget ensures quantifiable privacy protections. Because ϵ is a positive real number, it can be calculated how much the results $\Pr[\mathcal{M}(D_1) \in S]$ and $\Pr[\mathcal{M}(D_2)$ differ. Having this measurable privacy leakage, participants of a data survey can be ensured that the privacy leaking is bounded to e^{ϵ} value [117]. As a result, they will be aware of consequences of participating [117]. Thus, people might have less incentives to lie when participating in a study because they would know their data is protected [80] and they are able to make better informed decisions [117]. As a consequence, scientists and practitioners will be able to conduct and publish analysis, provide insights regarding their methodology and results [72].

Additionally, differential privacy comes with a composition theorem, which can be split into sequential and parallel composition. These compositions allow us to quantify the privacy-leakage after performing several operations on a data set, which guarantees that differential privacy is robust against attacks [117], where the adversary might try through posing many questions to gain insights about a data record. After privacy budget exceeds some threshold, the queries might be limited [117].

Another valuable supporting characteristic of ϵ-differential privacy is its compliance with privacy regulations. According to Wood et al., many organizations could rely on differential privacy in order to meet these regulations [117]. The General Data Protection Regulation (GDPR) is a privacy regulation applicable within Europe [15] and for the sake of simplicity, we will briefly discuss only how differential privacy applies to GDPR and not to other privacy regulations such as California Consumer Privacy Act (CCPA): GDPR should provide consistent data protection and ensure transparency to all European citizens regarding how their data is collected, analysed and used [15]. Differential privacy, as discussed in the previous sections, does not only protect the individual privacy but is indeed considered to be immune against linkage attacks. Additionally, organisations might escape being fined or losing business reputation as long as they are able to preserve the privacy of their customers and clients [22, 90]. What's more, using differential privacy, companies and institutions are no longer obligated to hide details regarding their privacy preservation technology [117]. They might publish their choice of privacy mechanisms and privacy budget parameters, as it is done within the Epsilon Registry [29], as well as to disclose their verifications and assessments, providing a guarantee that privacy is preserved even after multiple computations on the same data set and even after post-processing the output of a differentially private algorithm [24]. Organisations might disclose whether or not their differenitally private algorithm pass the tests, provided by scientists as counterexamples to verify their differentially private algorithm [24].

Up to this point, we hope the reader is able to identify that differential privacy provides the following benefits:

- Individual privacy can be protected.
- Mathematical assurance and proof that privacy is preserved.
- Quantifiable privacy leakage.
- Flexibility is provided.
- The composition of several differential privacy algorithms is again differentially private.
- Compliance with privacy regulations is ensured.
- Robustness against linkage and any future attacks.

We would also like to discuss shortly the ease of integration of differential privacy, that would serve us as motivation for exploring the answer of the last research question 1.2:

There are plenty of statements claiming differential privacy is not widely used in practice: The statistician Tore Dalenius stated: "differential privacy is an interesting concept, but of little value in practice" [18, 119]. However, according to the author's knowledge many scientists have tried to help applying differential privacy in practice: Scientists like Dwork, Wood and their scientific partners provided a huge amount of scientific works that aim to provide a comprehensive understanding how differential privacy works and how it should be integrated [27, 117]. Other scientists have also demonstrated that differential privacy can be applied to various applications such as statistical analysis and machine learning [117]. Additionally, as it will be stated in the chapter Related Work (3) there are platforms like Epsilon Registry, Tumult Analysis, OpenDP that can be used for ease of understanding and integration. Due to these findings, the paper tries to understand whether all theoretical contributions to the practical applications of differential privacy, are currently used. Thus, to check if the statement about "little value in practice" [18, 119] is no longer correct. That is what we have tried to find with the last research question, for which we conducted semi-structured interviews.

3. Related Work

For our research, we build upon the work of Dwork and Roth [27], who have made significant contributions to the field of privacy-preserving data analysis. They introduced the concept of ϵ-differential privacy, its mathematical guarantees by providing an explanation how much noise needs to be injected in order to preserve privacy and achieve meaningful statistics at the same time, and how to combine multiple differentially private algorithms to maintain privacy, etc. Li et al. provided theoretical foundations and practical guidance for applying differential privacy in practice [67]. Their work and the contributions from [118] ensured that relevant concepts are easily provided and gave us a good starting point for our research.

Initially, differential privacy was used with structured data sets [69, 62] however the work of Kamalika Chaudhuri and Anand D. Sarwate demonstrated the applicability of differential privacy to machine learning tasks [13]. Other scientists have also been focused on applying differential privacy to more complex data such as graphs and networks [105], images [129], streams, etc. [20]. Differential privacy in NLP tasks has been researched as well [34, 122, 50, 94]: Beigi et al. provided an algorithm that learns numeric text representations offering differentially private guarantees [3]. Feyisetan et al. also transformed text into word embedding, perturbed it with random values (noise), then replaced the text with the textual equivalence of the modified embedding [35]. His method is proved to be metric-differentially private [35]. Carvalho et al. demonstrated the use of binary word embeddings in order to reduce the cost of querying and sharing the vectors and to privatize the data before data leave a device [10].

Other scientists searched for other ways to adopt differential privacy. Many scientists came up with various differentially private relaxations and extensions in order to provide better utility and to escape the strictness of differential privacy [89]. Dwork et al. stated in 2019 that in the literature there are at least three variants of differential privacy [29]. At that point, "concentrated" differential privacy was examined [29]. Now there are roughly 201 different variants, that, according to Desfontaines, were proposed to facilitate the implementation of differential privacy to different cases and attacker models [20]. These variants are then summarized by Desfontaines and Pejo. [20]

While academia has opted for differential privacy applications, Shokri et al. provided a method to measure the privacy a model offers against membership inference attacks and showed that private information can be leaked from the original training data set used in machine learning models [98]. Carlini et al. also depicted the possibility of inference of sensitive data by attacking public language models [8]. Ding et al. proposed counterexample to test algorithms during development to figure out whether it meets the privacy guarantees of differential privacy because in their opinion many algorithms available does not satisfy differential privacy [24]. There are now tools like STATDP that can be used to detect whether

an algorithm violates differential privacy [93]. Another way to ensure algorithms correspond to differential privacy is through programming platforms, according to the scientists Ding et al.[24].

Dwork et al. created the Epsilon Registry with the intention to provide a platform where scientists and interested parties can expand their knowledge of how differential privacy could be implemented [29]. Berghel et al. presented the open-source framework Tumult Analytics - a platform for differential privacy deployment, which can be used by scientists and practitioners without expert knowledge about differential privacy [4]. The platform does not only offer easy to use interface, but also documentation and tutorials [4]. Another such programming framework is OpenDP, that aims to help researchers and practitioners to deploy and validate differentially private algorithms [38].

4. Research Methods

After the related work and the foundations have been presented, we would like to share how the research data has been collected and verified to provide a valuable analysis and results.

In order to answer the formulated research questions (1.2), this paper will rely on two steps: systematic literature review and semi-structured interviews.

4.1. Systematic Literature Review

A systematic literature review is necessary to familiarize the author with the relevant concepts of differentiated privacy and natural language processing, as well as to inform the author for the contributions of scientific experts in this field, and to gain additional knowledge that might serve for answering the posed research questions. Therefore, following the practices of [61], we searched for valuable and relevant academic papers and research. For this aim, we used the following databases: IEEE Xplore, Scopus, and Google Scholar. The search-strings, which we used, were defined based on inclusion and exclusion criteria in order to have a targeted selection of scientific papers:

Inclusion criteria

- We derived a keywords from the research questions (1.2).
- We selected papers published after 2006.
- We used backwards snowballing to identify additional relevant scientific works.

Exclusion criteria

- We excluded keywords such as graphs,trees and images ,due to the fact that a lot of papers describe the applicability of differential privacy within graphs and images, that are not relevant for our research.
- We excluded keywords such as (homomorphic) encryption and federated learning because a prevalent number of papers does describe differential privacy but very briefly, when other privacy-preserving technologies are presented as well.
- Scientific works before 2006 were not taken into account as stated in the Inclusion criteria.
- We excluded papers based on the title and on the abstract when we did not find them appropriate for our research.

The table 4.1 depicts the search strings, used for finding appropriate scientific literature, and the corresponding results in the different databases.

Table 4.1.: Selecting academic literature

Search strings	IEEE Xplore	Scopus	Google Scholar
"differential privacy" AND "nlp"	268	32	2520
(properties OR characteristics OR features OR guarantees) AND "differential privacy" AND "nlp"	253	16	2380
(properties OR characteristics OR features OR guarantees) AND "differential privacy" AND "nlp" AND NOT (graph OR image OR tree OR encryption OR "federated learning")	13	11	144
(properties OR characteristics OR features OR guarantees) AND "differential privacy" AND "use case" AND NOT (graph OR image OR tree OR encryption OR "federated learning")	66	9	337
"differential privacy" AND "nlp" AND "nlp applications"	4	3	188
"differential privacy" AND "nlp" AND "use case"	57	2	489

The first three rows from the table depict our approach to limit the number of scientific work and to receive more valuable results. After we reduced the number of results, we scanned the titles of the scientific works and selected only those, who seemed to be appropriate for our research. Then, we collected all results in a single file and got rid of the duplicates. The next step was to read the abstracts, introductions and conclusions of the selected papers to select those which are meaningful for our research objective. At the end, we ended up with roughly 102 papers. Those were read, but the ones that are most significant for our research were summarized in the chapter Related Work (3). The knowledge from these scientific works was then summarized, analyzed, and used for answering the first two research questions (1.2 and for providing a solid foundations, which were then used during the interviews.

The figure 4.1 summarizes how the literature review was conducted:

Figure 4.1.: Literature Review Process

4.2. Semi-structured Interviews

The semi-structured interviews were used to reach a targeted group of scientific experts and practitioners of the field of differential privacy and natural language processing, and to explore more deeply into the challenges and success factors identified during the literature review. Additionally, the interviews will serve as a verification of the results found during the first step of the research 4.1 and will provide us with more insights into the practical viewpoint of privacy experts and scientists.

We opted for a semi-structured interviews because according to [61] and author's opinion, they will provide a structure to the interview and might help extracting valuable information from them more easily.

We contacted 44 people in the period July 2023 - August 2023. We received 6 responses and conducted the semi-structured interviews with them via video conference, that lasted 42-43 minutes on average. We tried to formulate our interview questions in a way that participants feel free to express themselves freely without biasing their results. However, their answers were still directed by the questions we have previously formulated. The research we conducted is determined by our theoretical and practical interest and our theoretical pre-knowlegde, which was gained during the literature review. The questions we used for our interviews are presented in appendix A.

The interviewees were selected from the scientific papers, we used for the literature review - they were co-authors of the literature we used. We contacted them via their emails found on the papers. We sent them a message with information about the aim of our research, duration of the interviews and how their data will be anonymized. The people, who gave us a response, are professors, PhD students, postdoc students, and (staff) research scientists. All of them have a diverse background: some have more years of experience with differential privacy, others - with NLP; some have roughly equal experience with differential privacy and with NLP, others have only experience in differential privacy. A summary of the interview partners is presented in the table 4.2:

Table 4.2.: Interviewees and Interview Synopses

ID	*Current position*	*Country*	*Gender*	*Experience in privacy*	*Experience in NLP*	*Experience in NLP and DP*	*Communication Tool*	*Duration of the interview (h:mm:ss)*
#1	postdoc	United States	female	5 years	3 years	3 years	Zoom	0:52:03
#2	research scientist	United States	male	5 years	3 years	3 years	Zoom	0:31:19
#3	staff research scientist	United States	male	9 years	none	none	Google Meet	0:33:13
#4	PhD student	Singapore	female	-	-	2 years	Zoom	0:45:02
#5	professor	Germany	male	3 years	10 years	3 years	Zoom	0:42:57
#6	staff research scientist	United States	male	12 years	7 years	6 years	Google Meet	0:49:42

After the interviews were conducted,the recordings were played again and the interviews were transcribed. The next step was to analyse the conversations and extract meaningful value from them. The analysis aims to provide more details in case some theoretical knowledge was not well understood during the literature review and personal experience is needed in order to gain insights into the current challenges to differential privacy application within the NLP sphere. For this goal, we used the approach described in the book [6] to highlight part of the conversations which we found crucial for the research purposes - either some of the properties of differential privacy were more deeply discussed or challenges and success factors based on the experience of the interview participants were mentioned. During the interview and while analysing the conversations we were consistent and applied analytic approach [6] to clarify the misunderstanding from the literature review, to improve the results, and to introduce ourselves to real practical experiences. The highlighted texts from the transcribed interviews were extracted in a separate file, where we tried to combine statements into themes - what each of the interviewees said about each property of differential privacy, what their experience is, what challenges there are, etc. Then, we mapped these statements to the literature we used for this research, summarized the findings and the mappings in our

research paper and used the statements belonging to these themes as supporting claims.

The figure 4.2 shows how the interview analysis was conducted:

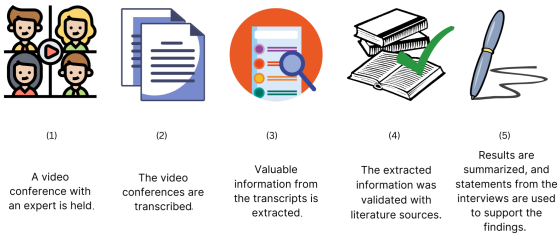

Figure 4.2.: Interview Analysis Process

After the research methodology has been presented, we would like to further discuss our main findings, limitations, and possible future research in the following chapters.

5. Properties of Differential Privacy

This chapter aims to address the first research question, defined in the introduction chapter (1.2). We were inspired by the approach of Gallersdoerfer and Matthes [40] to identify main and supporting characteristics. The supporting characteristics does not represent any requirements of using differentially private algorithms, but additional characteristics that motivate practitioners and scientists to further use and explore how differential privacy can be applied in various fields. These characteristics were presented as benefits in the previous chapter 2. The main characteristics are considered in our work as features that are needed in order to ensure that the usage of differential privacy is possible and makes sense. In summary, we searched for the properties that define the practical application setting of differential privacy. However, for the sake of simplicity and inspired by the approach of Gallersdoerfer and Matthes [40], we decided to put the found characteristics in order by categorising them the following way:

1. Strict properties that are needed for the application of differential privacy;
2. Lenient properties that might serve as a guidance whether or not it is advisable to use differential privacy;

These properties might not only help us to justify the usage of differential privacy but they may also serve as a preliminary certificate that an algorithm is a differentially private one.

After we classify the characteristics, we look how they can be appropriately mapped to NLP use cases in the next chapter 6.

5.1. Strict Properties

As a privacy-preserving technology, the core feature of differential privacy is privacy preservation, which is ensured through the addition of random values (noise) [30]. According to its definition (2.1.2), it ensures that an adversary would not be able to determine whether or not an individual is present or not in the data set: it provides plausible deniability to the individual [30]. Thus, the individual is protected from possible negative consequences that might arise due to individual's participation in a survey [117]. When an attacker can not infer whether an individual participated or not in the survey, some uncertainty about the responses of the individual remain [63]. From the definition, we can derive directly some strict properties - differential privacy requires private data, that needs to be protected, and neighbourhood relationships between two databases. Therefore, we will first discuss them in the following lines, and afterwards present additional strict and lenient properties.

5.1.1. Private Data

Private data was identified by us as a strict requirement because there is no need to use differential privacy at all, if a survey does not collect any sensitive information. Injecting random noise, when it is not needed, will only harm the utility of the analysis unnecessary, because, as stated earlier, the noise addition harms the utility [30]. To the best of author's knowledge, it might be challenging to determine which data should be protected and which do not, especially because there are a lot of definitions regarding privacy. In this paper, however, we consider any information that can harm a person as sensitive [85] due to the fact that differential privacy tries to quantify the risk of participation in a survey [30]. However, we do not want to discuss here how scientists and practitioners should define what "private" means. This may be part of other scientific work.

Private Data

As mention in the chapter Related Work 3, differential privacy was initially designed for tabular data sets and as the reader probably knows already, differential privacy tries to mask the presence of an individual in a certain database [30]. People, who observe the output of the database, are uncertain whether the individual is present or not. Thus, they might not infer personal characteristics, belonging to this individual, such as age or gender. That is true even though differential privacy tries to protect the overall information belonging to an individual, rather than specific features of this individual [59]. Thus, differential privacy might protect sensitive attributes [97].

The interviews we conducted with researchers and experts gave us an interesting viewpoint, when it comes to attributes within NLP. First of all, there is no clarity what an attribute within textual or speech data mean. If an attribute within NLP is the same as within tabular data set, then we should know how to work with differential privacy. Otherwise, we should firstly define what an attribute is, so one of the interviewee: "What is an attribute in NLP? Going from the database to language data where do we draw the analogies and what's different here". Other interviewee commented: "The whole thing with differential privacy is (that) we don't need to define an attribute. It's a worst case guarantee. So it's supposed to protect everything about a member. It's supposed to protect anything about a data record that would give away its membership. It's going to add noise to everything. So it's going to try and hide every, every attribute. So it's going to hide a style, it's going to hide all of these." According to this interviewee, only if we apply a weaker notion of differential privacy, we need to define what an attribute is and it might be advisable to consider not only the sensitive words but also the stylometric style within NLP context as something that needs to be protected.

Therefore, we first searched what does an attribute within text and speech mean: According to Wang et al, "Personal attributes are structured information about a person, such as what they like, what they have, and what their favorite things are. These attributes are commonly revealed either explicitly or implicitly during social dialogue" [114] After we clarified this, we also want to point out that when it comes to textual and speech data, not only personally identifiable information should be considered as private, but also the sentence structure might

be reveling as well [63]. The writing style might also be identifying, when knowing how a certain individual does structure his sentences [7]. Fernandes even stated: "Whilst intuition says that authors could simply mask their own writing style by choice, research has found that authors can still be identified by their stylistic traits when attempting to write anonymously" [33].Additionally, it is still unclear how a word perturbation would protect from data leakage [63]. According to Sousa and Kern even if we perturb a sensitive word from a text, the context or the surrounding information might, nevertheless, reveal the private data, which we try to protect [102]. However, this is not how all scientists believe: According to Fletcher, differential privacy assumes the worst case scenario where an attacker might have auxiliary information. Thus, an attacker is not able to infer any information from the perturbed word [37]. However, our example shows us how this could be possible: For instance, even when we modify a word that represents the exact living location of an individual, we might still be able to infer precisely where the person lives, especially if the context of the sentence describes the neighbourhood [102]. One of the interviewees also mentioned: "We are not saying that all these units of privacy (word-level, paragraph-level, user-level DP) are equally meaningful. Obviously, you would know that you're protecting the token or protecting the example. (...) If we want to apply DP properly, we want to try to protect individuals. So we want to protect users and participants in these types of data sets. So protecting tokens or examples and documents and paragraphs doesn't seem to hit the right spot."

However, if we decide to perturb all words from a text, this could be detrimental to the utility of the text data [97]. For instance, injecting random values to the whole sentence might disrupt language pattern learning, so Shi et [97]. Therefore, it is challenging to determine how many words need to be perturbed in order to ensure that privacy is guaranteed and meaningful analysis is provided, which we will also discuss in a later point. Another possible way to perturb all words and allow meaningful analysis is by generating privatized synthetic data set, so Igamberdiev et al. [54]. This view point was also discussed during the interviews we conducted with the aim to verify our findings. One interviewee said: "One easy way to get it [better utility] is to generate it [data] synthetically so you can actually fine tune an existing pre-trained language model. Pre-trained models are really good starting points and people showed that if you take differentially private SGD, fine-tune that model and generate text, the generated text is much better text. So instead of doing things like attribute based - those are very relevant pre-2020: people were looking for all sorts of things to give them this boost so that the privacy-utility trade off gets better, but now with pre-trained models that help us with good initialization, with all the public data that we have, you're actually able to get much better privacy-utility trade-offs without having to do attribute based notions, which is like really good."

Equal Privacy Guarantees

Not only does traditional differential privacy protect the individual data, but it also ensures that all individuals, represented in the database, will receive equal privacy guarantees, independent on how sensitive their own data might be [97]. This ensures, that all individuals receive equal treatment. Although, privacy leakage represents a risk for everyone, Reynolds et

al. and Baek et al. proved people do not act accordingly to their own perception about how to protect their private information and usually underestimate their privacy risks [91, 2]. Thus, it is impossible that people would have the same privacy preferences [56, 124]. The same holds, when we work with textual data belonging to different individuals. People still might have a different understanding of what is sensitive. Furthermore, when working with collocations even if they do not contain any sensitive information in it, they might give the opportunity for an attacker to infer private information such as location. For instance, "make a decision" and "take a decision" do not represent a data that needs to be protected. However, "make a decision" is used by American English speaking community, whereas "take a decision" is used by people who communicate through British English [78]. Another example is a police officer who regulates the traffic: In Russian, German and Serbo-Croatian, people use the word regulate when describing the policeman, whereas only in English, the word "direct" is used instead of "regulate" [78].

In the textual and speech domain, it is, therefore, hard to reason that all sentences or words would require the same privacy preferences. Based on the specific context, textual data might need different levels of privacy protection. Even if they come from the same domain, for instance, the medical one, it is obvious that cancer needs more privacy preservation than headache. Therefore, scientists such as Yue et al, used frequency in order to determine which words are sensitive and which not [125]. Other scientists such as Shi et al. formulated selective differential privacy - a differentially private relaxation that uses a policy function which would be chosen by each individual separately, and would determine which attributes are sensitive [97]. Additional benefit of the usage of selective differential privacy is that noise is only added to sensitive words and not to all [97]. Thus, the amount of noise is limited and utility increases [97]. This is needed especially because the addition of noise to large models could worsen the accuracy of the results, so Kerrigan et al. [59] Additionally, as mentioned earlier, injecting random values to the whole sentence might disrupt learning the language pattern [97].

Individual Data Records

Undeniably, each record from a database corresponds to a single individual. When we have a collection of documents, each belonging to a certain individual, then we can not say for sure that the sentence that individual I contributed to the database contains information about itself [7]. Often, textual and speech data might convey private information of other person or people, who do not represent the writer. Furthermore, if we assume, the writing style does identify the individual, then doubtlessly a record corresponding to a text of a single individual, does posses the private information of at least two people: "That in these conversations and emails and texts, there's usually more than one person that's involved. So we need to be very careful about how we talk about what are we protecting and how are we bounding the influence of the person to those models? ", shared a interview partner with us.

Independence between Attributes

Differential privacy treats each attribute as independent from the others. Therefore, correlation between attributes will contribute to low utility, because the injected noise will disrupt the correlation between them and affect the accuracy, so Li et al. [68]. Seleshi and Asseffa demonstrated differential privacy is not applicable to regressions. Even with small epsilon value, which should ensure high utility, the p-value is changed significantly leading to a wrong conclusion [95]. Li et al. also showed in 2011 that differential privacy does not ensure reliable privacy guarantees for the correlated data, because an adversary might gain more insights using correlations [68]. However, we might not neglect the fact that in a sentence the first word might determine what the next word would be. For instance, it is much more likely, to have the word's combination "kite surfing" than "kite swimming". Thus, words tend to have some correlation between each other [78]. In addition, word embeddings do convey some relationships with other words as well [82]. These assumptions were verified and even a better explanation was provided during the interview: ". If you're looking at it [differential privacy] at the level of words and you're looking for neighboring or similar words, then you are kind of considering those to be uncorrelated. If you look at the level of sentences though, you are considering some patterns in the words."

5.1.2. Database

Database

Differential privacy does require some notion of a database [55]. However, together with one of the interviewees, we discussed and want to point out that there is no requirement of specific database, but of some notion of a database that should be used when differential privacy is applied. As you probably remember from the previous chapters, differential privacy can be applied within diverse data sets - networks, images, streams, etc. However, as a reminder, we look at ϵ-differential privacy and how the requirements of it change when applied to textual and speech data. Therefore, when it comes to databases, we take the initial idea of tabular and static database, that was used when ϵ-differential privacy was defined.

In tabular data set, it is clear that the data set corresponds to a collection of records, where each record can be mapped to an individual and each dimension to an attribute or characteristics of that person [72]. In unstructured domain such as text and speech, a database might represent a collection of words, where each word belongs to a certain individual [63]. In this case, the formulation of the database would be similar to the one with structured data - a record would represent an individual and a single dimension will correspond to the word each individual gave to the survey researchers. However, this formulation is too strict because it can not be scaled and applied to sentences and paragraphs [63]. Another way to represent a database, is when a record corresponds to an individual's document [63] or text with variable length and tokens as attributes [97]. The disadvantage here is that dimensions can not be defined, because each text might have different length and additionally, the notion that two data sets are adjacent if they differ by a single individual or by a Hamming distance of 1 is no longer applicable [63].

Static Data Set

The original differential privacy is defined for static data set - a database that does not change throughout the data analysis or over time, no new data is added to it nor it is modified once it is created [30, 31, 63].A static data set is generally a representation of information gathered at a particular moment in time. Even though natural language processing requires a database for training and testing and this database may be static, unstructured data like text and voice recordings tend to vary over time [48]. The meaning of the words changes through the time and new words evolve over time [48]. Thus, it might be reasonable to collect data periodically, if a static data set is required, in order to ensure usability of the data.

Even though all data sets change over time - a data set needs to be updated when a data point is modified or a data point is included or excluded, language, on the other side, does require to be updated more regularly, because modifications occur more often due to cultural influence, social or historic factors, etc [75]. Tabular data set represents a collection of information from certain individuals from the population. Thus, not only language but population is also considered to be dynamic due to migration, reproduction, social factors, etc. However, the major difference is due to the fact that a value of a numerical data might change because of the time, however textual data might also change but it may still convey the same meaning as before due to lexical changes, according to Mantiri [75]. Furthermore, even though a word may not change, that does not hold for its meaning, so Mantiri: "In semantic change, the modern meaning of the word is different from the original usage."

No Access to the Raw Data

Data analysts should not have access to the intermediate operations M and to the data before noise has been added [14, 110]. Even if a data analyst wants to access the raw data in order to determine what kind of computation and statistics the data analysts should perform, and afterwards to use differentially-private algorithm, the whole process does not satisfy ϵ-differential privacy, because the choice of the analysis might reveal some private information as well, which on its own contradicts with the idea of preserving privacy [70]. "Having access to the raw data, would only put that data at more risk", mentioned an interviewee regarding the access to the data. This is so crucial requirement, especially for this decade, for the age of big data, due to the fact that a data scientists need to perform some queries in order to "look" at their data and decide what kind of computation and statistics should be performed [70]. Fortunately, machine learning algorithms do not work with raw data as input [112]. Interviewee #6 shared, additionally, according to his/her personal experience: "We should try to be very clever about how we do it. But in principle, we could use trusted execution environments". According to Habernal and Yin, when working with SGD the noise needs to be added when calculating the gradients because the training examples are only accessible in that point in time [45].

5.1.3. Neighbourhood Relationships

As mentioned in the previous subsections, definition of neighbouring databases is required in order to reason differential privacy. Even in the interviews, we find out that adjacency tend to be one of the crucial questions everyone should try to answer: "You could take differential privacy and apply it to another form of data by just really finding what a record means and what is a neighboring relation". Thus, we searched how this notion is defined within differential privacy in NLP.

According to scientists such as Wring and Rumsey, a row represents an individual [119]. Dwork also states a row stands for a single individual, where every row corresponds to a different individual [26]. Vatsalan et al. also consider a record as an individual and claim that according to differential privacy two databases are adjacent if they can be distinguished by at most one record [111]. Zhao and Chen provided another definition of adjacency, where data sets might differ at most by n elements, although most of the cases n is equal to 1 [129]. One interviewee also mentioned: "What you're trying to protect is the additional removal of a single record or one record changing. But in a lot of use cases in practice the situation is more complicated than this: You've got multiple records per person, for example. If you think of how many items a person bought on a website, maybe the person can buy multiple objects and then we have multiple records.(...) Do you want to be able to say we're hiding every single individual purchase on your website? Or do you want to say we want to hide anything related to any single individual? ". Furthermore, the researcher also shared: "I have seen papers where they protect records and it makes no sense to protect the records in that context." An example that was given was with a data set which contains the location data of individuals. Therefore, the definition of neighbourhood is different for each specific use case, which did further motivate us to map our findings to NLP use cases in the following chapter 6. According to other interviewee differential privacy also allow us to adjust the definition of differential privacy depending on what we want to protect: "It's going to be quite expensive to define units of privacy at the household level, but we could. Let's say these are texts between you and me and let's say we are in the same household. Obviously, you know, we're going to be talking about each other. And these conversations will include naturally things about your health, any issues you're facing, problems. So they can be sensitive. And just defining the unit of privacy as one person, may not be enough because these messages are being exchanged by multiple people"

Let's see how adjacency is defined within textual data: according to Fletcher et al., all occurrences of a specific term distinguish adjacent corpora [37]. Lyu at al. as our interviewees claim that neighbourhood has different notations in applications settings: We might have adjacency if two sentences differ by most n consecutive words ($n > 0$) [71]. According to Shi et al., two texts are neighbouring if they differ by n sensitive words ($n > 0$)[52]. Even though there is a meaningful definition of adjacency whenever differential privacy is applied within textual or speech data, it still remains a challenge defining neighbourhood. When we look at structured databases, it it not difficult to say how identical two data sets are. When it comes to language, we might experience some difficulties. Taking the examples provided by Cathal Horan, it is not an easy task to determine whether "cat" is closely related to "dog" or to "lion"

[51]. Even words like "machine" might have different similar words depending on the context they are in [51]. The word "file", for instance, can be used together with verbs like create, save and delete, which are only applicable in a discussion from a technical domain [78]. Adjacency can not be easily applied to unstructured and more precisely textual data. Therefore, the idea of reasoning adjacency through metric spaces.

Metric differential privacy - a local differentially private relaxation, can be used when there is no clarity how to use the adjacency property of differential privacy [63] and when the input data should be protected while useful analysis is still allowed [90]. The formula of metric differential privacy looks similar to the one of original differential privacy, where additionally the metric distance between any two words(sentences) is added [9] :

$$\Pr[\mathcal{M}(W_1) \in S] \leq e^{\epsilon d(w_1,w_2)} \Pr[\mathcal{M}(W_2) \in S]$$

Scientists such as Chatzikokolakis et al., believe that there is no intuitive definition of adjacency, if data cannot be represented as a database, but as a collection of “secrets”, and according to them, in such a case, it is appropriate to define how indistinguishable two secrets are [12]. Even though original differential privacy requires that the outputs of the analysis performed on two neighboring databases should be indistinguishable (and this should hold for structured as well as for textual and speech data), following the ideas of Chatzikokolakis et al., we require two secrets to be indistinguishable in order to allow for an intuitive definition of adjacency:

Reflexivity A text should be identical to itself.

Symmetry A text T_1 and T_2 are as identical as T_2 and T_1.

Transitivity If the texts T_1 and T_2, as well as T_2 and T_3 are identical, then the texts T_1 and T_3 should also be identical.

All these 3 relations and the distingushability notion between two texts, are easily transferable to the metric space. Therefore metric differential privacy is also used when working with textual data. Smaller metric distance means more indistinguishability. Difference between two databases because of a single individual transforms to difference between two databases because of some distance. [63], which should be carefully measured and presented, so Du et al. [25].

We also want briefly to mention that independent from the domain in which differential privacy is applied, it requires that two statistical outputs performed on adjacent databases should be indistinguishable. This way differential privacy guarantees that the contribution of an individual is hidden when its data is present in one of the databases and absent in the other one.

5.1.4. Sensitivity Measurement

In order to define how much noise needs to be injected to an output so that we preserve privacy while providing meaningful analysis, we need to know what sensitivity is and how to calculate it:

Sensitivity Sensitivity (Δf) of data set represents the maximum influence an individual I might have on an output [28].

It is worth mentioning that sensitivity does not depend on the database, but on the query we might want to perform on the database [28, 103] Thus, original differential privacy is independent from the distribution of the data [109]. Measuring the sensitivity might help to determine how sensitive a single data point is and accordingly to that to add noise in order to hide the contribution of the data point [18]. The higher the sensitivity, the more noise needs to be added [119].

Depending on the output of the analysis, we might want to use different noise injection mechanisms, that would have different requirements regarding the sensitivity:

Laplace mechanism The Laplace mechanism injects random values, taken from the Laplace distribution, to the data and it is usually used whenever the output of a data analysis would be numeric value and the sensitivity is bounded.

$$M(d) = f(d) + \text{Laplace}\left(\frac{\Delta f}{\epsilon}\right)$$

The Laplace distribution is a probability distribution, which is centred at zero and has a scale parameter b, when used together with differential privacy. The scale parameter is determined by the sensitivity and the privacy budget[42]:

$$b = \frac{\Delta f}{\epsilon}$$

Exponential mechanism The Exponential mechanism injects noise, determined by the sensitivity of the scored function, which measures the quality of the output [42]. In the end, the added noise corresponds to the maximal difference between two neighbouring inputs over the quality of the output [76]. The Exponential mechanism is usually used whenever the output of a data analysis would be non-numeric value and we have a finite range of possible options [76].

Gaussian mechanism The Gaussian mechanism injects noise, taken from the zero-mean Normal distribution, to the data. It is usually used when the sensitivity can not be measured.

$$M(d) = f(d) + \mathcal{N}(0, \sigma^2)$$

Gaussian mechanism is preferred to Laplace mechanism due to the fact the noise it produces is lower [17]. Thus, we might have higher utility. Another thing, we want to mention regarding Gaussian mechanism, is that it ensures privacy for ϵ values smaller than 1 [54]. For $\epsilon >= 1$, we might need to use Analytic Gaussian mechanism [54].

Thus, observing the presented mechanisms, we should be able to have a bounded sensitivity or finite range of possible options, which is usually beneficial when we work with categorical types of data. To facilitate the understanding of the reader we will provide some examples

that would demonstrate some rules when calculating sensitivity: The general rule for the sensitivity is that counting queries require sensitivity equal to 1 [103]. In order to facilitate the understanding of the reader regarding how the sensitivity is calculated, we will provide a simple example: If we have a database and we want to calculate how many people are male, then excluding a person from the data set might change the result of the query by 0, if the person is female, and by 1, if the person is male. As it can be seen from the definition of sensitivity, we search for the maximum influence. Therefore, in our example the sensitivity would be 1 [106]. However, if we pose several queries which might hold for the same individual (for instance, how many people are male and listen rock music), then this individual would maximum change the data set by 2, because the number of men and number of rock fans would both simultaneously change [106]. Similarly, when we work with metric spaces, the sensitivity is the maximum difference we might observe between any two outputs of functions performed on their databases [25].

As mentioned before, high-dimensional data usually is much more sensitive than a tabular data, which makes it challenging to calculate a bounded sensitivity [128]. It also might be demanding to identify all possible options from text or speech data in order to use the Exponential mechanism. This is caused by the fact that words usually have a lot of synonyms, sentences might have different sentence structure - sometimes the subject might be at the first position, sometimes at another position [128]. Additionally textual data allows people to express the same content differently - for instance wring a number with its digits or with letters [128]. According to Zhao et al., unbounded sensitivity occurs often within machine learning models [128]. However, "this is a very, very common pattern in differential privacy. You start out with a question that has unbounded sensitivity and then you bound it sort of manually", so one interviewee. The same holds independent from the fact whether we work with structured or unstructured data because ,just as mentioned several lines before, - sensitivity does not depend on the database [28].

When working with textual data, in DP SGD, the gradients are clipped so that the Euclidean norm(L2-norm) is bounded. Thus the sensitivity of the gradients is also bounded. And because we are using L2-norm, Gaussian mechanism is usually chosen due to the fact that Gaussian noise would result in lower sensitivity compared to Laplace mechanism. Therefore, less noise will be added. [45].

Unbounded latent vectors could be bounded by a clipping constant, where the representations can be clipped by their norm or by their value [54]. According to Zhu et al., “correlation between queries leads to a higher sensitivity” [131] and they propose several ways to reduce sensitivity [131]: 1) Transformation of the original database. They stated, however, that it might be challenging to figure out how to transform the database so that the sensitivity could be lower. 2) Partitioning the data set. This way, the sensitivity might remain the same as if a single query will be answered even when several queries are performed. A difficulty, here, might be to decide how the database needs to be partitioned, especially when having multiple queries and that needs to be done in advance, because inconsistency as a result of the multiple queries might occur. 3)Histograms are another way to limit the sensitivity, especially useful for original differential privacy whenever a max query will be performed [33]. 4) Synthetic

data set release also might cope with the high sensitivity. Problematic might be to provide efficiency and that usually the synthetic data set is published for a specific domain. When the noise that needs to be added is too large, there are additional approaches to combat that: Zhu et al., for instance, claim that dimensionality reduction would definitely be helpful [131]. Multidimensionality can be combated through iterative pruning and clipping by value, so Igamberdiev et al. [54]. Fletcher et al. use feature hashing to limit the amount of noise by mapping “a potentially infinite number of features into a known, bounded hash range” [37].

However, as it can be seen we need to know the types and the size of queries in advance in order to calculate the sensitivity reliably and be able to determine the appropriate noise level. According to Zhu et al, a challenge of differential privacy within NLP settings, represents the large size of the quires [131]. If many quires are to be performed on a database, more noise needs to be added: "If a recursive algorithm (...) repeats over ten rounds and noise has to be added to each round, the privacy budget will be divided into ten pieces", so Zhu et al [131]. Some possible solutions they mention are synthetic data set (but the time complexity will increase as well) and parallel composition (if it is possible - the types of quires should allow for parallel composition) [131].

According to Habernal, a query would be if “an analyst or adversary wants to get information about a data set” and queries could be numerical and arbitrary [42]. Numerical queries represent summary statistics such as mean, max, etc. [42]. Originally, differential privacy is also defined for these types of quires, according to the examples that are demonstrated in the scientific literature. In the case of NLP, these numerical queries could be to find the frequency of each word used in the text corpora [42] or find the average length of the documents [44]. Arbitrary queries, on the other hand, represent a transformation of a database [42]. Examples, Habernal gave, are to require a copy of a text corpora or the calculated gradients from a neural network, which predicts some attribute value based on given input values [42]. In DP SGD, an arbitrary query is used. Even though typical queries were presented by Habernal, predominant number of scientific papers regarding differential privacy and natural language processing do not discuss what type of queries are to be performed on their data sets and how do they decide what amount of noise should be added accordingly to the queries. So, a straightforward mapping between query in a tabular data set and in a language model was missing. Therefore, we addressed this question during the interviews and received the following answer: "When you’re training a model, you’re learning something from your data, so it’s kind of a statistics operation. So when you’re training you make sure that your training process gets noise added so that whatever you get as the output, which is your model parameters here, it is differentially private.So I’m taking gradient steps. So the data points, what they are contributing is the gradient. So I have to make sure I limit their contributions" Furthermore, it is worth mentioning, that “each query is a function f to be evaluated on the database” [84], however not every function is a query. Thus, it might be reasonable to differentiate between them. According to Habernal [42], natural language processing can apply differential privacy with queries (numerical) and functions(arbitrary). However, further specification of the queries and the functions that may be performed, will bring more light into the application of differential privacy in natural language processing.

Another major difference between pure differential privacy and differential privacy within NLP is that, differential privacy requires indistingushability between any pair of neighbouring data sets which is also known as global sensitivity [101]. Another way to represent this is that a small differences between any two databases should lead to small differences of the outputs of a query performed on the both databases. Thus, standard differential privacy is two strict for NLP settings because it allows equal privacy protection for all input databases independent how unrelated they might be [25]. An example Hu et al. give, is that global sensitivity will require the sentences "I will arrive at 4:00 pm" and "I will arrive at 100:00 pm" to be indistinguishable [52] Textual and speech data usually can be represented as high-dimensional data in the vector space. The higher the dimensionality, the more feasible it is to observe a sparse vector representations [127]. However, similar output is still required and more noise needs to be added, when the metric distance between two words is bigger [127, 25]. Thus, lower utility might be observed [25]. Additionally, small changes in the data, for instance through deleting or adding a word, can significantly impact the outputs of the query, by modifying many dimensions [127]. Thus, it is much more challenging applying differential privacy in high-dimensional data.

Zhu et al. proposes to reduce the scale of the data set in order to cope with data sparsity [130]. However, that would introduce additional noise and would lower the privacy level [130].

However, another notion of sensitivity exists, so that the strictness of standard differential privacy can be escaped: Local sensitivity requires indistigushability between the actual data set and its neighbouring databases [101, 130]. The difference between local and global sensitivity can be observed in their formulas [45]:

Global Sensitivity:

$$GS(M) = \max_{D,D'} \|f(D) - f(D')\|_1$$

Local Sensitivity:

$$LS(M) = \max_{D'} \|f(D) - f(D')\|_1$$

Although local sensitivity allows us to free ourselves from the strictness of standard differential privacy, in certain cases such as within DP SGD, global sensitivity might still be preferred because it is computationally efficient due to the fact the sensitivity is considered for the whole data set and it does not require to calculate the sensitivity for each data point. Additionally, another way to escape the strictness of the text neighborhood, is when working with domain specific texts, so that less noise is required [53].

Our interview partners also expressed their viewpoints that even though local sensitivity is a relaxation to escape the strictness of standard differential privacy, it might not be advisable to compare local and global sensitivity due to the fact they differ by their objective: the goal of global sensitivity is to provide strong privacy guarantees while the aim of local sensitivity is to inject less noise into the computation. This is something we want to point out: We do not claim that local sensitivity should be applied when working within NLP setting. However, up to this point, this technique was used to provide better utility and therefore might be

preferred to global sensitivity because as it will become clear in the following lines, accuracy and utility are also an important property of differential privacy.

Additionally, we do not only have difficulties when the sensitivity is high: Maheshwari et al. provides an example why it might be challenging to access the sensitivity of the data. For instance, they claim Lyu et al. measured wrongly how sensitive the data is, when they normalized the text vector embedding [73]. Furthermore, ADePT also has been proved to made a mistake when perturbing the data regarding to the sensitivity. ADePT used L2 sensitivity(which is computed using Euclidean distance) instead of L1 (which is computed using Manhattan distance) [44]. When we need to transfer the accurate sensitivity measurement requirement to NLP setting, it does not get easier: adding or removing a single word, might cause a variation across many dimensions and there are many possible combinations [127].Therefore, it might be even more challenging to reliably measure the sensitivity. Additionally, we already know, that words and sentences tend to require different privacy preservation and therefore, it is not so obvious what the effect on the output would be if we remove a single word or sentence [127]. Thus, the impact on the output might vary. Of course, we do not only have multidimensional noise only with high-dimensional textual and speech data. This may happen even when we work with tabular data sets. An example, one of the interviewee gave is: "If you have lots of columns, lots of dimensions, and you're having a lot of linear queries on it, often what you want to do is use the matrix mechanism, which is going to do some linear algebra transformations of your queries and then add multidimensional noise to the thing and then transform this back to the answers to your linear queries. If there might be a case to be made here that in machine learning contexts, it's much higher dimensionality than in a typical statistical context when you're adding noise to a vector with a few hundred dimensions, then this rarely happens for statistical use cases."

5.1.5. Accuracy and Utility

Another core feature of ϵ-differential privacy except privacy preservation is accuracy. This privacy-preserving technology "promises" to allow meaningful and valuable statistics and results, even though noise is added with the aim to preserve privacy [28]. Thus, it offers a measurable balance between privacy and utility. An output computed solely on random values preserves doubtlessly privacy, but does not provide any utility [28]. Dwork and Roth also claim deleting all private information does not provide any useful statistics either [27]. As stated earlier (2), the privacy budget ϵ determines the trade-off between privacy and utility and as every trade-off, this also mean we need to sacrifice the one of the two in order to achieve the other. Thus, ϵ-differential privacy makes sense to be used when precision is not vitally important for the analysis. However, the privacy budget can be adjusted so that individual privacy is protected and maximal utility is provided.

As mentioned earlier, the sensitivity of queries or analysis in high-dimensional data tends to be much higher compared to tabular data [130]. Modifying a single entry in high-dimensional data can have a more significant impact on the overall analysis, making it more challenging to find an appropriate noise level that balances privacy and utility [127]. Even if we do, it is still unclear whether a change of word with another one preserves privacy [102].Thus, in

order to find a balance between meaningful analysis and privacy preservation, it is not only required to specify the ϵ value, as in the original differential privacy but also to know how many words from the text corpora needs to be perturbed and whenever we use language models, fine-tuning differential privacy's parameters might also represent a difficulty: "If you're training model, you need to protect the model weights from inference attacks, the training data or the harder part is data publishing so there you need to find a clever way how to do that" Additionally, according to Habernal, it is still an open question whether a reasonable balance between privacy and utility in NLP might be provided due to the fact usually more noise is injected [42].

A way to find an optimization of the differentially private algorithm is to apply Empirical Risk Minimization technique (ERM) by defining a loss function to estimate the risk and trying to find a value that would minimize this risk [131]. For instance, the loss function might be the semantic difference between a private word and its perturbation. However, according to Zhu et al., higher dimensionality leads to poorer performance on ERM [131]. Thus, it might be much more challenging to find an optimal solution for textual data. "It is hard to figure out what ϵ means for textual data because the level of record you're considering and how much information is actually private from the data you're working with can all be a factor.", so one of the interviewee. Another one said: "It's definitely much harder to interpret. It's much harder to analyze what it really means."

On the other side, in order to preserve the semantic meaning of a word, thus to provide utility, several scientists such as Feyisetan, use metric differential privacy. Metric differential privacy ensures that perturbed words closer in the embedding space to the real word will have higher probability of being selected, which deviates from the original definition of differential privacy that requires that every outcome from a randomized mechanism is possible [35].

Another difficulty represents the sparsity of high-dimensional data that makes it harder to capture meaningful patterns or relationships among the features. According to Satyapriya et al., utility in NLP settings will be guaranteed if the syntactic structure is not damaged [65]. However, high-dimensional data is more prone to overfitting. Therefore, the analysis can lead to poor generalization.

Due to the privacy preservation differential privacy provides and its guarantee that a single individual could not change an output of an analysis significantly, because differential privacy tries to hide the individual's contribution, there is no use implementing differential privacy when a data analysts does want to observe outliers and not to hide their contribution [117]. Differential privacy should be used whenever the aim of the statistics is to infer something general about the population. Otherwise, it would not be beneficial to use differential privacy [117]. When it comes to machine learning, models tend to memorize the training data especially if the data is infrequently represented in the training data [63]. Differential privacy could also be used to protect against overfitting, to provide generalization [107, 132] and protection against bias, allowing the model to be more general and fair.

However, it is found that gradient clipping due to the incorporation of differential privacy might lead to unfair treatment for the underrepresented group [77]. It is however proved that using debiasing techniques together with differential privacy might increase the fairness and

the robustness against membership inference attacks[77] .

5.2. Lenient Properties

After we presented the strict properties that define the practical application of differential privacy, we would like to provide the relaxed one as well, which might serve as guidelines whether or not it is advisable to use differential privacy at all.

5.2.1. Large Size

Although differential privacy definition does not include any information regarding the size of the databases, scientists claim that differential privacy is particularly suitable when working with large data sets due to the fact that we want to infer something about the population [28]. Having large number of instances, would allow to preserve individual privacy and at the same time provide higher utility. The noise that needs to be added is relatively smaller than the one needed when using differential privacy for small data sets [117]. Wood et al claim that data sets with less or equal than $1/\epsilon$ samples can not provide utility [117]. On the other hand, when using large data sets, the injected noise is canceled out according to the Law of Large Numbers and the calculated answer diverges to the true answer [22]. Thus, better utility might be observed [22]. Practical examples such as Google and Apple which survey one dozen million individuals in order to ensure a large data set, also demonstrate that utility might be increased [124].

Large data sets are needed to train a language model. Unfortunately, large databases can also have a negative effect on the analysis - within language models they usually increase the time complexity, because especially when using differentially private stochastic gradient descent we need to add noise to every gradient calculation, which slows down the whole process [120]. "It's always good to have more data. But if you can really use it because it's slow - that's a problem. And that's the case in NLP right now.", so stated one of the interviewees. Additionally, even though noise injection is independent from the database and its size (it depends only on the type of query), the noise increases when we use differentially private stochastic gradient descent with larger size – again due to the added noise to each gradient calculation.

Furthermore, Igamberdiev et al. proved by their experiments that "a larger data set size does not necessarily mean better results at lower epsilon values, although the significantly larger data set does show the best results" [53]

5.2.2. Group Privacy

However, not only correlation between attributes might lead to lower utility or compromised privacy. Wood et al. claim the privacy risk for group of people sharing the same private information such as location increases proportionally with the number of people in this group [117]. Therefore, it is advisable to have a diverse data set in order to ensure that there is no large group represented in the database. Dwork states the differential privacy guarantee still

holds if there are several individuals who share the same private information, but the privacy loss will be maximum $e^(\epsilon * n)$, where n represents the number of individuals sharing the same private information and,of course, this is only desirable when n is a small number [28]. According to Brown et al., applying differential privacy in textual and speech data, would also decrease the privacy protection if several people share the same private information also known as "secret" [7]. Additionally, they criticize differential privacy due to the fact that a secret shared with a larger group of people would not mean that the secret becomes less private [7]. Other scientists have another view regarding group privacy: They claim that there are two things that can happen if a data set contains several documents that belong to a single individual [53]: 1) the privacy guarantee of people will degrade or 2) all documents that belong to the same individual could be merged into one and ϵ-differential privacy might be performed without decreasing the privacy protections. However, in this case, we do not assume there is another user who might have rewritten the sentence of a particular individual and provided this sentence to the database: "No one is accounting for it. You would not know it.", mentioned one of the interviewees.

Below, we present our summary table with the findings up to this point:

Table 5.1.: Summary of the results of RQ1

Properties of Differential Privacy	**NLP Mapping**
Individual and its private information	
individual attributes are preserved	individual attributes and stylometric style are preserved
random noise injection to a single attribute protects the attribute	random noise injection to a single attribute does not always protect the attribute
each individual receives equal privacy guarantees	language structures are differently privacy sensitive
each record conveys information of a single individual	text conveys information of at least one individual
individual's attributes are uncorrelated	words in a text exhibit correlation in their usage patterns
Database	
tabular dataset	different notations of database are applicable
static dataset	static dataset but dynamic language
no access to raw data	no access to raw data
Neighbourhood relationship	
intuitive definition of adjacency	complex definition of adjacency due to semantic similarities
indistinguishable outputs	indistinguishable outputs
Sensitivity measurement	
sensitivity needs to be bounded or constant (Laplace mechanism)	sensitivity is prone to be unbounded
finite range of possible options (Exponential mechanism)	finite range of possible options is challenging to find
advanced knowledge of queries	advanced knowledge of queries
global sensitivity	global sensitivity might not be appropriate
lower and higher dimensionality	usually higher dimensionality
Accuracy and utility	
meaningful analysis and privacy preservation should be provided	meaningful analysis and privacy preservation should be provided
balance between privacy and utility is challenging to find	balance between privacy and utility is much more challenging to find
used for generalization	used for generalization
Size	
large datasets provide better results but increase time complexity	large datasets provide better results but increase much more time complexity
Group privacy	
privacy guarantees decrease proportionally to the group size	privacy guarantees might decrease proportionally to the group size

6. NLP Use Cases

After looking at all the requirements, we searched for use cases in the papers and even asked interview participants to share their thoughts regarding the applicability of differential privacy in NLP use cases. Due to lack of scientific literature that describe the usability of differential privacy in NLP, we then decided to have a look into different NLP tasks, how they can be mapped to the use cases and the properties of ϵ-differential privacy, discussed in the previous chapter. To fulfil this goal, we looked into articles and scientific papers but we mainly used the work of [1] and whenever we found a certain task that is not too specific we decided to use this task for our analysis as a broader application where differential privacy could be applied. All findings are presented below:

6.1. NLP Use Cases

6.1.1. Training Models

Differential privacy might be used for training language models [52]. Differentially private stochastic gradient descent, for instance, is used on several language models [63]. The aim of the DP-SGD is to hinder the memorization and leakage of sensitive information, which are due to the training process [90]. Training a language model with differential privacy might hinder extraction attacks although it increases the time complexity and the utility get damaged [120]. Additionally, differential privacy should be used whenever the aim is to protect from privacy attacks. It might protect from mitigation data poisoning attacks within text classifications, according to Xu et al. [121], and from word substitution attacks, according to Wang et al [113]. Fine-tuning language models on sensitive data set might be another typical use case of differential privacy in machine learning [79, 52]. Additionally, crowd sourcing on sensitive data where labeled data set is required might be another use case of applying differential privacy in NLP, so Mouhammad et al. [81].

6.1.2. Public Release and Model Inference

Another NLP use case where differential privacy can be applied is for public release of data [63], of model parameters [79] or of a pre-trained model when the training data does contain sensitive information [45]. The same holds, when n-grams are publicly released [60]. A practical example might be whenever a user sends word embeddings for NLP processing in order to receive translation or classification result, or to check the grammatical correctness of the words [69]. An attacker might try to extract private information when the vector representations of the words are shared with other parties. Another example

might be whenever textual data such as doctors' notes, online reviews, social media posts or governmental records are released to third parties [63]. However, it is worth mentioning, that not only public release of a model is a valuable use case of differential privacy in NLP, but also there is a need to perform differential privacy within NLP whenever the aim is to have model inference [52]. One interviewee mentioned: "I don't know if anybody is actually taking these prompts, centralizing them and training on them. It's possible. I don't know. But when we start fine tuning or training and retraining on these prompts or doing any form of reinforcement learning human feedback in the loop, that's when I think DP would become interesting, even for the base pre-trained models." Unfortunately, there is a lot to be explored in this direction [52].

6.1.3. Sensitive Data Availability

Prevalent is the requirement of personal data that needs to be protected or as Carlini et al. and Song et al. have formulated it: differential privacy does make sense to be used in NLP when there is a potential risk of information leakage or unintended memorization [8, 100]. Researchers believe that whenever there is a private data that needs to be protected within textual data differential privacy might be used. Examples, Brown et al. gave, are use cases in which natural language is used that might contain sensitive information: call centers, medical applications, voice assistance, message auto completion and document translation [7]. Additionally, more than one interview participant mentioned a typical use case is in the medical domain in order to protect patient privacy. It is worth mentioning that some researchers still believe that would not be the perfect use of differential privacy because "in one extreme case, it was reported that the use of differential privacy to protect patient privacy in a medical experiment would have resulted in a number of deaths from over-medicating, due to the noise introduced by the differentially private mechanism" [33] Even though the performance of NLP models might not be satisfactory, scientists believe it is still a better idea to apply differential privacy than having no technology to protect data, but of course after careful consideration of its usage: "I would say that applying differential privacy to a sensitive application is still better than doing nothing."

6.1.4. Interaction with Language Models

Another use case when differential privacy and NLP can be used is when users interact with a natural language models [36]. There are two types of interactions: centralized where users upload data to a centralized model, and decentralized where computations are performed on a local machine [63]. Decentralized interactions might be usually preferred by users whenever their data is used to train and serve a model [90]. Therefore, depending on a specific use case, local differential privacy might be applied whenever the input data needs to be perturbed locally before being sent [90]. It provides strong privacy guarantees for individual data points, making it suitable for scenarios where preserving the privacy of each person's data is important. [90]. Otherwise, we might apply global differential privacy - perturb the data points when they have been already collected [90].

6.1.5. Indistinguishable Texts Generation

We might also need to use differential privacy when we want to publish anonymous reviews so that readers can inform themselves from the content but not infer who the author of the review is [115]. Panchal explains how differential privacy can be used in order to generate indistinguishable text from the original so that an adversary can not distinguish what is the real message [88]. Igamberdiev et al. claim rewriting texts is another use case where differential privacy can be used: "One particular method of applying DP to the domain of NLP is differentially private text rewriting, in which an entire document is rewritten with DP guarantees by perturbing the original text representations" [54].

6.2. NLP Tasks

6.2.1. Information Extraction

Information extraction represents a task that would extract keywords from a given input text [66]. To this task, we include named entity recognition (NER), conference resolution, part-of-speech tagging and text summarization. There might be other tasks that could be included into this section as well. However, we will stick to those mentioned in order not to obfuscate the reader.

Named entity recognition is a subtask of information extraction that involves identification of "names" and their classification to a predefined categories such as names, locations, organisations, countries [74].

Conference resolution is a subtask of information extraction because it might help us extract additional information belonging to a certain entity. Conference resolution identifies the relationship between nouns and pronouns and the corresponding entity [66]. For instance, in the sentence "John went to the supermarket even though he is injured", "he" refers to John [66]. That might help to gain the additional information that John is injured.

After briefly explaining some information extraction tasks, we would like to map them to our results up to this point 5.1. We want also to mention that named entity recognition and conference resolution rely on part of speech tagging in order to analyse the text and this will be discussed in the following subsection. However, what we want to point out here as a take away whenever we want to apply differential privacy to some information extraction tasks is that we need to define a database and what adjacency would be. The main challenge comes from the data collection process and how would we define which contribution belongs to a certain individual in order to protect the individual. More challenges are to be expected if the collected data is web data due to the fact a lot of information online is anonymous, so one interviewee: "If you collect data on the web, then (..) you don't have ownership. There's paragraphs on Wikipedia, but if it's Stack Overflow or Reddit, there's always an author". Whenever we have an authorship it might be easier to define user-level differential privacy because a user might contribute more than once to the data analysis process. However, when it comes to text summarization - a task that firstly identifies relevant information from an input text and secondly based on input, creates a summary, also known as text generation (where

we can use differential privacy as it will be discussed in the following subsection) - we might have single or multi-document input [99], which might determine how a database should be defined as discussed in the previous chapter 5: whenever we have a single-document input, then each individual should have contributed some corpora to this document. When we have a multi-document input we need to clarify first, whether each document belongs to only one individual and whether there are several ones that can be associated with a certain individual. Additionally, if we assume that we work with word-level differential privacy, then, of course, the stylometric style and the surrounding context might still identify the individual who provided the data. Here, we want to refer back to the example we gave in the previous chapter 5: if we have a word such as place of work or place of residence, the context might describe this place and reveal the word even if the word is perturbed. This might hold every time, when we apply word-level differential privacy independent from the use case. What we also find intriguing during our research, was the fact that we were not able to find out what was feasible and infeasible to expect from a perturbation method. This would make conference resolution challenging if the correlation between the pronoun "he" and the pronoun "she" is high and the pronoun "he" is substituted with the pronoun "she" - the conference resolution task would not possess high utility, and as mentioned earlier 5.1 meaningful analysis is other property of differential privacy that we need when we apply this privacy-preserving technology.

6.2.2. Semantic and Syntactic Analysis

Semantic and syntactic analysis is used to identify the structure and the meaning of a corpora in order to ensure that a machine would be able to understand [1]. Part of speech tagging is another important NLP task that might help to understand the input data by mapping each word to its grammatical label [1] and thus might help to understand the sentence structure and the corresponding meaning. Thus, it might contribute to conference resolution and named entity recognition. However, we want to point out that defining how similar two words are, might represent a challenge because metric differential privacy may assume that a word with grammatical label as noun might be highly correlated to a word with other grammatical label such as verb. For instance, burglar and shoplift or shoplifting. If that is something that is feasible, then we might have difficulties performing part-of-speech tagging whenever we apply word-level differential privacy. Word sense disambiguation is another NLP task that might help to understand the meaning of the words [1]. To fulfill this task, we should check the words in a dictionary or the meanings of the words should be learned during training [66]. However, there might be an obstacle to perform again word-level differential privacy before word sense disambiguation task because without replacing a word with a word with similar meaning, the word might be identified from the context: If we have the sentence: "I recommend you this book" and we change only the word "book" but we do not know that the word "book" is a noun, we might perturb the word "book" with another one such as "reserve" that would be highly correlated to the word "book" whenever both are considered to be verbs.Then the perturbation might result in "I recommend you this reserve". The problem arises here because as a human being we are able to infer that the sentence "I

recommend you this reserve" does not make sense and if we assume that people who apply word-level differential privacy would try to protect only sensitive attributes, we would be able to infer which word was perturbed and then search for highly correlated words such as synonyms or antonyms. Based on the findings, we might be able to infer the real word "book".

Additionally, we might have the same challenges with semantic and syntactic analysis as we do with information extraction - how database and neighbourhood will be defined depends on the data that we have collected and the way we proceeded while collecting this data.

6.2.3. Interactive Systems

Within interactive systems, we would discuss question answering, text generation, and dialog systems. Question answering is a task that provides answers to questions that are posed by the users who use the interactive system [66]. Text generation is a task that can be used to create a text or to provide the corresponding answer to the user in question answering. Problem we might have when incorporating differential privacy into question answering is that sometimes we need to memorize some data points, which differential privacy makes difficult, because it tries to hide each contribution. As one of the interviewee mentioned, we might need to memorize data point especially when they represent general facts that are rare events. For instance, giving the location of an endemic animal or plant should be information that should not be perturbed because it will provide wrong information, which is not the aim of question answering. A consequences of false information might lead to endangering the special animal or plant, which none of us want to do. Another example the interviewee gave is: "if I'm doing like closed book Q&A, I need to memorize facts: if I see that Mount Everest is eight kilometers tall, that is something that even if it's repeated once, I want it memorized. And the opposite is also true like again, if something is repeated many times or if there's a pattern, you usually want to learn it. But again with text that is not always the case. Let's say I have a lot of copyright material that is repeated many times: people showed that parts of the Harry Potter book are like repeated so many times that if you prompt a language model with like the first paragraph, it's going to spit out the whole chapter. So that is actually like something that no matter how many times it's repeated, I don't want it to be learned. So again, DP cannot help with that.", so one of the interviewees.

Q&A is a NLP application that could be used for model inference as mentioned in the previous section 6.1 and on its own is highly related to text generation. Text generation is another NLP task that is used to compose natural language text given some input and as one of the interviewee mentioned: "You can already with just some prompting, prepare it [the model] a little bit by giving it some examples, some example prompt with an answer. It can already do really well. It can write emails suddenly beautifully. It can write text messages and documents (...) . And those prompts can be sensitive". An example that might serve as a reference to the next paragraph is that during our interaction with language models like ChatGPT or BART, we might provide private information such as our health status and interests and expect the language model to provide us with a valuable answer.

When it comes to interactive systems, usually we also speak about dialog systems instead of types of interaction, where differential privacy can be applied. Brown and her colleagues presented the challenges when trying to preserve privacy within a chat: a chat participant might also reveal private information even though he/she does not mentioned anything sensitive, but his/her text represent a reference to what the other chat participant wrote [7]. Here, again, word-level differential privacy might not protect the stylometric style and the sensitive information as well due to the references the other chat participants can use. However, it is worth mentioning that if we try to use differential privacy for chat systems, then we do also need to define a new notion of a database - collection of documents belonging to single individuals is not representative for chats. Collections of chats each provided by single individual might also not be a representative database, because within a chat there are many people who interact with each other and as it holds for textual data - in their conversations people may discuss other individuals as well. Therefore, we discussed this issue with our interviewees and one of them shared the opinion that we should apply differential privacy at different level - for instance, household-level instead of individual-level - as discussed in the previous chapter 5. The interviewees also shared that we might have interactions with digital virtual personal assistants with whom we might again share sensitive information but we might also want that these personal assistants to interact with each other: "I want my assistant to be able to have a conversation with yours, but not necessarily reveal the information that it has learned from other conversations that it had with my family". These as mentioned for model inference, requires further exploration and up to this point should be considered as an open question [52].

6.2.4. Machine Translation

Within machine translation, where the aim is to translate one language into another [66], differential privacy may also make sense. If we collect pairs of languages that are provided by individuals, then we might infer the presence of certain individual by their choice for translation, because text and speech can not always be directly translated - we need as translators to find a proper way to use different words but to convey the semantic meaning [11]. Word-level differential privacy should not be considered possible in this case due to the fact languages do have a lot of idioms or phrases that can not be directly translated [11] and if we perturb only a single word, this might disrupt the semantic meaning which the words together convey. Another challenge that comes when scientists have to work with textual and speech data, according to one of the interviewees, is that no one assumes or can state for sure what privacy is guaranteed to people who does not use the official language but rather some dialects or who as foreigners do not speak the standard language 7. This viewpoint is particularly important for machine translation because:

- First of all, the individual's private data and his/her contribution need to be protected independent from the other individuals who decided to participate in the data collection process. The risk that an individual with Arabic dialect from Morocco and his/her private data would be exposed if other individuals who speak Arabic are from Iraq,

should be minimized [11].

- Second, as discussed 5.1, language tends to be dynamic - the meanings of the words change rapidly over time and additionally how a word changes has different meanings: a word might be unchanged but its semantic meaning can be different in the future, or a word might change without any shift in its semantics, or both - the spelling and the meaning might change.

Here, again we need to define what neighbouring databases mean. We might consider databases as neighbouring if they differ by a single user (however as discussed in the previous sections that depends on the data collection process) or if they can be distinguished by a number of consecutive tokens. An interviewee mentioned: "What you do from $D1$ to $D2$ is to change one sequence of tokens - this is what we call example-level differential privacy. But if you have the user information, you can also talk about user-level DP. And what should be done it to change all the training examples that belong to a user." However, latter in our discussion, we came to the conclusion that example-level differential privacy would protect the tokens but not the individual for sure: "If we want to apply DP properly, we want to try to protect individuals. So we want to protect users and participants in these types of datasets. So protecting tokens or examples and documents and paragraphs doesn't seem to hit the right spot."

6.2.5. Sentiment Analysis

Sentiment analysis analyzed the textual and the speech data in order to recognize the intend and the emotion the owner of the text has [66]. Thus, the subtasks might be emotion detection, hate speech detection, intend recognition, opinion mining, etc [1]. Applying differential privacy on a word level comes again with the challenge of similarities - antonyms might be highly related. Defining antonyms on a word level might also be a challenge - first of all because antonyms might be represented with the same word + not and still be highly related. Second, it is challenging to determine which word do you need to perturb in order to preserve privacy, when it comes to statements that express emotions. People might express their feelings and emotions with other words, and idioms instead of directly mentioning their emotion. For instance, people might say "I am over the moon" instead of "I am delighted". Vogel and Lange demonstrated in their work [112] that differential privacy can be applied to sentiment analysis tasks on Twitter. Based on their work, we were able to conclude that they might work with user-level differential privacy even though they did not directly mentioned it but we assume that it does make sense because each Twitter user might be distinguished and we might know which post or contribution belongs to which individual. Additionally, Vodel and Lange got rid of duplicates and retweets [112], which we mapped to our research as a technique to ensure that each user contributed to the database only once and to limit the challenge that textual data might convey information about other individuals. This way, we would say, people might provide at least some guarantees that try to ensure that each user's privacy protection would not decrease unintentionally and each user's contribution would represent information about the user providing this information. However, retweeting user A

directly is not the only possible way to convey the information about user A - people might also reformulate user A's statement. And here, we want to point out that there might be other posts that people need to handle in order to ensure better privacy-preservation. Another thing we want to discuss here, is that there is no clarity whether if we reformulate the statement of user A, the statement should be considered as ours. This distinction comes from the need to protect the sensitive data as well as the stylometric style of the individual. And, even though differential privacy should protect everything because it masks the individual contribution, it is not clear in this case who is the individual that is protected - is it user A or we who reformulated user A's statement?

After the extensive description of the properties of differential privacy that can be used to define its practical application setting and the mapping of this characteristics to NLP use cases, we found out that these properties are not clearly defined when differential privacy is used together with NLP in practice because the way they are formulated depends on the use case - how the data was collected and what we want to perform. Additionally, as we saw from the interviews, the trend in differential privacy in NLP is nowadays in training language models in a differentially-private manner, and generating synthetic data, which still needs to be further explored. Now, we would like to discuss our last research question and the corresponding results in the following chapter.

7. Challenges and Success Factors of Differential Privacy in NLP

As promised, in this chapter, we would like to present the current challenges and success factors of differential privacy within the NLP sphere. The findings were collected from systematic literature review which served as a basis and mainly from discussions with experts in this field through semi-structured interviews.

7.1. Challenges

7.1.1. Communication and Transparency

As mentioned in the chapter Foundations 2, many scientists summarized and structured in a clear and easy to follow way the foundations of differential privacy and created platforms that allow newcomers to understand and use differential privacy. The reason for these educational materials and platforms has arisen due to the identified challenges of implementing differential privacy: there was no clear explanation of differential privacy for non technical audience and there were communication difficulties between people coming from the computer science field or privacy field and other employees due to the fact they possess different knowledge [29]. Thus there was a gap in their understanding what differential privacy can do [29]. In addition, people to whom differential privacy is being introduced, tend to be skeptical at first [29]. We mentioned earlier, we are motivated to figure out whether these educational materials and platforms do create value in practice.

The findings from the interviews do show that privacy experts believe these educational materials and platforms do help to facilitate the understanding of differential privacy to non-technical audience. However, our interview participants also shared that people, who might possess already the basic understanding of what differential privacy is, still might experience difficulties when they start exploring the field and trying to go more into depth of the sphere. An interviewee shared: "Let's say, master's or new PhD students interested in this topic. They ask (...): where should I start? What should I take to get into that and get really familiar to the depth of everything? And then you run into issues." According to the interviewee, these students do possess different educational and practical experience: they might study linguistics, cryptography, computer science, etc. Another scientist did also mention: "People are scared of differential privacy. Especially NLP folks are not super open and not because they don't like it. It's just that they feel like there's too much theory and we don't know". Another problem concerning the understanding how differential privacy might be implemented do also come from the lack of transparency and reproducibility, that

according to Igamberdiev et al.,may be a challenge to adoption of differential privacy [54]. Therefore, they came up with DP-Rewrite - an open source framework with the goal to be a solution of these problems(thus, to provide transparency) [54]

7.1.2. Privacy-utility Trade-off

As mentioned in the chapter Properties of Differential Privacy 5, it is hard to determine what the privacy-utility trade-off should be. However, because we would not like to repeat ourselves, we would skip the explanation of why this should be considered as a challenge. We advise the reader to return to the chapter "Properties of Differential Privacy" 5 if more explanation to this part is needed. Here, we would only provide additional challenges that arise around the privacy budget and that were mentioned during the interviews: "It [differential privacy] really, really impacts performance more than applying differential privacy to tasks with tabular data. So I would say the biggest challenge would be finding a way to apply it at the appropriate granularity". Another researcher did also reveal: "The problem (...) is that it slows down. DP SGD is much slower by a constant factor than normal SGD." Another difficulty that was identified was the complexity of textual and speech data: "You have data that is very heavy tail and you see many things but few times. So you end up smoothing all of it over and you end up with very bland language if you just use differential privacy. So, you get very safe conservative sequences that have no taste, no diversity." One of the biggest problems regarding the privacy budget (privacy-utility trade-off) is that, as we defined in the previous chapters 5, differential privacy makes sense to be used when there is individual and private information that need to be protected but we might not have a comparable non-DP model because we might not be allowed to analyse the private data that is within this non-DP model [16]. Thus, assessing how utility is impacted would represent a challenge [16].

7.1.3. Semantic meaning

Another challenge that might be observed while using differential privacy in NLP is that the state of the art of technology still is not able to understand the semantic meaning of textual and speech data as we as humans can do. According to an interviewee, there is still a space that needs to be explored in order to ensure that two sentences are really similar: "There are things that aren't captured by that similarity and sometimes that similarity will miss things like the "NOT" case, because sometimes synonyms and antonyms are ranked very similar because they're related, but that's probably the most promising direction."

7.1.4. Private Data

It was also ubiquitous during the interview that there is no clarity how private and sensitive data is defined and actually "it's an arbitrary decision", so one of the interviews. Based on the conversation with the interviewee, we came to the conclusion that first of all, it is challenging to determine what is sensitive: "what is private and what is public itself is like a philosophical

question", so other interviewee. Second, how would we identify these private information and third, what would we do with the other data that was not identified but still is considered sensitive and how sensitive this information actually is. Fourth, who do need to decide all these questions so that all the involved parties are satisfied.

7.1.5. Data collection

Not only the boundaries between private and public data represent a challenge to adoption of differential privacy, but also the collection of the data itself should be considered as a difficulty because:

- Researchers do try differential privacy with publicly available data. Nobody will provide its private information just to give researchers and practitioners the opportunity to test whether a privacy-enhancing technology would preserve their privacy and when there is no such technology available that would ensure that the sensitive information will be provided to the researchers and practitioners, so one interviewee: "You don't have private data so how can you prove that your methods which are applied to movie reviews we'll actually work on this kind of sensitive data"
- Even if practitioners and researchers already possess private information, for instance because the customer allowed his private data to be used, we face the problem that the size of the available private information is small, which on its own may represent a challenge for training a language model and as mentioned earlier, larger data sets might provide better utility, which is not the case with small private database. And again, we might end up with slowing down the training of the model if we have large data sets, so one interviewee.

Another interview partner mentioned: "A lot of them [models in NLP] are actually trained on web data, on public data." but we also discussed that additional challenge represents the internet data because it is not clear what is public and what is private, but also it is not clear due to copyright issues who contributed this data or even some information might be provided by anonymous author. So how can we ensure that we really protect the individual in the database, if the models are trained on web data, where which data belongs to which individual is not so straightforward.

7.1.6. Performance

The slow down of the training is not desirable. Expersts into NLP shared that the performance of the model is impacted whenever differentially private SGD is used. As one of them said: "Training those models [large language models] alone is very expensive but training a model with differential privacy is even more expensive because adding noise to the data increases your costs." Nobody wants to have high costs. Therefore, everyone would rather try to control these costs [16]. As mentioned in the paper [16] and in one of the interviews, DP SGD can not be parallelized because the gradients are clipped per example. Thus, we might not take

advantage of the GPU parallel computing and even if there are ways to circumvent this challenge, additional work and expertise might be needed [16]. On the other hand, even if we can afford the cost and the additional work, the slow process might be a hurdle: "So if I have huge amount of data that it's going to take me three weeks to train on, even if I'm not using differential privacy, with DP it might take me a month, it might take me two months. So it might not even make sense to do it like that"

7.1.7. Application of Differential Privacy to Specific Use Cases

Lastly but not least, the experts who participated in the interviews, determine the connection between differential privacy and the specific use case where we want to apply differential privacy as a major challenge. Even scientific papers such as [16] demonstrate that people might know what differential privacy can be used for, but still the people, who need to make the valuable decisions such as what should be protected, what is the risk we should accept, what should represent a record, how would neighbouring databases will be defined, etc. experience difficulties. Thus, their thoughts helped us not only to verify our research findings and current challenges, but they also motivated us even further because their views demonstrate that our research might be valuable for researchers and practitioners. Below, we provide what some of the interviewees said regarding this challenge:

- "I mean everything depends on the use case."
- "Where the challenges really are and I see a strong need for tackling that is the very definition or the very promises of differential privacy, like coming from the databases and then the actual things like in a data set"
- "And the reason is that you don't have atomic data: in a table, you can see the rows. It's either duplicated or it's not. It's very clear what you're doing. But the text is like: do I look at a sentence or do I look at a paragraph or do I look at a document?(...) So as many times as someone else's data is repeated in another person's, their like epsilon is getting degraded. (...) But by default, you don't know that this is happening. So no one is accounting for it."
- "Somebody write some text regarding two-three other people. What does this mean to protect individuals in the data sets? Are you protecting just authors? Are you protecting the people whose data is being discussed in the text?"
- "What is [the privacy] unit x? And this is very much an application-dependent unit"

7.2. Success Factors

There are organizations that already use differential privacy as tool to preserve privacy. Such organisations are:

7.2.1. Apple Inc.

Apple Inc. uses local differential privacy to gain insights how to attract their customers by what they are doing and simultaneously preserve their individual privacy by perturbing their information locally - before their data leaves the customer's device [21]. Apple also uses differential privacy in its iOS 13 to protect the private information of users, while their digital virtual personal assistant is being trained [47]. Deletion of device identifiers and encryption mechanisms are additionally used to ensure privacy [21]. It is important to note that Apple limits the number of contributions per individuals so that they do not experience privacy degradation [21] as mentioned and discussed during the first research question. Apple also tries to keep the privacy budget relatively small, for instance 2 for Health Type Usage and 16 for Quick Type Suggestion (where a model tries to predict the next word a user wants to type based on previous words, context and writing style used)[19], and to collect a large amount of data so that the utility increases [21]. Thus, Apple takes into account the two properties that we identified as attractive for practitioners to use differential privacy. The database they use is defined as a collection of user input, which is represented by a limited number of bits or limited information that before noise addition has been encoded via a hash function SHA-256 [21]. The definition of the neighbouring data sets, is defined by how the company would determine what a record - privacy unit represent: Apple uses user information per day[19]

7.2.2. Google

Google also has "Community Mobility Reports", that measure how much time do people spend on public places and at home during the COVID 19 Pandemic, while ensuring privacy preservation by incorporating differential privacy [19]. Another successfully differentially private implementation of Google is their privacy-preserving technique - RAPPOR (RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response) that allows collecting locally perturbed values and strings to be used for statistics without compromising privacy [32]. Interestingly, RAPPOR might work with data that may periodically collect information for the same individual [32]. Another successful implementation of differential privacy from Google is the Gboard - the Google keyboard with features as predicting the next word a user might want to use, smart compose (generating inline suggestions to speed up typing), and On-the-Fly rescoring to create a list of potential word candidates during typing [123]. In this case, Google uses more than 20 language models and a combination of differential privacy and federated learning - another privacy-enhancing technology that preserves privacy by training models in a decentralized fashion [123]. For the training of the model, the company applies DP-FTRL - algorithm similar to DP-FedAvg, which is variant of DP SGD for user level differential privacy, which does not require uniform sampling [123]. User-level DP is applied and the user's participation is limited here as well [123]. In their paper, they claim that larger data size improves the utility of the Gboard task [123]. Another thing the company is doing is to pre-train their model on public data with the aim to increase utility [123]. Their differential privacy implementation seems to be successful, so one of the interviewees: "Those are successful implementations and people build upon that [pre-training language models

on public data]. People want to do it. They try and find good pre-trained models to do it." Other usage of differential privacy together with federated learning provided by Google is for Smart text selection, where users might receive a suggestion of texts that were previously copied by them [49]. During the interviews, the implementation of differential privacy in Google and Apple were also mentioned, thus verified to be successful.

7.2.3. Microsoft

Microsoft is another company that tries to incorporate differential privacy in several applications such as top-k results and histogram releases in LinkedIn, where user-level DP and adaptive quires are taken into account [92], but also for telemetry data collection such as app usage statistics [23], which was incorporated in Windows to improve user's experience and to solve application's issues [5]. Microsoft proposed in 2021 differentially private n-gram extraction - extremely suitable for text generation and sentence completion [60]. The company defines neighbouring databases that differ by a single user [60]. The database is a collection of Reddit posts [41]. Additionally, Microsoft also mentioned that their differentialy private algorithms have been used in practice for generating replies in Office [5].However, according to the company differential privacy is still not prepared to be used for a commercial product and research is needed to shed light into the mathematics and the computations involved, but also into the privacy regulations that need to be addressed [14].

8. Discussion

The aim of this chapter is to provide an objective overview of this research work.

As the reader already knows from the chapter Introduction 1, we were motivated to shed more light into applying differential privacy in NLP use cases. Therefore, we tried to identify the properties of differential privacy that define when does it makes sense to use this privacy-preserving technology and what are the requirements of differential privacy that need to be fulfilled in order to use it in practice. As a summary of our findings regarding these characteristics of differential privacy, the reader may have a look once again into the chapter Properties of Differential Privacy 5 and more precisely into the table 5.1. Our main findings are that it is meaningful to use differential privacy whenever

- there is an individual and private data that need to be protected
- we want to protect that data through addition of random values but still being able to analyse these data and to infer some statistics

and in order to apply differential privacy we would need

- a properly measured sensitivity (as in many works, wrongly calculated sensitivity led to wrongly classifying an algorithm as differentially-private one [44, 43])
- notions of database and adjacency that depend on the use case: whether we work with data gathered from the internet or data that was collected from individuals. What we want to protect would define our adjacency definition and the granularity at which we apply differential privacy.

In the chapters NLP Use Cases 6 and Challenges and Success Factors of Differential Privacy in NLP 7, we depicted, for instance, that companies like Google, define user-level differential privacy to protect an individual's participation in a database - in their use case it is even more easier to define user-level differential privacy because they combine differential privacy with federated learning where each user trains a model with his/her local data and the model parameters are then send to a central model [123].

Additionally, for this work, we defined two additional characteristics that we identified as appealing to practitioners who want to incorporate differential privacy - large data sets as the utility is less damaged and diverse data sets that would ensure that a group of individuals would not receive less privacy protection.

Then, we tried to provide different NLP use cases where differential privacy can be applied and tried to explain how the previously formed properties were met and how this privacy-preserving technology has been implemented successfully and what challenges may

arise or may have arisen. With our answers to the formulated research questions 1.2, we tried to facilitate the adoption of differential privacy in NLP in practice by making the requirements/properties of differential privacy and the challenges of incorporating it more clear. We also hope that our research helped for bridging the research gap we identified 1 - thus, making the connection between pure differential privacy and differential privacy that is used in NLP more straightforward. Additionally, we hope we helped people understand what do they need to consider when they want to apply differential privacy to a specific NLP use case [16] because according to scientists such as Cummings et al., this represents a challenge of applying differential privacy in practice [16]. Of course, in order to provide an objective overview of our research work, we need also to outline both the limitations of this study and potential avenues for future research that may emerge from our findings.

8.1. Limitations

The focus of the literature we examined was predominately concentrated on textual data within NLP and only a few of the scientific papers we read discussed speech data as well - another vital component of NLP that has its role as well and should not be underestimated. Furthermore, it is important to note that all the papers reviewed were exclusively in English, potentially restricting the number of scientific literature that could contribute to this research. However, we should not exclude the fact that English was a preferred language due to the fact that a lot of work regarding the topic is written in English.

Another constrain we would like to admit is the limited number of interviews conducted, which makes it harder to gain a more objective view of the usage of differential privacy within NLP. Unfortunately, due to the scope of the interviews conducted, most of the challenges and the use cases are subjected to the personal experience of the interview participants. Additionally, even though we aimed to address researchers and potential practitioners, the interviews were mainly conducted with scientists doing research in this field, rather than practitioners who use the technology in their businesses. Although we can not deny the invaluable insights gained from the researchers, the absence of the view point of practitioners introduced a significant imbalance in the study's perspective because practical experiences and real-world challenges might not have been discussed in their depth and breadth. As a consequence, the broader practical challenges that practitioners often encounter in the field were not fully presented, thus offering only a partial understanding of the challenges of implementing differential privacy within NLP in practice.

We also want to point out that a tutorial regarding privacy preservation within the NLP sphere in practice might have had a huge contribution to our work. However, the tutorial was not available during the time of doing the research and of summarizing the results in this document [42].

Lastly but not least, we would like to point out that our research was focused on differential privacy, thus a lot of additional steps that are needed in order to preserve privacy were not investigated and not discussed in this paper. However, it should remain clear that differential privacy is only one of the things that need to be fulfilled in order to preserve privacy.

8.2. Future Work

As mentioned in the previous chapter 5, perturbing a single word from a text does not protect this sensitive word that could be identified from the context and it might not protect the individual due to the fact the individual might be identified by its stylometric style on its own. Even though, there are research groups who investigate differential privacy not on a word level, we believe there is still a need for further investigation into applications of differential privacy in NLP under another granularity such as sentence DP or paragraph DP or under another approaches such as text rewriting. Scientists who participated in the interviews even stated for their research on text rewriting: "I don't think it's a success story yet." Another future work is coming from the fast improvement of language models like ChatGPT. There is a lot to be explored in this sphere: What could we do to preserve the private information that is typed by the users using ChatGPT? Would differential privacy be efficient to protect all language models that were not trained and fine-tuned, but were generated by prompting ChatGPT or prompting at all?

Up to this point, it seems like differential privacy and NLP are working with free word combinations, which allows to replace a word from a combination without changing the meaning of the sentence, so Mckeown et al. [78]. There are also idiomatic expressions (idioms) and collocations [78] which are used in natural language and should be considered. According to Mckeown et al., collocations might still be manageable within differential privacy and NLP, but more challenging would be the work with idioms which usually can not be perturbed [78]. In such cases, the possible replacements of a word are limited [78]. Another gap identified by Mckeown et al. is the dialectical forms which are used in the language [78]. There is still limited research on these topics. As one of the interviewee said: "If you have diverse data that has users with dialects, or people who speak non standard English, you're always going to end up getting worse utility on them. So that's another unfairness that DP has in NLP" Furthermore, another future research might investigate and provide more information how NLP and differential privacy are applicable in other languages except for English [126].

We also would like to point out that in our work we decided not to discuss what sensitive and private information is. As the majority of the interview participants mentioned the boundaries between private and public are not clearly defined. However, we believe that if practitioners and researchers have a grounded understanding and based on this are able to classify more precisely data points, that would make the usage of differential privacy at all more appropriate. Of course, we agree with the scientists who claim private information is not exactly 0 or 1, but is somewhere in between depending on the individual's perception. However, we still believe that such a research might be helpful for organisations, considering implementing differential privacy, with reasoning and justifying their choices of privacy protection. Furthermore, privacy preservation requires continuous observation because privacy protection does change over time, so Kan [57]. According to his opinion, we need to be up to date regarding how are legal and regulatory systems, management and business involved in protecting individual privacy [57].

When it comes to privacy, there is still a lot to be investigated in this sphere. First of all, scientists such as Mattern et al., claim that there is still no standard how to measure privacy

when working with textual data [76]. Second, as mentioned before, there is still no clarity what the value of the privacy budget should be in practice in order to protect from data leaking, even though it is considered that for machine learning an ϵ value smaller than 1 would provide strong privacy preservation, whereas an ϵ value bigger than 10 would not ensure much privacy [45]. Third, there is still less work regarding implementing differential privacy in NLP use cases such as model inference, so Hu et al. [52]. Therefore, further investigation is still required.

From the interviews we also came to the conclusion, that although there are educational materials and platforms that could facilitate the understanding of differential privacy to non technical audiences, there is gap in the educational materials, that provide easy to follow explanation and usage of differential privacy for technical audiences. For instance, people who are from the computer science field belong to the technical audience, however they might not have sufficient knowledge about privacy. Even in companies there is a gap between privacy experts and software developers. Lack of reproducibility and transparency, as discussed earlier, is still needed to help bridge this gap: "I believe we will move to the point where everything will be the gold standard: here's everything like in security - here's my code, here's my implementation and find the bug there. We should do that as a community", mentioned one of the interviewees. Another researcher gave its hope for this transparency: "Discussing how applying it [differential privacy] would impact the models or running some experiments to see what happens when it's applied because one of the places where it's most important is on the very large models and that's also a place where it's very challenging to work on it because training those models alone is very expensive. But training a model like that with differential privacy is even more expensive because adding noise to the data increases your costs."

As discussed in the previous chapter, there is still room for improvement regarding the similarity measurement between text corporas. There are several ways to look at similar texts: they might have similar structure/ syntax or they might have similar meaning with or without similar structure.

The work [16] also raises the problem that there is no enough research on how data scientists can decide what kind of query or statistics they want to perform on the database without having direct access to it. As discussed earlier, this is requirement that should be fulfilled independently from the domain where differential privacy is applied. Otherwise, data scientists might unintentionally reveal some data.

Many scientists stated during the interview process that there is a room for research in regard to finding other privacy-enhancing technologies that would provide better utility, and not "just brutally shuffle things at random noise" (as one interviewee shared with us), but also these technologies would be better applicable to textual data: "I think we might end up coming up with some new guarantees at least for text, like hopefully that would be easier to interpret. Maybe for another few years, you would just get better at utility privacy trade offs, minorities, stuff like that with DP SGD, but hopefully down the road in five years, we see some new guarantees that are better suited for language, and then we can go from there."

The successful implementation of differential privacy in giant companies like Apple, Meta

and Google demonstrated to us the usage of differential privacy is not as ubiquitous as all differentially private researchers would want it to be. Small and medium companies also have data publications and work with private data. Therefore, there is a need to find a way to bridge this gap or at least to answer the following questions: Is it worth to apply differential privacy within small and medium-sized companies and organizations? If not, is there a way to circumvent the challenges or are there technologies that when combined could ensure similar guarantees as differential privacy?

Further investigation of the speech domain is also needed. There were not many scientific papers that discussed how to apply differential privacy to speech data. What are the similarities and the differences between differential privacy within textual and within speech data? Such research might shed light into additional challenges to implement differential privacy in NLP - we should not neglect the fact that NLP works with both - text and speech.

9. Conclusion

In this scientific work, we were inspired to make the connection between pure ϵ-differential privacy and differential privacy that is used within NLP more straightforward. To fulfil this goal, we first started with a literature review, where we tried to identify the properties of ϵ-differential privacy that determine when does it make sense to use this privacy-enhancing technology and that are needed in order to apply differential privacy in practice. Our main findings describe the characteristics of differential privacy that make it meaningful to use differential privacy. Such features are the presence of an individual and private data that need to be protected, and the ability to analyse the data in cases when the data is preserved via the injecting of random values to it. In our work, we delve deeper by inspecting how an individual is defined, how his/her presence in the database is protected, and whether the privacy guarantees are met. Additionally, our main findings do also depict what do we need in order to use differential privacy - properly measured sensitivity, notion of a database and definition of adjacency. In our work, we consider how hese features are defined when the standard ϵ-differential privacy is used and what the differences and the similarities are when trying to implement differential privacy in NLP - how adjacent databases that differ in a single row can be mapped to adjacent databases in the context of NLP. Furthermore, we defined another two characteristics that we identified as appealing to practitioners who intend to use differential privacy. However, due to the fact, there are not mandatory features in order to work with differential privacy, we characterized them as lenient properties. These properties are the presence of a large database, as the utility is less damaged, and diverse database, that does not ensure less privacy guarantees to a group of individuals that share the same characteristics. We again examine how these properties can be mapped to NLP and what the differences and the similarities are.

Second, we searched for NLP use cases of differential privacy and for the practical view points of privacy experts in order to find out how the found properties of differential privacy are defined and implemented in practice. During our research, we found out that the way these properties are formulated in practice depends on the specific use case practitioners are confronted with. Therefore, we were even more encouraged to shed more light into how differential privacy could be applied to NLP use cases. We achieved this goal by extracting possible use cases from the papers we collected for our literature review, and by looking at different NLP tasks and discussing how the found characteristics can be mapped to these tasks and use cases. Additionally, through our discussions with researchers, we were able to find out the current challenges that may arise or may have arisen and that represent a barrier to adoption of differential privacy. Some of the challenges, such as ensuring good communication and transparency, privacy-utility trade-off and performance issues, are not new and there has been a lot of research in order to tackle these difficulties, but there is still

room to explore. We also discussed with our interview partners the successful stories of implementing differential privacy in NLP and which properties were met and how.

After presenting our main findings, we would like to share our hopes that our main findings will make the connection between differential privacy and differential privacy that is used in NLP more clear and will shed more light into how to apply differential privacy in NLP use cases, what is required, what should be considered by practitioners, and what the challenges might be. In addition, we hope we were able to bridge the research gap, we identified and formulated for our research and we hope our contribution to inspire more practitioners to apply differential privacy in practice, thus to broaden the applicability of differential privacy.

A. Interview Questions

A.1. Background

- What is your current position?
- How many years of experience with NLP do you have, if any?
- How familiar are you with Differential Privacy?

A.2. Differential Privacy Requirements [Artifact Review]

- Looking at standard Differential Privacy, do you agree with all of the requirements listed?
- Are there any requirements missing?
- Now looking at the NLP side: does the mapping from standard requirements make sense?
- Are there any changes that should be made?

A.3. Use Cases in NLP

- Can you name any NLP use cases in which Differential Privacy would be applicable? What makes Differential Privacy suitable?
- Could you elaborate upon any challenges or difficulties you foresee here?
- Could you discuss any potential limitations or risks of using Differential Privacy in NLP?
- Are there any technical challenges to be expected?
- From an implementation standpoint, what are the key considerations for organizations looking to deploy differentially private NLP systems? Are there any requirements or best practices?
- How can one evaluate privacy in NLP systems that employ Differential Privacy mechanisms?

- Can you share any real-world examples or case studies where differential privacy has been successfully implemented in NLP applications? What were the lessons learned from those experiences?

A.4. Looking Beyond

- How do you see the current landscape of privacy regulations and standards affecting the implementation of differential privacy in NLP?
- In your experience, do you find that tools and educational materials for explaining Differential Privacy are used in practice to facilitate the understanding to non-technical audiences?
- What are the key areas of NLP research that require further exploration to incorporate differential privacy?
- Generally speaking, what are the future prospects and challenges for implementing differential privacy in NLP? How do you envision the field evolving in the coming years?
- Is there any aspect that might have been missed that should be discussed?
- Can you recommend anyone else who may contribute to this interview study?

Bibliography

[1] K. Arabi. "Automated Refinement of an Ontology of NLP Research Concepts".

[2] Y. M. Baek, E.-m. Kim, and Y. Bae. "My privacy is okay, but theirs is endangered: Why comparative optimism matters in online privacy concerns". In: *Computers in Human Behavior* 31 (Feb. 2014), pp. 48–56. DOI: 10.1016/j.chb.2013.10.010.

[3] G. Beigi, K. Shu, R. Guo, S. Wang, and H. Liu. *I Am Not What I Write: Privacy Preserving Text Representation Learning*. en. arXiv:1907.03189 [cs]. July 2019. URL: http://arxiv.org/abs/1907.03189 (visited on 05/24/2023).

[4] S. Berghel, P. Bohannon, D. Desfontaines, C. Estes, S. Haney, L. Hartman, M. Hay, A. Machanavajjhala, T. Magerlein, G. Miklau, A. Pai, W. Sexton, and R. Shrestha. *Tumult Analytics: a robust, easy-to-use, scalable, and expressive framework for differential privacy*. en. arXiv:2212.04133 [cs]. Dec. 2022. URL: http://arxiv.org/abs/2212.04133 (visited on 05/24/2023).

[5] S. Bird. *Putting differential privacy into practice to use data responsibly*. Oct. 2020.

[6] V. Braun and V. Clarke. "Using thematic analysis in psychology". In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa. eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa. URL: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

[7] H. Brown, K. Lee, F. Mireshghallah, R. Shokri, and F. Tramèr. "What Does it Mean for a Language Model to Preserve Privacy?" In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 2280–2292.

[8] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. *The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks*. arXiv:1802.08232 [cs]. July 2019. DOI: 10.48550/arXiv.1802.08232. URL: http://arxiv.org/abs/1802.08232 (visited on 05/28/2023).

[9] R. S. Carvalho, T. Vasiloudis, and O. Feyisetan. *BRR: Preserving Privacy of Text Data Efficiently on Device*. en. arXiv:2107.07923 [cs]. July 2021. URL: http://arxiv.org/abs/2107.07923 (visited on 05/24/2023).

[10] R. S. Carvalho, T. Vasiloudis, and O. Feyisetan. *TEM: High Utility Metric Differential Privacy on Text*. arXiv:2107.07928 [cs]. July 2021. DOI: 10.48550/arXiv.2107.07928. URL: http://arxiv.org/abs/2107.07928 (visited on 05/27/2023).

[11] cbtranslateday. *Challenges in Translation*. 2021. URL: https://www.translateday.com/challenges-in-translation/ (visited on 05/11/2021).

[12] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi. "Broadening the Scope of Differential Privacy Using Metrics". en. In: *Privacy Enhancing Technologies*. Ed. by E. De Cristofaro and M. Wright. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 82–102. ISBN: 978-3-642-39077-7. DOI: 10.1007/978-3-642-39077-7_5.

[13] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. *Differentially Private Empirical Risk Minimization*. arXiv:0912.0071 [cs]. Feb. 2011. DOI: 10.48550/arXiv.0912.0071. URL: http://arxiv.org/abs/0912.0071 (visited on 05/28/2023).

[14] M. Corporation. *Differential Privacy for Everyone*. 2012. URL: https://download.microsoft.com/download/D/1/F/D1F0DFF5-8BA9-4BDF-8924-7816932F6825/Differential_Privacy_for_Everyone.pdf.

[15] R. Cummings and D. Desai. "The role of differential privacy in gdpr compliance". In: *FAT'18: Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2018, p. 20.

[16] R. Cummings, D. Desfontaines, D. Evans, R. Geambasu, M. Jagielski, Y. Huang, P. Kairouz, G. Kamath, S. Oh, O. Ohrimenko, N. Papernot, R. Rogers, M. Shen, S. Song, W. Su, A. Terzis, A. Thakurta, S. Vassilvitskii, Y.-X. Wang, L. Xiong, S. Yekhanin, D. Yu, H. Zhang, and W. Zhang. *Challenges towards the Next Frontier in Privacy*. 2023. arXiv: 2304.06929 [cs.CR].

[17] R. Danger. *Differential Privacy: What is all the noise about?* 2022. arXiv: 2205.09453 [cs.CR].

[18] F. K. Dankar and K. E. Emam. "The application of differential privacy to health data". In: *Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, March 30, 2012*. Ed. by D. Srivastava and I. Ari. ACM, 2012, pp. 158–166. DOI: 10.1145/2320765.2320816. URL: https://doi.org/10.1145/2320765.2320816.

[19] D. Desfontaines. *A list of real-world uses of differential privacy*. 2023. URL: https://desfontain.es/privacy/real-world-differential-privacy.html (visited on 03/21/2023).

[20] D. Desfontaines and B. Pejó. *SoK: Differential Privacies*. arXiv:1906.01337 [cs]. Nov. 2022. URL: http://arxiv.org/abs/1906.01337 (visited on 05/27/2023).

[21] *Differential Privacy*. Apple Inc. URL: https://www.apple.com/ru/privacy/docs/Differential_Privacy_Overview.pdf.

[22] C. Dilmegani. *Differential Privacy: How It Works, Benefits And Use Cases in 2023*. 2021. URL: https://research.aimultiple.com/differential-privacy/ (visited on 06/01/2021).

[23] B. Ding, J. Kulkarni, and S. Yekhanin. *Collecting Telemetry Data Privately*. 2017. arXiv: 1712.01524 [cs.CR].

[24] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer. "Detecting Violations of Differential Privacy". en. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. Toronto Canada: ACM, Oct. 2018, pp. 475–489. ISBN: 978-1-4503-5693-0. DOI: 10.1145/3243734.3243818. URL: https://dl.acm.org/doi/10.1145/3243734.3243818 (visited on 05/24/2023).

[25] M. Du, X. Yue, S. S. Chow, and H. Sun. "Sanitizing Sentence Embeddings (and Labels) for Local Differential Privacy". In: *Proceedings of the ACM Web Conference 2023*. 2023, pp. 2349–2359.

[26] C. Dwork and R. Pottenger. "Toward practicing privacy". In: *Journal of the American Medical Informatics Association* 20.1 (2013), pp. 102–108. DOI: 10.1136/amiajnl-2012-001047. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84871917950&doi=10.1136%5C%2famiajnl-2012-001047&partnerID=40&md5=184ac831aa20d09571911b94cd613de2.

[27] C. Dwork and A. Roth. "The algorithmic foundations of differential privacy". In: *Foundations and Trends in Theoretical Computer Science* 9.3-4 (2013), pp. 211–487. DOI: 10.1561/0400000042. URL: https://dblp.org/rec/journals/fttcs/DworkR14.html.

[28] C. Dwork. "Differential Privacy". en. In: *Automata, Languages and Programming*. Ed. by M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 1–12. ISBN: 978-3-540-35908-1. DOI: 10.1007/11787006_1.

[29] C. Dwork, N. Kohli, and D. Mulligan. "Differential privacy in practice: Expose your epsilons!" In: *Journal of Privacy and Confidentiality* 9.2 (2019).

[30] C. Dwork, F. McSherry, K. Nissim, and A. Smith. "Calibrating Noise to Sensitivity in Private Data Analysis". en. In: (2006).

[31] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum. *Differential Privacy Under Continual Observation*. URL: https://www.wisdom.weizmann.ac.il/~naor/PAPERS/continual_observation.pdf (visited on 05/31/2023).

[32] Ú. Erlingsson, V. Pihur, and A. Korolova. "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response". en. In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. Scottsdale Arizona USA: ACM, Nov. 2014, pp. 1054–1067. ISBN: 978-1-4503-2957-6. DOI: 10.1145/2660267.2660348. URL: https://dl.acm.org/doi/10.1145/2660267.2660348 (visited on 05/24/2023).

[33] N. Fernandes. "A novel framework for author obfuscation using generalised differential privacy". In: (Mar. 2022). DOI: 10.25949/19434467.v1. URL: https://figshare.mq.edu.au/articles/thesis/A_novel_framework_for_author_obfuscation_using_generalised_differential_privacy/19434467.

[34] N. Fernandes, M. Dras, and A. McIver. "Generalised differential privacy for text document processing". In: *Principles of Security and Trust: 8th International Conference, POST 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6–11, 2019, Proceedings 8*. Springer International Publishing, 2019, pp. 123–148.

[35] O. Feyisetan, B. Balle, T. Drake, and 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. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 178–186. ISBN: 978-1-4503-6822-3. DOI: 10.1145/3336191.3371856. URL: https://doi.org/10.1145/3336191.3371856 (visited on 05/27/2023).

[36] O. Feyisetan, T. Diethe, and T. Drake. "Leveraging Hierarchical Representations for Preserving Privacy and Utility in Text". In: *2019 IEEE International Conference on Data Mining (ICDM)*. ISSN: 2374-8486. Nov. 2019, pp. 210–219. DOI: 10.1109/ICDM.2019.00031.

[37] S. Fletcher, A. Roegiest, and A. K. Hudek. "Towards protecting sensitive text with differential privacy". In: *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2021, pp. 468–475.

[38] M. Gaboardi, M. Hay, and S. Vadhan. "A Programming Framework for OpenDP". en. In: ().

[39] U. Gallersdoerfer and F. Matthes. "Towards Valid Use Cases: Requirements and Supporting Characteristics of Proper Blockchain Applications". In: *2020 Seventh International Conference on Software Defined Systems (SDS)*. 2020, pp. 202–207. DOI: 10.1109/SDS49854.2020.9143999.

[40] U. Gallersdörfer and F. Matthes. "Towards Valid Use Cases: Requirements and Supporting Characteristics of Proper Blockchain Applications". In: *2020 Seventh International Conference on Software Defined Systems (SDS)*. Apr. 2020, pp. 202–207. DOI: 10.1109/SDS49854.2020.9143999.

[41] S. Gopi, P. Gulhane, J. Kulkarni, J. Shen, M. Shokouhi, and S. Yekhanin. "Differentially private set union". In: *37th International Conference on Machine Learning, ICML 2020*. Vol. PartF168147-5. 2020, pp. 3585–3594. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096198208&partnerID=40&md5=a4277d3e6d532581fad%5C-5670b6e3ccc91.

[42] I. Habernal. *EACL 2023 Tutorial on Privacy-preserving NLP: Formal guarantees with differential privacy*. Youtube. 2023. URL: https://www.youtube.com/watch?v=HCSqVwikv4U.

[43] I. Habernal. "How reparametrization trick broke differentially-private text representation learning". In: *arXiv preprint arXiv:2202.12138* (2022).

[44] I. Habernal. "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*. Ed. by M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih. Association for Computational Linguistics, 2021, pp. 1522–1528. DOI: 10.18653/v1/2021.emnlp-main.114. URL: https://doi.org/10.18653/v1/2021.emnlp-main.114.

[45] I. Habernal, F. Mireshghallah, P. Thaine, S. Ghanavati, and O. Feyisetan. "Privacy-Preserving Natural Language Processing". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. 2023, pp. 27–30.

[46] K. Hammoud, S. Benbernou, M. Ouziri, Y. Saygın, R. Haque, and Y. Taher. *Personal information privacy: what's next?* CEUR Workshop Proceedings. 2019. URL: https://ceur-ws.org/Vol-2622/paper5.pdf (visited on 07/06/2023).

[47] K. Hao. *How Apple personalizes Siri without hoovering up your data*. Dec. 2019.

[48] B. Harper. *Natural Language Processing. Bridging computers and human languages*. 2019. URL: https://www.oak-tree.tech/blog/data-science-nlp (visited on 10/11/2019).

[49] F. Hartmann and P. Kairouz. *Distributed differential privacy for federated learning*. Mar. 2023.

[50] R. Hathurusinghe, I. Nejadgholi, and M. Bolic. "A Privacy-Preserving Approach to Extraction of Personal Information through Automatic Annotation and Federated Learning". In: *Proceedings of the Third Workshop on Privacy in Natural Language Processing*. Online: Association for Computational Linguistics, June 2021, pp. 36–45. DOI: 10.18653/v1/2021.privatenlp-1.5. URL: https://aclanthology.org/2021.privatenlp-1.5 (visited on 05/28/2023).

[51] C. Horan. *Training, Visualizing, and Understanding Word Embeddings: Deep Dive Into Custom Datasets*. 2023. URL: https://neptune.ai/blog/word-embeddings-deep-dive-into-custom-datasets (visited on 04/21/2023).

[52] L. Hu, I. Habernal, L. Shen, and D. Wang. *Differentially Private Natural Language Models: Recent Advances and Future Directions*. 2023. arXiv: 2301.09112 [cs.CL].

[53] T. Igamberdiev and I. Habernal. "Privacy-Preserving Graph Convolutional Networks for Text Classification". In: *2022 Language Resources and Evaluation Conference, LREC 2022*. 2022, pp. 338–350. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85144381170&partnerID=40&md5=ee695cbc525844a94014eb960927b9de.

[54] T. Igamberdiev and I. Habernal. *DP-BART for Privatized Text Rewriting under Local Differential Privacy*. 2023. arXiv: 2302.07636 [cs.CR].

[55] N. Johnson, J. P. Near, and D. Song. "Towards practical differential privacy for SQL queries". In: *Proceedings of the VLDB Endowment* 11.5 (2018). Publisher: VLDB Endowment, pp. 526–539.

[56] Z. Jorgensen, T. Yu, and G. Cormode. "Conservative or liberal? Personalized differential privacy". In: *2015 IEEE 31st International Conference on Data Engineering*. ISSN: 2375-026X. Apr. 2015, pp. 1023–1034. DOI: 10.1109/ICDE.2015.7113353.

[57] K. Kan. *Seeking the Ideal Privacy Protection: Strengths and Limitations of Differential Privacy*. IMES Discussion Paper Series 23-E-02. Institute for Monetary and Economic Studies, Bank of Japan, Jan. 2023. URL: https://ideas.repec.org/p/ime/imedps/23-e-02.html.

[58] C. Kapelke. *Using differential privacy to harness big data and preserve privacy*. 2020. URL: https://www.brookings.edu/articles/using-differential-privacy-to-harness-big-data-and-preserve-privacy/ (visited on 08/11/2020).

[59] G. Kerrigan, D. Slack, and J. Tuyls. *Differentially Private Language Models Benefit from Public Pre-training*. arXiv:2009.05886 [cs]. Oct. 2020. URL: http://arxiv.org/abs/2009.05886 (visited on 07/06/2023).

[60] K. Kim, S. Gopi, J. Kulkarni, and S. Yekhanin. "Differentially Private n-gram Extraction". In: *Advances in Neural Information Processing Systems*. Vol. 7. 2021, pp. 5102–5111. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85131728803&partnerID=40&md5=482c6479e4d58d2dd1778c9fdb29d2d5.

[61] B. A. Kitchenham, D. Budgen, and P. Brereton. *Evidence-Based Software Engineering and Systematic Reviews*. Chapman & Hall/CRC, Oct. 2015. ISBN: 978-1-4822-2865-6.

[62] O. Klymenko, S. Meisenbacher, and F. Matthes. "Differential Privacy in Natural Language Processing The Story So Far". en. In: *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 1–11. DOI: 10.18653/v1/2022.privatenlp-1.1. URL: https://aclanthology.org/2022.privatenlp-1.1 (visited on 05/24/2023).

[63] O. Klymenko, S. Meisenbacher, and F. Matthes. "Differential Privacy in Natural Language Processing: The Story So Far". In: *arXiv preprint arXiv:2208.08140* (2022).

[64] M. Koppel, S. Argamon, and A. R. Shimoni. "Automatically Categorizing Written Texts by Author Gender". In: *Literary and Linguistic Computing* 17.4 (Nov. 2002), pp. 401–412. ISSN: 0268-1145. DOI: 10.1093/llc/17.4.401. eprint: https://academic.oup.com/dsh/article-pdf/17/4/401/3345463/170401.pdf. URL: https://doi.org/10.1093/llc/17.4.401.

[65] S. Krishna, R. Gupta, and C. Dupuy. "ADePT: Auto-encoder based differentially private text transformation". In: *arXiv preprint arXiv:2102.01502* (2021).

[66] A. Kumar. *Natural Language Processing (NLP) Task Examples*. 2023. URL: https://vitalflux.com/natural-language-processing-nlp-task-examples/ (visited on 09/03/2023).

[67] N. Li, M. Lyu, D. Su, and W. Yang. "Differential privacy: From theory to practice". In: *Synthesis Lectures on Information Security, Privacy, & Trust* 8.4 (2016). Publisher: Morgan & Claypool Publishers, pp. 1–138.

[68] X. Li, C. Luo, P. Liu, and L.-E. Wang. "Information entropy differential privacy: A differential privacy protection data method based on rough set theory". In: *Proceedings - IEEE 17th International Conference on Dependable, Autonomic and Secure Computing, IEEE 17th International Conference on Pervasive Intelligence and Computing, IEEE 5th International Conference on Cloud and Big Data Computing, 4th Cyber Science and Technology Congress, DASC-PiCom-CBDCom-CyberSciTech 2019*. 2019, pp. 918–923. DOI: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00169. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85075119281&doi=10.1109%5C%2fDASC%5C%2fPiCom%5C%2fCBDCom%5C%2fCyberSciTech.2019.00169&partnerID=40&md5=a8e4d98bae10c1ff6319740c239910f2.

[69] Y. Li, T. Baldwin, and T. Cohn. "Towards robust and privacy-preserving text representations". In: *arXiv preprint arXiv:1805.06093* (2018).

[70] K. Ligett. *Tutorial on Differential Privacy*. 2013. URL: https://simons.berkeley.edu/talks/tutorial-differential-privacy (visited on 12/11/2013).

[71] L. Lyu, X. He, and Y. Li. "Differentially Private Representation for NLP: Formal Guarantee and An Empirical Study on Privacy and Fairness". In: *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*. Ed. by T. Cohn, Y. He, and Y. Liu. Vol. EMNLP 2020. Findings of ACL. Association for Computational Linguistics, 2020, pp. 2355–2365. DOI: 10.18653/v1/2020.findings-emnlp.213. URL: https://doi.org/10.18653/v1/2020.findings-emnlp.213.

[72] A. Machanavajjhala, X. He, and M. Hay. "Differential Privacy in the Wild: A Tutorial on Current Practices & Open Challenges". en. In: *Proceedings of the 2017 ACM International Conference on Management of Data*. Chicago Illinois USA: ACM, May 2017, pp. 1727–1730. ISBN: 978-1-4503-4197-4. DOI: 10.1145/3035918.3054779. URL: https://dl.acm.org/doi/10.1145/3035918.3054779 (visited on 05/24/2023).

[73] G. Maheshwari, P. Denis, M. Keller, and A. Bellet. "Fair NLP Models with Differentially Private Text Encoders". In: *arXiv preprint arXiv:2205.06135* (2022).

[74] A. Mansouri, L. S. Affendey, and A. Mamat. "Named entity recognition approaches". In: *International Journal of Computer Science and Network Security* 8.2 (2008), pp. 339–344.

[75] O. Mantiri. "Factors Affecting Language Change". In: *http://ssrn.com/abstract=2566128* (Mar. 2010). DOI: 10.2139/ssrn.2566128.

[76] J. Mattern, Z. Jin, B. Weggenmann, B. Schölkopf, and M. Sachan. "Differentially Private Language Models for Secure Data Sharing". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*. 2022, pp. 4860–4873. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146633916&partnerID=40&md5=b28ad0d30b672916e3a00a3c0a1cc79b.

[77] C. Matzken, S. Eger, and I. Habernal. *Trade-Offs Between Fairness and Privacy in Language Modeling*. 2023. arXiv: 2305.14936 [cs.CL].

[78] K. R. McKeown and D. R. Radev. "Collocations". In: *Handbook of Natural Language Processing. Marcel Dekker* (2000), pp. 1–23.

[79] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. *Learning Differentially Private Recurrent Language Models*. en. arXiv:1710.06963 [cs]. Feb. 2018. URL: http://arxiv.org/abs/1710.06963 (visited on 05/24/2023).

[80] F. McSherry and K. Talwar. *Mechanism Design via Differential Privacy*. URL: http://kunaltalwar.org/papers/expmech.pdf (visited on 05/28/2023).

[81] N. Mouhammad, J. Daxenberger, B. Schiller, and I. Habernal. *Crowdsourcing on Sensitive Data with Privacy-Preserving Text Rewriting*. 2023. arXiv: 2303.03053 [cs.CL].

[82] D. Mwiti. *How to Create Word Embeddings With TensorFlow*. 2023. URL: https://www.machinelearningnuggets.com/how-to-create-word-embeddings-with-tensorflow/ (visited on 01/10/2023).

[83] A. Narayanan and V. Shmatikov. "Robust De-anonymization of Large Sparse Datasets". en. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. ISSN: 1081-6011. Oakland, CA, USA: IEEE, May 2008, pp. 111–125. ISBN: 978-0-7695-3168-7. DOI: 10.1109/SP.2008.33. URL: http://ieeexplore.ieee.org/document/4531148/ (visited on 05/24/2023).

[84] K. Nissim, S. Raskhodnikova, and A. Smith. "Smooth Sensitivity and Sampling in Private Data Analysis". In: *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*. STOC '07. San Diego, California, USA: Association for Computing Machinery, 2007, pp. 75–84. ISBN: 9781595936318. DOI: 10.1145/1250790.1250803. URL: https://doi.org/10.1145/1250790.1250803.

[85] P. Ohm. "Sensitive information". In: *S. Cal. L. Rev.* 88 (2014), p. 1125.

[86] A. Ouadrhiri and A. Abdelhadi. "Differential privacy for fair deep learning models". In: *15th Annual IEEE International Systems Conference, SysCon 2021 - Proceedings*. Vol. 2021-January. 2021. DOI: 10.1109/SysCon48628.2021.9591252. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119285088&doi=10.1109%5C%2fSysCon48628.2021.9591252&partnerID=40&md5=d7a4298f036c3b13c66fb8%5C-fbe0eba632.

[87] X. Pan, M. Zhang, S. Ji, and M. Yang. "Privacy risks of general-purpose language models". In: *Proceedings - IEEE Symposium on Security and Privacy*. Vol. 2020-May. 2020, pp. 1314–1331. DOI: 10.1109/SP40000.2020.00095. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090943319&doi=10.1109%5C%2fSP40000.2020.00095&partnerID=40&md5=7d0ce7ce852cf9aa5a82932015d51035.

[88] K. Panchal. "Differential privacy and natural language processing to generate contextually similar decoy messages in honey encryption scheme". In: *arXiv preprint arXiv:2010.15985* (2020).

[89] N. Ponomareva, H. Hazimeh, A. Kurakin, Z. Xu, C. Denison, H. B. McMahan, S. Vassilvitskii, S. Chien, and A. Thakurta. *How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy*. arXiv:2303.00654 [cs, stat]. Mar. 2023. DOI: 10.48550/arXiv.2303.00654. URL: http://arxiv.org/abs/2303.00654 (visited on 05/30/2023).

[90] C. Qu, W. Kong, L. Yang, M. Zhang, M. Bendersky, and M. Najork. "Natural Language Understanding with Privacy-Preserving BERT". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. arXiv:2104.07504 [cs]. Oct. 2021, pp. 1488–1497. DOI: 10.1145/3459637.3482281. URL: http://arxiv.org/abs/2104.07504 (visited on 05/30/2023).

[91] B. Reynolds, J. Venkatanathan, J. Gonçalves, and V. Kostakos. "Sharing Ephemeral Information in Online Social Networks: Privacy Perceptions and Behaviours". en. In: *Human-Computer Interaction – INTERACT 2011*. Ed. by P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, and M. Winckler. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, pp. 204–215. ISBN: 978-3-642-23765-2. DOI: 10.1007/978-3-642-23765-2_14.

[92] R. Rogers, S. Subramaniam, S. Peng, D. Durfee, S. Lee, S. K. Kancha, S. Sahay, and P. Ahammad. *LinkedIn's Audience Engagements API: A Privacy Preserving Data Analytics System at Scale*. 2020. arXiv: 2002.05839 [cs.CR].

[93] S. Roy, J. Hsu, and A. Albarghouthi. "Learning Differentially Private Mechanisms". In: *2021 IEEE Symposium on Security and Privacy (SP)*. ISSN: 2375-1207. May 2021, pp. 852–865. DOI: 10.1109/SP40001.2021.00060.

[94] T. Sasada, M. Kawai, Y. Taenaka, D. Fall, and Y. Kadobayashi. "Differentially-Private Text Generation via Text Preprocessing to Reduce Utility Loss". In: *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. Apr. 2021, pp. 042–047. DOI: 10.1109/ICAIIC51459.2021.9415242.

[95] B. Seleshi and S. Asseffa. *A Case Study on Differential Privacy*. URL: http://www.diva-portal.org/smash/get/diva2:1113852/FULLTEXT01.pdf (visited on 07/05/2023).

[96] M. Senekane. "Deployment of Differential Privacy for Application in Artificial Intelligence". In: *International Conference on Electrical, Computer, and Energy Technologies, ICECET 2021*. 2021. DOI: 10.1109/ICECET52533.2021.9698473. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85127086680&doi=10.1109%5C%2fICECET52533.2021.9698473&partnerID=40&md5=d61e962499e%5C-0b8014cb5718f3c68f3c5.

[97] W. Shi, A. Cui, E. Li, R. Jia, and Z. Yu. "Selective differential privacy for language modeling". In: *arXiv preprint arXiv:2108.12944* (2021).

[98] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. *Membership Inference Attacks against Machine Learning Models*. arXiv:1610.05820 [cs, stat]. Mar. 2017. URL: http://arxiv.org/abs/1610.05820 (visited on 05/29/2023).

[99] A. Singh. *Text summarization using NLP*. Apr. 2020.

[100] C. Song and A. Raghunathan. *Information Leakage in Embedding Models*. arXiv: 2004.00053 [cs, stat]. Aug. 2020. URL: http://arxiv.org/abs/2004.00053 (visited on 07/06/2023).

[101] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and D. Megías. "Individual differential privacy: A utility-preserving formulation of differential privacy guarantees". In: *IEEE Transactions on Information Forensics and Security* 12.6 (2017). Publisher: IEEE, pp. 1418–1429.

[102] S. Sousa and R. Kern. "How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing". In: *Artificial Intelligence Review* 56.2 (2023). Publisher: Springer, pp. 1427–1492.

[103] B.-C. Tai, S.-C. Li, and Y. Huang. "K-aggregation: Improving accuracy for differential privacy synthetic dataset by utilizing k-anonymity algorithm". In: *Proceedings - International Conference on Advanced Information Networking and Applications, AINA*. 2017, pp. 772–779. DOI: 10.1109/AINA.2017.97. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85019705748&doi=10.1109%5C%2fAINA.2017.97&partnerID=40&md5=88b9e5af1d37d669ee17da5986a055b6.

[104] N. Talagala. *Data as The New Oil Is Not Enough: Four Principles For Avoiding Data Fires*. 2022. URL: https://www.forbes.com/sites/nishatalagala/2022/03/02/data-as-the-new-oil-is-not-enough-four-principles-for-avoiding-data-fires/ (visited on 03/02/2022).

[105] C. Task and C. Clifton. "A Guide to Differential Privacy Theory in Social Network Analysis". en. In: *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Istanbul: IEEE, Aug. 2012, pp. 411–417. ISBN: 978-1-4673-2497-7. DOI: 10.1109/ASONAM.2012.73. URL: http://ieeexplore.ieee.org/document/6425731/ (visited on 05/24/2023).

[106] C. Task. *A Practical Beginners' Guide to Differential Privacy - CERIAS Security Seminar, Purdue University*. Youtube. 2012. URL: https://www.youtube.com/watch?v=Gx13lgEudtU&t=411s.

[107] T. O. Team. *The OpenDP White Paper*. 2020. URL: https://projects.iq.harvard.edu/files/opendp/files/opendp_white_paper_11may2020.pdf (visited on 05/11/2020).

[108] P. Thaine. *Differentially Private Natural Language Processing*. 2019. URL: https://medium.com/privacy-preserving-natural-language-processing/differentially-private-natural-language-processing-4f18912c5de0 (visited on 01/28/2019).

[109] A. Triastcyn and B. Faltings. "Bayesian differential privacy for machine learning". In: *International Conference on Machine Learning*. PMLR, 2020, pp. 9583–9592.

[110] M. C. Tschantz, D. Kaynar, and A. Datta. *Formal Verification of Differential Privacy for Interactive Systems*. arXiv:1101.2819 [cs]. Jan. 2011. DOI: 10.48550/arXiv.1101.2819. URL: http://arxiv.org/abs/1101.2819 (visited on 06/06/2023).

[111] D. Vatsalan, R. Bhaskar, A. Gkoulalas-Divanis, and D. Karapiperis. "Privacy Preserving Text Data Encoding and Topic Modelling". In: *Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021*. 2021, pp. 1308–1316. DOI: 10.1109/BigData52589.2021.9671552. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-

85125312891&doi=10.1109%5C%2fBigData52589.2021.9671552&partnerID=40&md5=17875bce4d317abb5e3275fd1cf4c4af.

[112] F. Vogel and L. Lange. *Privacy-Preserving Sentiment Analysis on Twitter*. 2023. URL: https://dbs.uni-leipzig.de/file/SKILL2023_private_twitter_sentiment-6.pdf (visited on 08/28/2023).

[113] W. Wang, P. Tang, J. Lou, and L. Xiong. "Certified Robustness to Word Substitution Attack with Differential Privacy". In: *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*. 2021, pp. 1102–1112. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85114253246&partnerID=40&md5=d087d82a1568914773cc824129f7988b.

[114] Z. Wang, X. Zhou, R. Koncel-Kedziorski, A. Marin, and F. Xia. *Extracting and Inferring Personal Attributes from Dialogue*. 2022. arXiv: 2109.12702 [cs.CL].

[115] B. Weggenmann, V. Rublack, M. Andrejczuk, J. Mattern, and F. Kerschbaum. "DP-VAE: Human-Readable Text Anonymization for Online Reviews with Differentially Private Variational Autoencoders". In: *Proceedings of the ACM Web Conference 2022*. WWW '22. Virtual Event, Lyon, France: Association for Computing Machinery, 2022, pp. 721–731. ISBN: 9781450390965. DOI: 10.1145/3485447.3512232. URL: https://doi.org/10.1145/3485447.3512232.

[116] C. Weiss, F. Kreuter, and I. Habernal. *To share or not to share: What risks would laypeople accept to give sensitive data to differentially-private NLP systems?* 2023. arXiv: 2307.06708 [cs.CL].

[117] A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, J. Honaker, K. Nissim, D. O'Brien, T. Steinke, and S. Vadhan. "Differential Privacy: A Primer for a Non-Technical Audience". en. In: *SSRN Electronic Journal* (2018). ISSN: 1556-5068. DOI: 10.2139/ssrn.3338027. URL: https://www.ssrn.com/abstract=3338027 (visited on 05/24/2023).

[118] A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, J. Honaker, K. Nissim, D. O'Brien, T. Steinke, and S. Vadhan. *Differential Privacy: A Primer for a Non-Technical Audience*. en. SSRN Scholarly Paper. Rochester, NY, 2018. DOI: 10.2139/ssrn.3338027. URL: https://papers.ssrn.com/abstract=3338027 (visited on 05/27/2023).

[119] C. Wright and K. Rumsey. *The Strengths, Weaknesses and Promise of Differential Privacy as a Privacy-Protection Framework*. URL: https://math.unm.edu/~knrumsey/pdfs/projects/DifferentialPrivacy.pdf (visited on 06/03/2023).

[120] P. Wu. *The Privacy Risk of Language Models*. 2022. URL: https://www.private-ai.com/2022/06/17/the-privacy-risk-of-language-models/ (visited on 06/17/2022).

[121] C. Xu, J. Wang, F. Guzmán, B. Rubinstein, and T. Cohn. "Mitigating Data Poisoning in Text Classification with Differential Privacy". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021, pp. 4348–4356.

[122] Z. Xu, A. Aggarwal, O. Feyisetan, and N. Teissier. *A Differentially Private Text Perturbation Method Using a Regularized Mahalanobis Metric*. arXiv:2010.11947 [cs, stat]. Oct. 2020. DOI: 10.48550/arXiv.2010.11947. URL: http://arxiv.org/abs/2010.11947 (visited on 05/28/2023).

[123] Z. Xu, Y. Zhang, G. Andrew, C. A. Choquette-Choo, P. Kairouz, H. B. McMahan, J. Rosenstock, and Y. Zhang. *Federated Learning of Gboard Language Models with Differential Privacy*. arXiv:2305.18465 [cs]. July 2023. URL: http://arxiv.org/abs/2305.18465 (visited on 08/25/2023).

[124] M. Yang, L. Lyu, J. Zhao, T. Zhu, and K.-Y. Lam. *Local Differential Privacy and Its Applications: A Comprehensive Survey*. en. arXiv:2008.03686 [cs]. Aug. 2020. URL: http://arxiv.org/abs/2008.03686 (visited on 05/24/2023).

[125] X. Yue, M. Du, T. Wang, Y. Li, H. Sun, and S. S. Chow. "Differential privacy for text analytics via natural text sanitization". In: *arXiv preprint arXiv:2106.01221* (2021).

[126] J. Yuki. "Clinical Natural Language Processing for Japanese: Challenges and Opportunities". In: (Feb. 2023). DOI: 10.36227/techrxiv.22032656.v1. URL: https://www.techrxiv.org/articles/preprint/Clinical_Natural_Language_Processing_for_Japanese_Challenges_and_Opportunities/22032656.

[127] W. Zhang, J. Zhao, F. Wei, and Y. Chen. "Differentially Private High-Dimensional Data Publication via Markov Network". In: *EAI Endorsed Transactions on Security and Safety* 6.19 (Jan. 2019). DOI: 10.4108/eai.29-7-2019.159626.

[128] F. Zhao, X. Ren, S. Yang, Q. Han, P. Zhao, and X. Yang. "Latent Dirichlet Allocation Model Training With Differential Privacy". In: *IEEE Transactions on Information Forensics and Security* 16 (2021). Conference Name: IEEE Transactions on Information Forensics and Security, pp. 1290–1305. ISSN: 1556-6021. DOI: 10.1109/TIFS.2020.3032021.

[129] Y. Zhao and J. Chen. "A survey on differential privacy for unstructured data content". In: *ACM Computing Surveys (CSUR)* 54.10s (2022). Publisher: ACM New York, NY, pp. 1–28.

[130] T. Zhu, G. Li, W. Zhou, and P. S. Yu. *Differential Privacy and Applications*. Vol. 69. Advances in Information Security. Springer, 2017. ISBN: 978-3-319-62002-2. DOI: 10.1007/978-3-319-62004-6. URL: https://doi.org/10.1007/978-3-319-62004-6.

[131] T. Zhu, G. Li, W. Zhou, and P. S. Yu. "Differentially Private Data Publishing and Analysis: A Survey". In: *IEEE Transactions on Knowledge and Data Engineering* 29.8 (Aug. 2017). Conference Name: IEEE Transactions on Knowledge and Data Engineering, pp. 1619–1638. ISSN: 1558-2191. DOI: 10.1109/TKDE.2017.2697856.

[132] T. Zhu, D. Ye, W. Wang, W. Zhou, and S. Y. Philip. "More than privacy: Applying differential privacy in key areas of artificial intelligence". In: *IEEE Transactions on Knowledge and Data Engineering* 34.6 (2020). Publisher: IEEE, pp. 2824–2843.

[133] A. Zimovnov. *Simple Deep Neural Networks for Text Classification*. Youtube. 2018. URL: https://www.youtube.com/watch?v=wNBaNhvL4pg.