Towards Task-Agnostic Privacy- And Utility-Preserving Models

Yaroslav Emelyanov
Sberbank, Moscow, Russia
emelyanov.y.i@sberbank.ru

Abstract

Modern deep learning models for natural language processing rely heavily on large amounts of annotated texts. However, obtaining such texts may be difficult when they contain personal or confidential information, for example, in health or legal domains. In this work, we propose a method of de-identifying free-form text documents by carefully redacting sensitive data in them. We show that our method preserves data utility for text classification, sequence labeling, and question answering tasks.

1 Introduction

Data privacy has become an important topic recently, and new regulations that govern the processing and usage of consumer personal data arise every year. We use the term *sensitive data* to define all information that contains personal data or other potentially compromising data that can lead to the re-identification of individuals or organizations. With the advent of machine learning, privacy and compliant data governance have become even more important.

Free-form text documents often contain sensitive data. For example, legal documents contain full information about individuals or organizations, dialogues contain references to different people and possibly information that can lead to their identification, such as addresses or job positions. NLP applications such as text classification or question answering require annotated data, and as Feyisetan et al. (2019) argues, the privacy cost of annotating texts must be considered when developing new applications.

Anonymization, or *de-identification*, is viewed as one of the ways to make working with non-public datasets both safer and more compliant. It is also important to preserve utility in de-identified data for further usages, e.g. machine learning or analytics. Traditional approaches like k-anonymity (Sweeney, 2002) were shown to be ineffective when applied to high-dimensional media (Aggarwal, 2005) for example, free-form texts. Moreover, unlike structured data present in relational databases, anonymization of unstructured text documents poses additional challenges, because locations of textual spans which constitute sensitive data must be inferred at runtime.

Previous work either focused on preventing the model from memorizing data (Kerrigan et al., 2020) or considered only the text classification task. In this paper, we aim to solve the challenges of preserving both privacy and utility in free-form text documents with respect to different NLP tasks. We propose a simple “find-and-replace” method for automatically de-identifying documents and evaluate the utility of the de-identified data with respect to different downstream tasks, namely named entity recognition, question answering and text classification. To the best of our knowledge, this is the first work that evaluates the utility of the de-identified data for sequence labeling and reading comprehension tasks.

Specifically, our contributions are as follows:

1. We develop and evaluate the deep-learning-based method for de-identifying text documents.
2. We conduct series of experiments and show that our method mostly preserves utility for different NLP tasks: NER, QA, and text classification.
3. We investigate how our method impacts the end task performance for different model architectures.

2 Background

2.1 Preserving Privacy In Texts

There are several research directions in privacy in NLP. One research direction covers generating synthetic training data, thus abandoning original

Proceedings of Recent Advances in Natural Language Processing, pages 394–401
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_045

Task	Dataset	# instances	# classes	domain
NER	RuReBus (Ivanin et al., 2020)	218 documents	8	State documents
NER	FactRuEval (Starostin et al., 2016)	255 documents	3	News
Question Answering	SberQuAD (Efimov et al., 2020)	45328 questions	–	Wikipedia
Text Classification	In-house data	4000 sentences	13	Legal documents

Table 1: Datasets used in experiments

data completely. Krishna et al. (2021) shows that this approach performs well for text classification and protects against membership inference attacks. However, in narrow domains such as legal contracts, maintaining internal coherency is important for information retrieval tasks, but generating long coherent texts is still a challenging NLP task (Tan et al., 2021).

Another area of research focuses on adding noise during training to prevent models from memorizing their training data. For example, Kerrigan et al. (2020) add noise to gradients during training to prevent large generative models from "memorizing" training data. However, NLP applications like NER or QA are powered by large amounts of annotated training data and data annotation happens *before* model training. This means that more people, including annotators, should have access to private data, which adds additional scrutiny to the dataset development process (Feyisetan et al., 2019).

The third branch of privacy research can be viewed as a form of noise added to the original data. The system that adds this noise must satisfy the requirement, stated in Dwork (2006):

Anything that can be learned about a respondent from the statistical database should be learnable without access to the database.

While this assumption is not achievable in practice, because we have to extract *some* value from the database, several methods were proposed to satisfy more relaxed privacy guarantees, e.g. (ε, δ) differential privacy (Geng and Viswanath, 2013).

Finally, a practical approach for texts de-identification via targeting named entities was proposed. In this approach, the auxiliary NER model is trained to recognize sensitive spans in the document that are further redacted. Stubbs et al. (2015); Marimon et al. (2019) organized challenges to develop the best de-identification system for English or Spanish medical texts, however, they did not explore the utility of anonymized data.

Our method falls into the latter research direction, however, we treat text de-identification as an *auxiliary* part of our work, focusing on measuring *utility* in the de-identified data.

2.2 Measuring Utility In The De-Identified Data

Rahman et al. (2018) show that models trained via privacy-preserving methods may poorly generalize to the original data. This means that data utility is reduced during the de-identification procedure.

Several approaches to measure utility in de-identified data were proposed. For example, Sánchez et al. (2014) follow the information-theoretic approach to ensure that de-identified entities do not exhibit any information that would help the attacker to de-identify data, even when the attacker has access to large amounts of open information from the Internet. Another approach is to measure utility as the quality of models trained on the downstream tasks on the de-identified data (Xu et al., 2020a). While the latter approach gives an intuitive and practical definition of utility, it is not clear how utility estimates depend on the end task, data, and model architecture. We adopt the latter approach and investigate impact of task, data and model choice in the section 4.

3 Proposed Method

We aim to develop a method that performs fine-grained substitutions of text spans comprising sensitive information. Contexts of sensitive data are kept intact and therefore our method preserves as much original text as possible.

3.1 Extracting Sensitive Information

Searching for sensitive information, such as names or IDs of individuals, can be difficult in free-form texts. We formulate it as a named entity recognition (NER) task (Mamede et al., 2016), for which various methods were proposed. For example, IDs and other numerical information may be found by regular expressions, while medical diagnoses may be looked up in dictionaries. However, as Yadav and Bethard (2018) suggest, deep learning methods mostly outperform gazetteer-based or feature-based

Original text	**Pseudonymized text**
Context: В Миссолонги Байрон заболел лихорадкой, продолжая отдавать все свои силы на борьбу за свободу страны. In Missolonghi, Byron fell ill with a fever, continuing to devote all his strength to the struggle for the freedom of the country.	**Context:** В Сельниково Колико заболел лихорадкой,продолжая отдавать все свои силы на борьбу за свободу страны. In Selnikovo, Koliko fell ill with a fever, continuing to devote all his strength to the struggle for the freedom of the country.
Question: Чем заболел Байрон в Миссолонги? What made Byron sick at Missolonghi?	**Question:** Чем заболел Колико в Сельниково? What made Koliko sick at Selnikovo?
Answer: лихорадкой fever	**Answer:** лихорадкой fever

Table 2: Consistent pseudonymization for paragraph and question for SberQuad sample and its English translation. All mentions of Байрон (Byron) (highlighted with dashed underline) are pseudonymized consistently. In addition, all mentions of Миссолонги (Missolongi) are also anonymized consistently.

models, and we opt for the neural-network-based approach.

Traditionally (Tjong Kim Sang and De Meulder, 2003), the performance of NER models is evaluated by micro-averaged f1-measure over extracted spans. Several datasets (Stubbs et al. (2015), Garat and Wonsever (2019)) exist to evaluate de-identification models, however, they do not provide annotations that would enable measuring utility in the de-identified data.

3.2 Replacing Sensitive Information

Several ways of replacing sensitive information were proposed in the literature (Carrell et al., 2020; Jiang et al., 2019). We study two methods:

1. `Sanitization` strategy redacts the sensitive information and replaces it with the token describing the label of the replaced entity, e.g. replace “John Smith” with generic label `PERSON`, “Organization inc.” with generic label `ORGANIZATION`.

2. `Pseudonymization` replaces real entities with synthetically generated but semantically and grammatically sound values. Pseudonymization is often used in practice when releasing private data for research or third parties (Stubbs et al., 2015). Table 2 provides an example of pseudonymization strategy for SberQuAD dataset.

To implement `Sanitization` strategy, we need only the label of the entity the token corresponds to, which is available from NER model prediction at inference time. However, such replacement erases coreference links throughout the document, which may be important for the downstream task. In section 4 we provide results and show when this strategy impairs the performance on the downstream tasks.

`Pseudonymization` strategy is more difficult to implement because tokens of different entity types should be replaced differently. For example, a sequence of random digits comprises a number, but sequence of random characters does not always comprise a valid person or organization name. We generate synthetic values for `Pseudonymization` strategy as follows:

1. For numerical spans, e.g. numbers, IDs, and dates, we generate random numbers of the same length.

2. For textual spans, e.g. names, addresses, we use lookup from dictionaries. We make a random selection from the dictionary independently for every word in the span.

For tasks that require reasoning over input text, like question answering (Rajpurkar et al., 2016), train instances should maintain internal coherency: inconsistent changes in context and question would leave the question unanswerable. To solve this problem, `Pseudonymization` strategy maintains a mapping between original and pseudonymized values during its work, which allows coherent replacements of entities values. During anonymization of the datasets, each dataset instance is processed independently, meaning that mentions of the same person in different instances will be anonymized differently. We do not add

Dataset	Architecture	Training data	$M'(T'_{tgt})$	$M'(V'_{tgt})$	$M'(V_{tgt})$	Δ	SOTA
FactRuEval	`BERT-FC`	original	0.99	0.85	0.850±0.003	-	0.86
		pseudonimized	0.99	0.81	0.757±0.032	**-0.093±0.035**	Starostin et al. (2016)
		sanitized	0.98	0.82	0.401±0.047	-0.449±0.050	
RuReBus	`BERT-FC`	original	0.737	0.540	0.540±0.005	-	0.56
		pseudonimized	0.727	0.527	0.530±0.003	-0.010±0.008	Ivanin et al. (2020)
		sanitized	0.925	0.526	0.528±0.002	-0.012±0.007	
SberQuAD	`BERT-QA`	original	0.891	0.825	0.825±0.002	-	0.848
		pseudonimized	0.851	0.815	0.821±0.002	**-0.004±0.004**	Efimov et al. (2020)
		sanitized	0.872	0.784	0.793±0.002	-0.032±0.004	
In-house legal documents	Gradient Boosting	original	0.988	0.942	0.942±0.001	-	-
		pseudonimized	0.981	0.942	0.937±0.003	**-0.005±0.004**	
		sanitized	0.931	0.925	0.905±0.001	-0.037±0.002	

.

Table 3: Micro f1 measures and relative differences with the models trained on original data. Higher Δ is better. For models trained on original data, $T' \equiv T$, $V' \equiv V$

any coreference information in `Sanitization` strategy.

3.3 Re-Identification Risks

Our de-identification system relies on the NER model. Deleger et al. (2013) show that even double manual de-identification is not perfect, so de-identification errors will inevitably occur. Such errors may lead to *re-identification* of the de-identified subject, for example, when not all of their mentions were de-identified. Scaiano et al. (2016) propose to use all-or-nothing recall to evaluate de-identification models, because if even one mention of the person within the document was not de-identified, adversary may be able to re-identify the person. For example, if document has 10 mentions of the person, of which 9 were anonymized, then all-or-nothing recall is 0, while regular recall is 0.9. We measure all-or-nothing recall in our experiments. Meystre et al. (2014); Scaiano et al. (2016) argue that re-identification risks are further reduced for `Pseudonymization` strategy compared to `Sanitization`: when encountering specific names in the pseudonymized document, it is difficult to tell whether they were left intact by an imperfect anonymizer system or they are the result of pseudonymization.

4 Experiments

4.1 Experimental Protocol

Let S be our anonymization strategy (for example, sanitization or pseudonymization) and $D_{target} = (T_{target}, V_{target})$ be the target dataset for which we train model on a downstream task. Target dataset is split into train and validation sets named as T and V. Let M be the model trained on T_{target}. Our goal is to evaluate how M's performance on V_{target} depends on the dataset M is trained on: original T_{target} or de-identified T'_{target}. Inspired by this goal, we design the following experimental protocol:

1. Anonymize T_{target}, V_{target} and obtain $T'_{target} = S(T_{target}), V'_{target} = S(V_{target})$
2. Train model on original data: $M = M(T_{target})$, get validation metrics $M(V_{target})$
3. Train model on de-identified data: $M' = M(T'_{target})$, get validation metrics for both original and de-identified data: $M'(V'_{target})$, $M'(V_{target})$
4. Compare results of models: $\Delta = M'(V_{target}) - M(V_{target})$

We repeat this experiment for every strategy S and compare results for each strategy with the baseline that was trained on the original data. Our anonymization procedure may yield distribution shift which may result in imperfect generalization to the original validation set and therefore in negative values of Δ. $\Delta \geq 0$ means that data utility was completely preserved. We repeat each experiment 3 times and report mean and variance for $M'(V_{target})$ across them in Table 3.1.

4.2 NER Models For Extracting Sensitive Data

For experiments on publicly available data, we use Collection3 corpus created by Mozharova and Loukachevitch (2016) to develop our anonymizer system. This corpus has the same annotation schema as FactRuEval and is approximately 7 times larger. We train vanilla BERT-based NER model on Collection3 corpus and obtain the micro f1 measure 0.931. To measure re-identification risks, we manually review 200 randomly chosen documents from anonymized SberQuAD train set and find that all-or-nothing recall is 0.93.

For NER model for in-house data de-identification, we use an in-house corpus of 3040 documents with 426 272 annotated entities. This corpus has annotation schema similar to Ontonotes dataset (Weischedel et al., 2011). We train BERT-CRF model on this corpus using 90-10 train-test split, evaluate the model using micro f1-measure over spans and get the value of 0.93. We performed an additional evaluation to see how well our model finds sensitive data. We asked domain experts to manually annotate sensitive information in 30 legal documents of various types and then checked our system's output against these annotations. We found that of 1030 entities, 1009 were anonymized, resulting in 0.98 recall and 0.95 all-or-nothing recall.

4.3 Downstream Tasks And Models

To show that our method transfers across tasks and domains, we use different NLP tasks and datasets:

FactRuEval (Starostin et al., 2016) is a NER dataset developed to evaluate fact extraction from Russian news articles. It is annotated similarly to Tjong Kim Sang and De Meulder (2003) and has PER, LOC, ORG entities.

RuReBus (Ivanin et al., 2020) is another NER dataset consisting of state documents and reports. It has more diverse annotation schema then FactRuEval. It is annotated with custom annotation schema that includes entities like METRICS, ACTIVITY, QUALITATIVE. Unlike FactRuEval, most of the classes in this dataset are not considered as sensitive data, except for INSTITUTION class, which is similar to ORG class in FactRuEval.

SberQuAD (Efimov et al., 2020) is Russian extractive question answering dataset similar to SQuAD (Rajpurkar et al., 2016). It has 9.080 unique paragraphs and 50.364 questions, about 20% of answers and about 72 % of paragraphs contain named entities that should be anonymized, e.g. people, locations, or organizations. Unlike Rajpurkar et al. (2016), SberQuAD does not have unanswerable questions.

For the **text classification** task, an internal dataset of 5 000 texts annotated with 13 different classes was used. Data instances are segments of legal documents and classes represent types of these segments.

We use train-dev splits provided by the authors for all publicly available datasets. For text classification and NER tasks, we use micro averaged f1 measure. For SberQuAD, we use f1 measure as in the SQuAD dataset.

Comparative statistics of all datasets are shown in Table 1.

For NER datasets, we use simple BERT for token classification as described in Devlin et al. (2019). For SberQuAD, we use the same architecture as Devlin et al. (2019) used for SQuAD. For the in-house text classification dataset, BERT performed on par with gradient boosting (Ke et al., 2017) on top of tf-idf vectorization, and we choose boosting for its simplicity. It is also interesting to investigate how our anonymization methods affect end task performance for different vectorization methods.

4.4 Utility Tests

As described in subsection 4.1, we measure the utility of anonymized data as the performance drop between models trained on original and anonymized datasets. All experiments were implemented with AllenNLP (Gardner et al., 2018) framework. We present our results in Table 3.1.

In all experiments, baseline models trained on original data achieved performance close to currently reported state-of-the-art results. We note that in all experiments anonymization impairs end task performance, although results vary depending on the task and dataset.

Experiments on **RuReBus** dataset showed only a slight performance drop, however, all models achieve relatively low scores compared to the FactRuEval dataset. We attribute low scores to the inconsistent annotations in the RuReBus dataset. We attribute the low difference between performance on pseudonymized and sanitized data to the annotation schema: most entities in the schema are not considered sensitive information. However, for entity INSTITUTION, which is close to ORG entity, performance drop is significant: from 0.436 f1-measure in original data to 0.348 in sanitized data.

Similarly, in **text classification** task performance changes are also small compared to SberQUAD and FactRuEval tasks. This can be explained by the nature of the task: Xu et al. (2020b); Marivate and Sefara (2020) show that text classi-

Architecture	Training data	$M'(T'_{tgt})$	$M'(V'_{tgt})$	$M'(V_{tgt})$	Δ
`BERT-FC`	original	0.99	0.85	0.85	-
`BERT-FC`	sanitized	0.98	0.82	0.40	-0.45
`BERT-FC`	pseudonimized	0.99	0.81	0.76	**-0.09**
`BERT-BiLSTM`	original	0.96	0.75	0.77	-
`BERT-BiLSTM`	sanitized	0.93	0.77	0.15	-0.62
`BERT-BiLSTM`	pseudonimized	0.95	0.70	0.61	**-0.16**
`w2v-CNN-BiLSTM`	original	0.89	0.70	0.69	-
`w2v-CNN-BiLSTM`	sanitized	0.83	0.72	0.12	-0.57
`w2v-CNN-BiLSTM`	pseudonimized	0.77	0.55	0.46	**-0.23**

Table 4: Results for different architectures for FactRuEval dataset. Higher Δ is better.

fication task is robust to different kinds of noise, including word substitution, which is close to our anonymization procedure.

Results on **SberQuAD** dataset confirm our hypothesis that consistent anonymization is important for question answering: difference between pseudonymized and sanitized data is higher than in previous experiments. About 28% of the dataset was kept intact by anonymization and therefore performance drop for anonymized instances will be even larger.

The largest difference in performance between models trained on sanitized and pseudonymized data is in **FactRuEval** dataset. This difference can be attributed to the annotation schema and nature of the task: Bernier-Colborne and Langlais (2020) showed that that NER models rely more on entity text and less on the entity context. This intuition explains performance drop for models trained on sanitized data: they have not seen any real entities during training and can find entities based only on the context during the evaluation on the original data. However, models trained on pseudonymized data are able to generalize from synthetic entities to real ones.

Our experiments suggest that task, data and annotation schema impact downstream task model sensitivity to data anonymization.

4.5 Impact Of The Downstream Model Architecture

In lieu of the current NLP state, we perform most of our experiments using BERT-based models. However, we also explore how anonymization impacts end task performance for different model architectures. We choose FactRuEval dataset for this experiment because in prior experiments we showed that it is more sensitive to data anonymization. We use three popular NER architectures:

`BERT-FC` is a vanilla BERT for token classification model (Devlin et al., 2019), both pre-trained layers and projection layer are fine-tuned during training. We use `RuBERT` initialization trained by Kuratov and Arkhipov (2019) in all experiments because it performed the best on the original data in all experiments.

`BERT-BiLSTM` is an architecture with the 2-layer bidirectional LSTM applied on top of BERT embeddings. During training, BERT parameters are frozen and only LSTM layers are tuned. We use the same `RuBERT` initialization.

`w2v-CNN-BiLSTM` is a popular architecture that uses fixed word embeddings together with character embeddings to encode each token and BiLSTM on top of them to encode context. We use word embeddings trained by Grave et al. (2018) and keep them frozen during training due to the small size of the training corpus.

We provide results in Table 4. As in subsection 4.4, pseudonymization enjoys lower performance drops for all architectures. We notice that performance drops for BERT-based models are lower. This can be explained by the number of OOV words that generates our pseudonymization procedure: synthetic names or addresses are randomly sampled from the large dictionaries, so they are mostly not present in the embeddings table even for large pre-trained word embeddings. We calculated that only 38% of all names and 15% of all surnames from our dictionaries are present in the pre-trained embeddings. Our results support the claim made by Hendrycks et al. (2020), who showed that pre-trained transformers are more robust to distribution shifts.

5 Suggestions To Practitioners

Our experiments highlight several characteristics of anonymization procedure, downstream task and model architecture that should be taken into account when using anonymized data for training NLP models. Our suggestions are as follows:

1. Noise-robust downstream tasks are also robust to anonymization.

2. Downstream tasks that do not require reasoning over named entities are robust to anonymization.

3. Downstream tasks that require reasoning over named entities also require coherent pseudonymization to maintain data consistency.

4. Pseudonymization works better than sanitization, although it is more difficult to develop.

5. Transformer-based models generalize the best between original and anonymized data.

6 Conclusions And Future Work

In this work, we consider the practical side of anonymizing unstructured documents while simultaneously preserving their utility for different downstream tasks. New policies regarding personal data make privacy research a more important topic over the years. We anticipate that in the near future de-identification of sensitive data before training will become a necessity. We hope our work will pave the way for investigating broader impact and limitations of free-form text anonymization.

We demonstrate that pseudonymization mostly preserves data utility for different extractive NLP tasks. We show it is possible to achieve close results with the model trained only on the de-identified data. However, it is not yet clear whether our results transfer to generative tasks and more complex settings, for example, scenarios with multiple languages like machine translation or multilingual datasets. We believe this is the promising research direction.

References

Charu C Aggarwal. 2005. On k-anonymity and the curse of dimensionality. In *VLDB*, volume 5, pages 901–909.

Gabriel Bernier-Colborne and Phillippe Langlais. 2020. Hardeval: Focusing on challenging tokens to assess robustness of ner. In *LREC*.

D. Carrell, B. Malin, David Cronkite, J. Aberdeen, C. Clark, Muqun Rachel Li, Dikshya Bastakoty, Steve Nyemba, and L. Hirschman. 2020. Resilience of clinical text de-identified with "hiding in plain sight" to hostile reidentification attacks by human readers. *Journal of the American Medical Informatics Association : JAMIA*.

Louise Deleger, Katalin Molnar, Guergana Savova, Fei Xia, Todd Lingren, Qi Li, Keith Marsolo, Anil Jegga, Megan Kaiser, Laura Stoutenborough, and Imre Solti. 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association*, 20:84–94.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Cynthia Dwork. 2006. Differential privacy. volume 4052 LNCS, pages 1–12. Springer Verlag.

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. Sberquad–russian reading comprehension dataset: Description and analysis. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 3–15. Springer.

Oluwaseyi Feyisetan, Thomas Drake, Borja Balle, and Tom Diethe. 2019. Privacy-preserving active learning on sensitive data for user intent classification.

Diego Garat and Dina Wonsever. 2019. Towards de-identification of legal texts.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv*.

Quan Geng and P. Viswanath. 2013. The optimal mechanism in \$$(\epsilon, \delta)$\$$- differentialprivacy$.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness.

VA Ivanin, EL Artemova, TV Batura, VV Ivanov, VV Sarkisyan, EV Tutubalina, and IM Smurov. 2020. Rurebus-2020 shared task: Russian relation extraction for business. In *Computational Linguistics and Intellectual Technologies*, pages 416–431.

Dehuan Jiang, Yedan Shen, Shuai Chen, Buzhou Tang, Xiaolong Wang, Q. Chen, Ruifeng Xu, J. Yan, and Yi Zhou. 2019. A deep learning-based system for the meddocan task. In *IberLEF@SEPLN*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. volume 30, pages 3146–3154.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. 2020. Differentially private language models benefit from public pre-training. pages 39–45. Association for Computational Linguistics.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. Adept: Auto-encoder based differentially private text transformation. *arXiv preprint arXiv:2102.01502*.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

N. Mamede, J. Baptista, and Francisco Dias. 2016. Automated anonymization of text documents. *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1287–1294.

M. Marimon, A. Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@SEPLN*.

Vukosi Marivate and Tshephisho Sefara. 2020. Improving short text classification through global augmentation methods. pages 385–399.

S. Meystre, Shuying Shen, Deborah Hofmann, and A. Gundlapalli. 2014. Can physicians recognize their own patients in de-identified notes? *Studies in health technology and informatics*, 205:778–82.

Valerie Mozharova and Natalia Loukachevitch. 2016. Two-stage approach in russian named entity recognition. Institute of Electrical and Electronics Engineers Inc.

Md.Atiqur Rahman, Tanzila Rahman, R. Laganière, and N. Mohammed. 2018. Membership inference attack against differentially private deep learning model. *Trans. Data Priv.*, 11:61–79.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

D. Sánchez, Montserrat Batet, and A. Viejo. 2014. Utility-preserving privacy protection of textual healthcare documents. *Journal of biomedical informatics*, 52:189–98.

Martin Scaiano, G. Middleton, Luk Arbuckle, V. Kolhatkar, L. Peyton, M. Dowling, D. Gipson, and K. Emam. 2016. A unified framework for evaluating the risk of re-identification of text de-identification tools. *Journal of biomedical informatics*, 63:174–183.

A. Starostin, V. Bocharov, S. Alexeeva, A. Bodrova, A. Chuchunkov, S. S. Dzhumaev, Irina Efimenko, D. Granovsky, V. Khoroshevsky, I. Krylova, M. Nikolaeva, I. Smurov, and S. Toldova. 2016. Factrueval 2016: Evaluation of named entity recognition and fact extraction systems for russian. In *Annual International Conference Dialogue*.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.

Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4313–4324, Online. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. *OntoNotes: A Large Training Corpus for Enhanced Processing*.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020a. A differentially private text perturbation method using regularized mahalanobis metric. In *Proceedings of the Second Workshop on Privacy in NLP*, pages 7–17, Online. Association for Computational Linguistics.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020b. A differentially private text perturbation method using regularized mahalanobis metric. pages 7–17. Association for Computational Linguistics.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.