CMPUT 622 Project Report: Privacy-Preserving language model for sentiment analysis on COVID-19 tweets

Seeratpal Jaura
CCID: seeratpa
Department of Computing Science
University of Alberta, Edmonton, Alberta
seeratpa@ualberta.ca

Sohyun Park
CCID: sohyun2
Department of Computing Science
University of Alberta, Edmonton, Alberta
sohyun2@ualberta.ca

Abstract

Sentiment analysis is an important natural language processing (NLP) task; however, privacy preservation is a big concern. Our work focused on privacy preservation for sentiment analysis during the COVID-19 situation and treated emotion and sentiments as a piece of sensitive information that needs to be protected. To ensure the privacy of sensitive data, we implemented the Differentially Private (DP) Bidirectional Encoder Representations from Transformers (BERT) model. Furthermore, we presented and implemented the keyword mask DP BERT model to protect the sentimental expression. Our results demonstrated that the keyword masked DP BERT model achieved higher accuracy than the DP BERT model. Moreover, the Keyword mask model solved the memorization problem of baseline and DP BERT model.

1 Introduction

Sentiments refer to "a thought, opinion, or idea based on a feeling about a situation, or a way of thinking about something" by the Cambridge dictionary [1]. Sentiment analysis is an important task in the natural language processing (NLP) task that aims to determine sentiments in the text as positive, negative and neutral. However, privacy preservation is a key challenge in opinion mining tasks such as sentiment analysis [1]. During the global pandemic (COVID-19), people from all over the world heavily relied on social media for expressing their strong emotions and concerns. Specifically, the situation was very uncertain from March 2020 to April 2020. People have expressed their sensitive sentiments in public related to their mental health, physical health and financial conditions. Analyzing such text with natural language processing techniques would help the government make policies accordingly.

Previous studies have shown that machine learning models hold valuable intellectual property [2]. Neural Networks tend to *unintentionally memorize* by learning the out-of-distribution data that seems irrelevant to the task [3]. An adversary could perform model extraction in NLP to gain private information by performing queries. This leads to a novel challenge to analyze sentiments using neural networks without compromising the privacy of sensitive individual data.

The problem we are interested in exploring is based on a scenario. Consider a private firm/company, who wants to analyze how the employee's life was affected by the COVID-19. Particularly the company might be interested in investigating if their employees had positive or negative sentiment in the COVID-19 situation. In this scenario, we will be training the model on private data, and the model must not learn the data for inference. In such a case, adding noise while model training would

[1] https://dictionary.cambridge.org/dictionary/english/sentiment

help preserve sensitive information related to the user since the data is private. For our course project, we don't have access to private data. So we are considering COVID-19 tweet data to be private data where we want to protect the user attributes and sentiments during model training.

Users willingly share their information on social media, therefore public tweets do not always pose a privacy risk. However, tweets on sensitive subjects like mental, physical health and financial condition of people during an unusual situation of COVID-19 need to handle with care. As suggested by [4] that the care must be taken because the account users can make their account private at any point in time, which would make their previously collected tweets as private data. Our work deals with protecting the privacy of emotions and sensitive information (including user attributes).

2 Motivation

Emotions reveal personal feelings about an issue and have the strong potential to expose private information (like age, gender, and mental health). Especially in the COVID-19 situation, accusation towards a particular race on social media becomes inevitably severe [5], and this could stigmatize someone as a racist too easily. Most people would not want to be held under this disgraceful prejudice due to the expression of emotions at the moment. Taking an example from our COVID-19 tweets data for illustration, "Tweet: alive again. No more No more China shanghai chinavirus corona covid_nineteen crownvirus two thousand and twenty covidnineteen covdnineteen asia mask washyourhands : Shanghai, China" [6]. In this example, certain words like "chinavirus" target a particular race, and it could easily intimidate a specific group of people. Therefore, we'd like to deal with emotions as private information and prevent privacy leakage during the model training on the dataset.

Furthermore, our task is motivated to preserve the privacy of sensitive data that could potentially reveal user identity. Using an example from our dataset, *"A Palestinian man, Sabri Dweikat, seventy, has been making traditional straw brooms for forty years, in the West Bank city of Nablus. The production and sale of straw brooms were reduced due to the coronavirus disease"* [6] reflects a lot of sensitive information that could reveal the identity of a person. This leads to severe privacy information leakage. Moreover, the paper "I Am Not What I Write: Privacy-Preserving Text Representation Learning" clearly states that sometimes tweets on sensitive subjects reveal more sensitive information than the user intends to provide [7]. Taking another example from our dataset, *"Day thirty-two of madness.. and I'm sick not corona fools just allergies and diabetes shit... Phoenix, Arizona"* [6]. An adversary could use this information to find some secrets about the user. One such secret could be determining if the user is paying higher or lower health insurance.

Therefore, we focus on protecting the sentiments and sensitive information within the text.

3 Related Work

The work done on sentiment analysis has shown that the user attributes can be easily detected and extracted during these analyses [8]. More recently, attention has focused on neural networks because of their capabilities; however, they pose a high capacity for memorization [9]. Memorization in neural networks could lead to severe privacy leakage. The hidden representations of neural network models impose a risk on privacy by providing the ability to an attacker to predict sensitive information [10].

Furthermore, Carlini et al. [3] addresses the issue of "unintended memorization" by neural networks models during the training process. The paper reflects the key privacy concern of generative sequence models, which often memorize the out-of-distribution data that seems irrelevant to the task. This employs the risk of learning sensitive information. An adversary could query a threat model many times to uncover private information such as bank details, health records and unique personal identity information. In their experiments, they inserted the secret information they referred to as "canaries" in their paper to showcase that neural networks quickly memorize out-of-distribution data during training even though the model does not overtrain. This accidental exposure of private information raises a concern to safeguard information from unintentional memorization.

Previous works have shown from their experiments that differential privacy is an effective defence against individual privacy attacks [3][9]. They have applied privacy techniques on Long Short-term memory (LSTM) based architecture. Differential Privacy (DP) [11] guarantees privacy by using the notion that the individual information does not get harmed during any analysis (more details are under the methodology section). Mireshghallah et al. [9] proposed a training model using adversarial learning to improve the privacy of neural representation in the language classification task. Bidirectional Encoder Representations from Transformers (BERT) is now widely used for machine learning tasks. Basu et al. [4] also studied the effects of Differential Privacy and Federated Learning on contextualized language models BERT, ALBERT, RoBERTa and DistilBERT for Depression and Sexual Harassment related data.

4 Proposed Solution

To overcome the above challenges, our project has implemented three models: the baseline BERT model, the Differentially Private Bidirectional Encoder Representations from Transformers (BERT) model, and Keyword masked Differentially Private BERT model. We utilized the pre-trained BERT model to perform sentiment analysis as it is less computationally expensive and requires fewer resources to optimize its performance.

Dataset: For this task, since we don't have access to private data, we consider the COVID-19 tweet data to be private as it has sensitive information. We performed privacy-preserving sentiment analysis on the "Coronavirus (COVID-19) geotagged tweets dataset" from 20 March 2020, to 31 March 2020 available on the *IEEE* website [6]. Due to Twitter's content distribution policy, the dataset only has tweet ids, so we reconstructed the tweets from tweet ids using the Hydrator application [2]. Hydrating tweets is the process of obtaining tweets texts from tweets' unique ids. We make our dataset with 19,126 tweets. This number came out when we found a balance point between avoiding model overfitting and CUDA memory that we can use through Google Colab [12]. We labelled positive sentiments as 1 and negative sentiment as 0 using TextBlob [13], which is mentioned in the IEEE COVID-19 dataset webpage [6]. Although we cannot say making a label using TextBlob is perfect, it gave us the label results that seem correct to us. There are 15,711 positive tweets and 3,415 negative tweets in our dataset. Note that the dataset has about 82% of positive labels and about 18% of negative sentiments, showing us that the dataset was imbalanced.

4.1 Models

This section applies to the models for our task of privacy preservation on COVID-19 tweets data.

4.1.1 Baseline model

The Baseline model is implemented using the state-of-the-art pre-trained Bidirectional Encoder Representations from Transformers (BERT) model and fine-tuned on our task of sentiment analysis on COVID-19 tweets. Bidirectional Encoder Representations from Transformers (BERT) is a language representation model [14] developed by Google. BERT was pre-trained on two tasks. First, on a language modeling task where 15% of tokens were masked, it was trained to predict based on context. Secondly, BERT was trained to predict if a chosen next sentence was probable or not given the first sentence on the sentence prediction task. BERT learns contextual embeddings for words during the training process. After pre-training, the model is fine-tuned with fewer resources on smaller datasets to optimize its performance on specific tasks.

4.1.2 Differentially private BERT model

We implemented a differentially private BERT model(DP BERT model) to protect the sensitive information within the dataset while training on the neural network model. Differential Privacy is a mechanism that deals with plausible deniability by stating that the individual data is protected. It resists attacks on personal sensitive information.

Definition according to Dwork and Roth [11]: *A randomized algorithm M with domain* $\mathbb{N}^{|X|}$ *is* (ϵ, δ)*-differentially private if for all S* $\subseteq$ *Range(M) and for all x, y* $\in \mathbb{N}^{|X|}$ *such that* $\|x - y\|_1 \leq$ *1 :*

$$Pr[M(x) \in S] \leq exp(\epsilon)Pr[M(y) \in S] + \delta$$

[2] https://github.com/DocNow/hydrator

where ϵ refers to the privacy budget and δ refers to error probability. In other words, we can say that the dataset x and y that differs by one row would yield similar results. The adversary cannot infer any private individual information while querying the dataset. One of the flexible features of M is that it does not impose any implications on its nature.

To make the DP BERT model, we used the Opacus library, which is developed by Facebook [15]. Opacus library helps us train the PyTorch models with differential privacy. It ensures differential privacy by adding the noise to the model's parameter gradient to update weights. Thus, it adds noise to data; it adds Privacy to the model, making the resulting model safe to release.

In terms of architecture, the DP BERT model is similar to the vanilla BERT model except that the DP BERT model attaches a privacy engine to its optimizer. The Opacus library applies privacy to every step of the optimizer. Thus, the model parameters are updated with epsilon values. The library helps add the right amount of noise because adding more noise than required noise could greatly imbalance the results, while small noise puts more privacy risk. Therefore, it computes the per-sample gradients of each sample in a mini-batch to overcome this issue. Then it clips their gradients individually and aggregates them into the single batch gradient. In the final step, it adds the Gaussian noise to the parameters of every step[15].

4.1.3 Keyword mask DP BERT model

The DP BERT model is very general as a privacy-protecting method. However, even with the DP BERT model, it is still possible to guess certain words in the dataset because it applies the same epsilon values for each step. If the attackers use the prediction score of the model, then it is possible to guess if the dataset contains a specific word or not. In this project, we mainly aim to protect the users from the attacker who wants to know the user's sentimental expression regarding the specific situation.

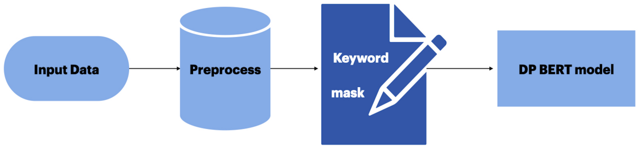

Figure 1: Pipeline for Keyword mask DP BERT model

Therefore, we developed a keyword mask for sentimental data. Using word2vec cosine similarity, we made a mask containing sentimental words(negative/positive) in the dataset and concealed them from the dataset. As a result, machines do not train with certain sentimental and sensitive words in specific situations. This method makes it impossible for the attackers to guess which words are belong to the dataset with a prediction score. This technique can also solve the memorization problem of the deep neural network because the keyword mask model trains words representing particular sentiment, not the original ones.

Meanwhile, the cosine similarity has a high probability of giving us the antonyms as similar words due to the characteristics of the vectors.

Given two vectors of attributes, A and B,

$$cosine\ similarity = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||\ ||\mathbf{B}||}$$

For this reason, we consider the antonyms of each positive and negative word when we calculated the keyword mask. Then the formula should change for including the other attribute, C, using Triangle inequality.

Originally, the triangle inequality for angles is

$$|\angle AC - \angle CB| \leq \angle AB \leq \angle AC + \angle CB$$

.

While an angle in $[0, \pi]$ radians increasing, the cosine function decreases. Thus, the inequalities is reversed when we apply the cosine to each value:

$$cos(\angle AC - \angle CB) \geq cos(\angle AB) \geq cos(\angle AC + \angle CB)$$

With the cosine addition and subtraction, we can write the two inequalities:

$$cos(A, C) \cdot cos(C, B) + \sqrt{(1 - cos(A, C)^2 \cdot (1 - cos(C, B)^2)} \geq cos(A, B)$$

$$cos(A, B) \geq cos(A, C) \cdot cos(C, B) - \sqrt{(1 - cos(A, C)^2) \cdot (1 - cos(C, B)^2)}$$

This form of the triangle inequality can calculate the maximum (or minimum) similarity of two objects A and B when we know the similarities to a reference vector C. It lets us create a more robust keyword mask that mainly contains synonyms of a particular word using its antonyms.

5 Experiments

5.1 Training Data and Batching

We trained on the COVID-19 geotagged tweets dataset consisting of about 20,000 tweets paired with sentimental labels. Tweets are tokenized and encoded using BertTokenizer with the pre-trained model "bert-base-case" provided from the transformer library. We split Training data and Testing data with a ratio of 8 to 2; therefore, there are 15300 tweets in the training dataset and 3826 tweets in the testing dataset. The tweets and labels are batched together, and the batch size is 4, while the virtual batch size for using opacus(==0.15.0) privacy engine is 32.

5.2 Hardware and Schedule

We trained our models on Google Colab GPU session, and the models consume approximately 3GB of GPU memory for each. The epochs are 4, and the Logging interval is 100, which are all the same for every model we used. The training time of the suggested models was not significantly different from the baseline model. For each epoch, the models took 3824 training steps, and it took about 14 minutes. Therefore, for training the model, we need approximately 45 minutes overall.

5.3 Optimizer and Privacy engine

We used AdamW[16] optimiser with $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e - 8$, which is

$$x_t \leftarrow x_{t-1} - \alpha \frac{\beta_1 m_{t-1} + (1 - \beta_1)(\nabla f_t + w x_{t-1})}{\sqrt{v_t} + \epsilon}$$

where f is cost function, $x(t)$ is weight, β_1 and β_2 are parameters that control how quickly the averages decay, and α is learning rate.

We applied the privacy engine provided by Opacus library to optimizers of our DP BERT model and keyword mask DP BERT model. We used $\frac{1}{training\ dataset\ size}$ as δ value, which is the parameter for privacy accounting. And the maximum gradient norm for the privacy engine is 0.1. For the sample rate of the engine, we used virtual batch size, batch size, and the training dataset size. The sample rate is like below:

$$sample\ rate = \frac{batch\ size(== 4)}{training\ dataset\ size} * \frac{virtual\ batch\ size(== 32)}{batch\ size(== 4)}$$

5.4 Epsilon noise

We employ three numbers of epsilon noise during training:

epsilon noise 0.5 It is the lowest value of epsilon. With this value of epsilon, we expect more privacy, however, that privacy will come at the cost of accuracy. Thus, a low epsilon value (more privacy) would yield lower accuracy.

epsilon noise 5.5 Another value we consider for the privacy budget is epsilon noise of 5.5. We applied this epsilon value to find the optimal privacy budget for our project in order to find the balance point between the privacy budget spent and the accuracy achieved. We hope to achieve higher accuracy than the epsilon value of 0.5 since we relaxed the privacy budget. Furthermore, we hope that the model loss will also decrease.

epsilon noise 7.5 The third value that we consider for epsilon is 7.5 for our experiments. We also wanted to measure how does the accuracy and loss value will vary by increasing the epsilon for our task. We expected the best accuracy of the model with least loss among three epsilon cases when using this value of epsilon.

5.5 Expressions for prediction score

We chose five sentences for employing our experiments regarding prediction score based on the keyword mask which we got during the implementation:

Sentence 1	"Corona time gains We have no control over the situations the world is up against"
Sentence 2	"This is how Chinese people caught Corona Virus They eat anything they worse than white people chinavirus"
Sentence 3	"an empty Escalator at M Purple Line station. Where are the passenger? Afraid of CoVidnineteen ?"
Sentence 4	"I am working away from home to earn a living and the covidnineteen causes a total lockdown in my state"
Sentence 5	"The virus is too dangerous and may last for many years. I m going back to Neptune. corona China"

Table 1: Sentence from dataset [6] that have labels classified as negative

These sentences are picked up based on the labels classified as negative (from dataset [6]). Considering that the situation we supposed for the project causes severe problems in people's lives, we decided it is essential to deal with negative feelings more than positive ones. In addition, within this dataset (even if it is not perfectly classified), negative emotions have more scarcity, so we thought the consequences of our application of techniques to negative emotions were important than vice versa. Therefore, we focused more on negative sentences than positive sentences for the experiments.

6 Results

We considered two models - differential privacy BERT model and keyword mask differential privacy BERT model - based on BERT, a state-of-the-art model for implementing privacy in natural language processing(NLP) tasks. We compare their respective performance and with baseline model.

6.1 Model performance with Epsilon noise

From the experiments, our baseline model's loss and accuracy values are 0.424 and 83%. And when we used the DP BERT model, the loss went up, and the accuracy went down. It happened because we applied epsilon noise for the DP BERT model, which is related to the model capacity. Therefore, the lower the noise is, the smaller the model capacity to memorization, making the accuracy go lower compared to the baseline model.

Meanwhile, we can see that the accuracy of the keyword mask DP BERT model for epsilon value 0.5 is 55.9%, significantly higher than the DP BERT model with the same epsilon value(0.5). Given that the loss goes more elevated than the DP BERT model's thus there is a high probability that the model is overfitted to the dataset. However, the problem is inevitable: First, the dataset is relatively small. In addition, we changed specific words to representative words in that small dataset. Second, the labels are not perfectly matched with the sentences, and there are lots of sentences that have more negative facts than positive ones but are classified as positive. Thus, we can say the model cannot train appropriately. If we can have a more extensive dataset and solve the dataset labeling problem, the problem of the loss value will go down.

Even considering a problem with the dataset, the keyword mask DP BERT model provides higher accuracy than the DP BERT model at all epsilon values. Furthermore, it is consistent with our intention not to harm the model's accuracy while some privacy noises are applied.

Baseline			
loss	0.424		
Acc	83%		
Differential Privacy BERT Model			
Epsilon	0.5	5.5	7.5
loss	1.422	1.357	1.327
Acc	17.8%	75.8%	82.5%
Keyword Mask DP BERT Model			
Epsilon	0.5	5.5	7.5
loss	1.540	1.369	1.301
Acc	**55.9%**	**82.7%**	**83.2%**

Table 2: Loss and Accuracy of the 3 models

6.2 Prediction score

With the baseline model, we found that the sentences we mentioned at **5.5** are classified as negative with an evidently higher value than prediction scores for the positive possibility of the sentence. The prediction scores of negative for the sentences are almost the highest values that the model can yield, so we can say with the baseline model it is straightforward to know if the sentences are in our dataset, so the model trained with them or not.

In the case of the DP BERT model, the score gap between the positive and negative labels narrowed. However, since the words in the sentences do not have negative meanings in general situations, we can infer relatively easily that the model has learned the sentence. We can certainly prove that our thoughts are reasonable through the results of the keyword mask DP BERT model on the sentences.

We gave the sentences that don't apply the keyword mask as inputs to the keyword mask DP BERT model. Unlike the baseline and DP BERT models, the keyword mask DP BERT model classifies the sentences as positive. Furthermore, the negative and positive labels gap is more significant than the DP BERT model. The result addresses that we can expect the sentiments to be protected using the keyword mask DP BERT model.

sentence	model	**negative**	positive
sentence 1	Baseline	1.500	0.749
	DP-BERT	0.534	0.377
	Keyword Mask	0.259	0.645
sentence 2	Baseline	1.500	0.749
	DP-BERT	0.569	0.232
	Keyword Mask	0.327	0.733
sentence 3	Baseline	1.500	0.749
	DP-BERT	0.512	0.327
	Keyword Mask	0.254	0.600
sentence 4	Baseline	1.500	0.749
	DP-BERT	0.447	0.356
	Keyword Mask	0.120	0.392
sentence 5	Baseline	1.500	0.749
	Dp-BERT	0.447	0.356
	Keyword Mask	0.120	0.589

Table 3: Prediction scores of 3 models with epsilon 5.5

The table above shows the prediction scores of 3 models as an example while applying epsilon 5.5 to the DP BERT model and keyword mask DP BERT model. We chose to use epsilon value 5.5 for giving the example as our hypothesis is clearly shown by it.

7 Conclusion and Future Work

The situation of Covid-19 is now inevitable, and even more, it requires us to manage our lives while considering the circumstances made by viruses in the future. This situation will make it easier for people to feel frustrated and depressed, and social networking services (SNS) will be a place to vent these feelings out. Even though this was a momentary expression of emotion for a particular situation, things can happen that have a prejudice against a person's mental health status or stigmatize them as racist. Therefore, we suggested employing a method that can protect the sensitive information and, most importantly, sentiments of people for sentimental classification deep learning models.

For our project, we set two objectives; firstly, the Differentially Private model's performance should be the same or at least similar to the baseline model even after applying epsilon noise. The reason to focus on this objective is that models' privacy and performance trade-offs can be obstacles to commercializing privacy-applied models. Secondly, we aim to prevent attackers from inferring the emotions expressed by users. Deep neural networks are prone to memorization problems, making it possible for attackers to guess the words based on the model's prediction score. Thus, we suggest a method that can solve this problem.

To conclude, we have successfully achieved results for both of our goals. Our experiments highlighted that the DP BERT models do not reach the same level of accuracy as the baseline model. In addition, we found out the DP BERT model is not a perfect way to solve the memorization problem. The Keyword mask BERT model provides a significant direction to solve these two problems of performance and memorization. The keyword mask DP BERT model confirmed that even if a small epsilon value(0.5) was used, the accuracy did not decrease significantly compared to the DP BERT model. Furthermore, Keyword mask DP BERT models tend to yield better accuracy even with the lowest value of epsilon used in our experiments. With the baseline and DP BERT models, we can infer that specific sentences from the dataset are based on their prediction scores. In contrast, the keyword mask DP BERT model makes it impossible to guess if those sentences are from the dataset. Keyword Mask BERT model shows significant outcomes when classifying given sentences as positive or negative because it makes different decisions from other models, as shown in our results section.

Future Work: In our experiments, we have only considered the BERT model and protected the sensitive information while training the BERT model for our task. One could consider applying differential privacy to the Long Short Term Memory (LSTM) model which is also heavily used as a sentiment analysis task and compare how does the model change would affect the privacy-utility trade-off.

Moreover, due to the time constraints on our project, we only considered the time frame from 20 March 2020, to 31 March 2020 for COVID-19 sentiment analysis on tweets. As a result, we realized that our dataset is imbalanced that has more positive sentiment than negative sentiment because it was the onset period of the COVID-19 situation and people were still hopeful. Future work could consider applying our DP BERT model and Keyword masked DP BERT model to the COVID-19 related dataset that is more balanced.

8 Github Repository

The implementation of our methodology is available at `https://github.com/JauraSeerat/CMPUT-622-Project.git`

Acknowledgement

We are grateful to our course professor Dr. Nidhi Hegde for guiding us in getting a good direction to our task.

References

[1] Tim Mittermeier, Matthias Frank, Sabine Ullrich, Gabi Dreo Rodosek, and Michaela Geierhos. A multimodal mixed reality data exploration framework for tactical decision making. In *2021 International Conference on Military Communication and Information Systems (ICMCIS)*, pages 1–8. IEEE, 2021.

[2] Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer. Thieves on sesame street! model extraction of bert-based apis. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=Byl5NREFDr`.

[3] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 267–284, USA, 2019. USENIX Association. ISBN 9781939133069.

[4] Priyam Basu, Tiasa Singha Roy, Rakshit Naidu, Zümrüt Müftüoglu, Sahib Singh, and Fatemehsadat Mireshghallah. Benchmarking differential privacy and federated learning for BERT models. *CoRR*, abs/2106.13973, 2021. URL `https://arxiv.org/abs/2106.13973`.

[5] Akash Dutt Dubey. The resurgence of cyber racism during the covid-19 pandemic and its aftereffects: Analysis of sentiments and emotions in tweets. *JMIR Public Health Surveill*, 6 (4):e19833, Oct 2020. ISSN 2369-2960. doi: 10.2196/19833. URL `http://publichealth.jmir.org/2020/4/e19833/`.

[6] Rabindra Lamsal. Coronavirus (covid-19) geo-tagged tweets dataset, 2020. URL `https://dx.doi.org/10.21227/fpsb-jz61`.

[7] Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. Privacy preserving text representation learning. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, page 275–276, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368858. doi: 10.1145/3342220.3344925. URL `https://doi.org/10.1145/3342220.3344925`.

[8] Kennedy J. *Evolution of the PAN Lab on Digital Text Forensics*, pages 461–485. Springer International Publishing, Cham, 2019. ISBN 978-3-030-22948-1. doi: 10.1007/978-3-030-22948-1_19. URL `https://doi.org/10.1007/978-3-030-22948-1_19`.

[9] Fatemehsadat Mireshghallah, Huseyin Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert Sim. Privacy regularization: Joint privacy-utility optimization in LanguageModels. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3799–3807, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.298. URL `https://aclanthology.org/2021.naacl-main.298`.

[10] Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1001. URL `https://aclanthology.org/D18-1001`.

[11] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL `http://dx.doi.org/10.1561/0400000042`.

[12] Ekaba Bisong. *Google Colaboratory*, pages 59–64. Apress, Berkeley, CA, 2019. ISBN 978-1-4842-4470-8. doi: 10.1007/978-1-4842-4470-8_7. URL `https://doi.org/10.1007/978-1-4842-4470-8_7`.

[13] Steven Loria. textblob documentation. *Release 0.15*, 2, 2018.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

[15] A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao, G. Cormode, and I. Mironov. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.

[16] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. URL `http://arxiv.org/abs/1711.05101`.