CAPE: Context-Aware Private Embeddings for Private Language Learning

Richard Plant, Valerio Giuffrida, Dimitra Gkatzia
Edinburgh Napier University
{r.plant, v.giuffrida, d.gkatzia}@napier.ac.uk

Abstract

Neural language models have contributed to state-of-the-art results in a number of downstream applications including sentiment analysis, intent classification and others. However, obtaining text representations or embeddings using these models risks encoding personally identifiable information learned from language and context cues that may lead to privacy leaks. To ameliorate this issue, we propose *Context-Aware Private Embeddings (CAPE)*, a novel approach which combines differential privacy and adversarial learning to preserve privacy during training of embeddings. Specifically, CAPE firstly applies calibrated noise through differential privacy to maintain the privacy of text representations by preserving the encoded semantic links while obscuring sensitive information. Next, CAPE employs an adversarial training regime that obscures identified private variables. Experimental results demonstrate that our proposed approach is more effective in reducing private information leakage than either single intervention, with approximately a 3% reduction in attacker performance compared to the best-performing current method.

1 Introduction

Deep learning has provided remarkable advances in language understanding and modelling tasks in recent years (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020). However, this increased utility may harm user privacy, as neural models trained with datasets containing personal identifiable information can unintentionally leak information that users may prefer to keep private (Carlini et al., 2019; Song et al., 2017). Even seemingly innocuous collections of metadata (Xu et al., 2008) such as data provided by the users (e.g. at registration time on social media) or data which has been cleansed of identifying attributes (Sun et al., 2012), can provide *latent* information for the re-identification of participants.

Using social media data can also raise ethical considerations (Townsend and Wallace, 2016). Users may have edited or deleted posts that models continue to rely on in existing datasets, and may unintentionally reveal information they would rather keep private (Bartunov et al., 2012; Pontes et al., 2012; Goga et al., 2013). Research has shown practical attacks that exploit trained models to establish whether a particular individual formed part of a model's training dataset, in an attack known as membership inference (Leino and Fredrikson, 2020; Truex et al., 2019). Personally identifiable attributes such as age, gender, or location can be reliably reconstructed given the output of such a model (Fredrikson et al., 2015; Zhang et al., 2020). Neural representations of input data, including language embeddings, have proven to be a vulnerability for these inferences (Song and Raghunathan, 2020), thus privacy-preserving techniques should be applied to these text representations when they form part of a machine learning pipeline.

To minimise the risk of such attacks in uncovering sensitive information, previous work has employed an adversarial training objective (Coavoux et al., 2018; Li et al., 2018) by modifying the loss function of the model to impose a penalty when a simulated attacker task, such as predicting a private variable from the input sequence, performs well. However, this approach provides no formal privacy guarantees nor privacy loss accounting system. Phan et al. (2020) proposed an approach which implements classical differential privacy in an adversarial learning paradigm, however, this work relies on adversarial objectives to promote robustness to adversarial samples rather than privacy.

Providing a privacy guarantee leads to the notion of differential privacy (DP), as defined by Dwork and Roth (2013). This definition quantifies privacy loss as the maximum possible deviation between the same aggregate function applied to two datasets which differ only in a single record, which can be expressed by the variable ϵ.

Definition 1.1 (ϵ-differential privacy)**.** *The level of private information leaked by a computation M can be expressed by the variable ϵ where for any two data sets A and B, and any set of possible outputs $S \subset Range(M)$,*

$$Pr[M(A) \in S] \leq Pr[M(B) \in S] \times exp(\epsilon \times |A \oplus B|)$$

This notion of ϵ-differential privacy has been extended to text embeddings through the application of calibrated noise (Fernandes et al., 2019; Beigi et al., 2019). Lyu et al. (2020) proposed a method based on local differential privacy—an extension to the schema under which noise is applied to the input data before it leaves the user's device and is encountered by the model owner—producing a private representation which can be sent to a server for classification. However, this approach uses simulated attacker performance as a test benchmark for private information leakage, rather than during training to improve privacy outcomes.

Determining the state-of-the-art in a task of relatively recent provenance and with somewhat limited practical research such as this proves challenging, however we consider the adversarial learning approach of Coavoux et al. (2018) and the local DP approach of Lyu et al. (2020) the focus of the most current research (Alnasser et al., 2021; Dayanik and Padó, 2021; Kaneko and Bollegala, 2021; Friedrich et al., 2019; Vu et al., 2019).

Contributions: In this work, we propose an approach that combines perturbed pre-trained embeddings with a privacy-preserving adversarial training function that helps preserving the encoded semantic links in the input text while obscuring sensitive information. We demonstrate that our approach achieves comparable task performance against a competitive baseline while preserving privacy. We experiment with a dataset that contains personally identifiable information namely gender, location and birth year. To minimize harm, we experiment with a publicly available English-language dataset (Hovy et al., 2015). Specifically:

- We introduce CAPE, "Context-Aware Private Embeddings"[1], an approach that applies both DP-compliant perturbations and an adversarial learning objective to privatize the embedding outputs of pre-trained language models.
- We establish metrics for testing the privacy result of our system against non-DP-compliant models by offering an empirical framework for determining the level of success of simulated attacks.
- We find that attacker inferences demonstrate differing levels of accuracy depending on the type of the private attribute targeted.
- We establish superior privacy outcomes for our method compared to a sample adversarial learning approach (Coavoux et al., 2018) and a perturbation-only method (Lyu et al., 2020) representing the dominant approaches currently applied to other task domains.

2 Methodology

We consider the possibility that an attacker may have access to the intermediate feature representations extracted from text from a published language model along with a supervision signal that may allow them to train a model to recover private information about the text author, possibly garnered from access to a secondary data source as demonstrated in Narayanan and Shmatikov (2008) and Carlini et al. (2020). To mitigate this risk, we introduce a DP-compliant layer to the feature extractor that perturbs the representations by adding calibrated noise. We train a second classifier to predict known private variables in addition to our main target task classifier, then pass the error gradient from the secondary classifier through a reversal layer to promote embedding invariance to the private features. Figure 1 shows the system architecture.

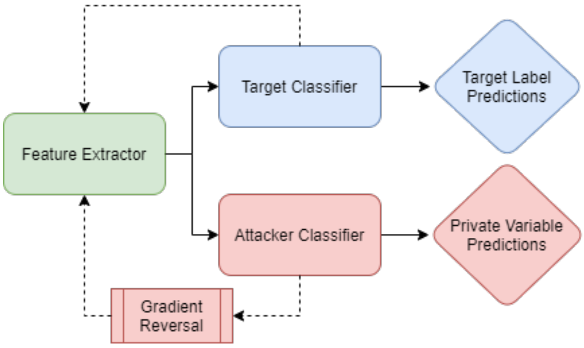

Figure 1: CAPE model diagram. Solid lines indicate data flow, dotted lines indicate gradient updates.

2.1 Task Formulation & Data

We experiment with multi-class sentiment analysis on the UK section of the Trustpilot dataset (Hovy

[1] Code base available at https://github.com/NapierNLP/CAPE

et al., 2015), which provides text reviews with an attached numerical rating from 1-5 as well as three demographic attributes: gender, location and birth year. Sentiment analysis from text reviews represents a popular task to which pre-trained language models are well suited. We use the gender as reported in the dataset, as a binary attribute, while birth years are separated into six equal-sized age range bins (<1955, 1955-1963, 1964-1971, 1972-1978, 1979-1985, >1986), and locations are translated from latitude/longitude pairs into Geohash strings with a precision of two characters, which results in five potential location classes. This dataset covers multiple regions and languages, however for ease of implementation we include only English language results from the UK region in these experiments. A summary of this dataset is included in Table 1.

Label	Private Info	Size
rating	gender birth year location (lat/long)	train: 28,000 test: 15,000 validation: 7,000

Table 1: Summary of Trustpilot UK dataset

We treat gender here as a binary categorical variable, since this is the way the value is represented in the dataset. We recognise that this dualism may not fully represent the range of potential gender expressions (Cao and Daumé III, 2020), and would advocate for a wider conception of potential gender representations in further dataset releases. Age of the respondent is listed in the dataset as a year of birth. We separate these values into six equal-sized bins, assigning each bin an integer ID which replaces the year in our input data. The location variable is encoded as a Geohash string [2] of length 2, which translates into a precision of ±630 km. This level of precision avoids the risk of under-poulated classes; with a more extensive dataset it would make sense to increase precision by extending the length of the Geohash string. This set of strings (a total of five possible strings) for our dataset fraction, is also given a categorical integer ID. Thus bucketed, these attributes are suitable variables for classification modelling.

In our initial baseline experiment, we train a feature extractor consisting of a pre-trained BERT model (Devlin et al., 2019) along with two dense layers in order to extract useful features from the input text x. We obtain the final hidden state of the pre-trained model for each token in the input, then take a mean average over the sequence to produce an embedding for the full text, such that:

Parameter	Search Values	Optimal
Hidden units	128, 256, 512	256
Dropout	0, 0.2, 0.4	0.4
Learning rate	0.01, 0.001, 0.0001	0.0001

Table 2: Model hyper-parameters

$$x_e = f(x) \tag{1}$$

Sentiment analysis is then carried out by a classifier which learns to predict the review rating label y given the embedding vector.

Layer size, dropout rate and other hyper-parameters were optimised with a grid search, selecting the most effective with respect to the target task F1 score metric. Optimal parameters are shown in Table 2.

A sample setup as created for CAPE model testing is shown in Appendix A. Adversarial only, differentially-private only, and baseline setups are similar, omitting the noise layer, attacker classifier, or both respectively.

We simulate a task that an attacker may wish to perform on the input text by training a secondary classifier along with the target task that attempts to predict the value of private information variables z. Following Coavoux et al. (2018), we target several features of the respondent as extracted from the dataset, namely gender, location, and birth year. These features, while in reality not being private by virtue of being public information provided by users, represent good proxies for sensitive attributes that users may not wish to be inferred from similar public datasets. In this sense, they provide a useful benchmark of the potential privacy risk, while allowing us to avoid unethical inferences concerning private attributes not shared by the user.

2.2 Adversarial Training

In order to promote invariance in the text representation with respect to our private variables, we adopt the approach pioneered by Ganin et al. (2017). Initially designed to promote domain-independent learning, this system involves training a secondary objective to predict features we do not wish to be distinguishable via gradient descent, then passing the loss through a gradient reversal

[2] https://github.com/vinsci/geohash/

layer into a target task objective, represented in our experiments by the feature extractor.

For a single instance of our data (x_e, y, z) the adversarial classifier optimizes:

$$\mathcal{L}_a(x_e, y, z; \theta_a) = -logP(z|x_e; \theta_a) \qquad (2)$$

Hence, the combination of both target and attacker classifiers lead to the following objective function, where θ_r, θ_p, θ_a represent the parameters of the feature extractor, classifier and adversarial classifier respectively:

$$\begin{aligned}\mathcal{L}(x_e, y, z; \theta_r, \theta_p, \theta_a) = &- logP(y|x_e; \theta_r, \theta_p) \\ &- \lambda logP(\neg z|x_e; \theta_a)\end{aligned} \qquad (3)$$

where $\neg$ indicates that the log likelihood of the private label z is inverted, and λ is the regularization parameter scaling the gradient from our adversarial classifier.

The combined classification section therefore consists of two separate classification heads, one for our base task and one for our simulator attacker task. Each consists of two densely-connected layers separated by a dropout layer. The attacker classifier includes a gradient reversal layer which flips the sign of the gradient during the backwards pass.

2.3 Embedding Perturbation

Since it is also desirable to provide a measure of general privacy alongside the specific attacker task we simulate in our adversarial training, we initially adopted the local DP method of Lyu et al. (2020) to perturb the feature representations we produce. Converting the generated embedding into a DP-compliant representation requires us to inject calibrated Laplace noise into the hidden state vector obtained from the pre-trained language model as follows:

$$\tilde{x}_e = x_e + n \qquad (4)$$

where n is a vector of equal length to x_e containing i.i.d. random variables sampled from the Laplace distribution centred around 0 with a scale defined by $\frac{\Delta f}{\epsilon}$, where ϵ is the privacy budget parameter and Δf is the sensitivity of our function.

Since determining the sensitivity of an unbounded embedding function is practically infeasible, we followed the initial work in constraining the range of our representation to [0,1], as recommended by Shokri and Shmatikov (2015). In this way, the maximum L1 norm of our function summed across n dimensions of x_e is 1. However, as detailed in Maheshwari et al. (2022), Lyu et al. make a fundamental error in their algorithm, resulting in a real sensitivity of $1 * n$. We adopt instead the corrected methodology proposed by Maheshwari et al., resulting in a maximal sensitivity of 2 and concordant noise sampled from a distribution of $Lap(\frac{2}{\epsilon})$. Results obtained using the original methodology are preserved in Appendix B.

```
Algorithm 1: Context-Aware Private Embeddings (CAPE)
Input: Input data $x$, Label $y$, Private label $z$
Extract features from input: $x_e = f(x)$;
Normalise representation: $x_e \leftarrow x_e/\|x_e\|_1$;
Apply perturbation: $\tilde{x}_e = x_e + Lap(\frac{\Delta f}{\epsilon})$;
Train classifiers: $\mathcal{L}(\tilde{x}_e, y, z; \theta_r, \theta_p) =$
    $-logP(y|\tilde{x}_e; \theta_r, \theta_p) - \lambda logP(\neg z|\tilde{x}_e; \theta_a)$
```

2.4 Context-Aware Private Embeddings

To preserve the general privacy benefits of DP-compliant embeddings with invariance to the specific private variable identified for adversarial training, we combine both processes in a system we call Context-Aware Private Embeddings (CAPE). Algorithm 1 presents the joint adversarial training scheme with perturbed embedding sequences derived from our feature extractor.

3 Evaluation and Results

3.1 Evaluation

We evaluate performance on the target task (i.e. sentiment analysis) and on our simulated attacker task (i.e. classifying each private attribute) with the accuracy metric, as well as providing a measure of the F1-score along with standard deviation of those results. It should be noted that lower results for the attacker classifier denote greater empirical evidence of privacy (*i.e.*, the attacker cannot predict the target variable), and therefore the lowest score in each scenario is indicated in bold, whereas the highest score for the target task is likewise indicated.

All evaluations were performed by randomly selecting 70% of the data for training (the remaining 30% for testing). We compute mean and standard deviation of the F1-score over 4 runs.

3.2 Results

Results for our target and attacker tasks using the English-language only reviews drawn from the UK section of the Trustpilot dataset are listed below. Table 3 shows the results for each system, with ϵ and λ parameters static at 0.1 and 1.0 respectively. These values are derived from a set of experiments with a range of privacy parameter values as detailed in Table 4.

	Target			Attacker		
Approach	Acc.	F1	SD	Acc.	F1	SD
Location						
Base	**0.818**	**0.776**	0.006	0.853	0.785	0.001
Adv.	0.817	0.762	0.005	0.846	0.783	0.009
DP	0.744	0.635	$5e^{-5}$	0.852	0.783	0.004
CAPE	0.746	0.637	0.005	**0.844**	**0.780**	0.002
Gender						
Base	**0.820**	**0.772**	0.003	0.659	0.762	0.006
Adv.	0.818	0.772	0.009	0.632	0.752	0.007
DP	0.744	0.635	$5e^{-5}$	0.605	0.754	0.005
CAPE	0.746	0.637	0.005	**0.603**	**0.751**	0.001
Age Range						
Base	0.816	0.773	0.008	0.234	0.210	0.018
Adv.	**0.824**	**0.779**	0.007	0.188	0.098	0.004
DP	0.744	0.635	$5e^{-5}$	0.179	0.054	0.001
CAPE	0.746	0.637	0.005	**0.177**	**0.053**	0.002

Table 3: Results for the target task and the simulated attacker task. SD = Standard Deviation of F1 score over four cross-validation runs. CAPE outperforms all other approaches in terms of privacy-preservation for all variables.

3.3 Influence of privacy parameters

In order to determine the impact of increasing the stringency of privacy guarantees on performance, we tested our CAPE model with the gender private variable using several values of ϵ while maintaining a value of 1.0 for λ. A similar experiment was carried out for values of λ with ϵ static at 0.1. Results for both experiments are shown in Table 4.

		Target		Attacker	
	Value	Acc.	F1	Acc.	F1
Baseline		0.873	0.816	0.723	0.691
ϵ	0.01	0.741	0.631	0.602	0.747
	0.1	0.747	0.639	0.603	0.751
	0.5	0.744	0.635	0.604	0.753
	1.0	0.749	0.641	0.611	0.758
λ	0.1	0.747	0.739	0.606	0.754
	0.5	0.741	0.630	0.599	0.750
	1.0	0.740	0.620	0.584	0.748
	1.5	0.723	0.614	0.552	0.688

Table 4: Impact of privacy parameters. Lower ϵ and higher λ values lead to increased privacy, but increase performance impact.

4 Discussion and Conclusion

These results demonstrate the enhanced privacy afforded by the CAPE approach over either privacy approach applied in isolation. We provide evidence that adversarial training can produce superior outcomes to a DP-only approach, if we consider the private variable targeted in training. Adding DP noise clearly harms performance outcomes, indicating that we require further work to implement alternate processes for perturbing embeddings. Perturbed embeddings generated in Euclidean space perform more poorly as the privacy guarantee increases, so projecting embeddings into Hyperbolic space (Dhingra et al., 2018) or implementing a search mechanism to select semantically-similar vectors that represent real words (Feyisetan et al., 2020) could produce better outcomes with lower privacy budgets.

Interestingly, we find that different private attributes are predictable by an attacker at different rates—while the attacker can predict the correct gender or location class effectively, results for age range are barely above random chance. It may well be the case in the UK that word choice varies more between areas and genders than age cohorts, for example, a reviewer who cites the product's "lush vanilla taste" may reside in the West of England, while calling a bad service "shite" may indicate they are Scottish. This is an interesting counter-finding to Welch et al. (2020) which found better embedding performance with age- and gender-aware representations in a global population. Differing privacy requirements for separate attributes are a feature of multiple variations on differential privacy regimes (Kamalaruban et al., 2020; Alaggan et al., 2017; Jorgensen et al., 2015).

We note finally that English exhibits fewer grammatical markers that indicate gender than some other languages (Boroditsky and Schmidt, 2000), a peculiarity which may affect the utility of the model in significant ways. Further exploration on different language families will shed light on how privacy-preserving methods can assist in concealing private information.

Acknowledgements

DG's contribution is supported under the EPSRC projects CiViL (EP/T014598/1) and NLG for low-resource domains (EP/T024917/1).

References

Mohammad Alaggan, Sébastien Gambs, and Anne-Marie Kermarrec. 2017. Heterogeneous differential privacy. *Journal of Privacy and Confidentiality*, 7(2).

Walaa Alnasser, Ghazaleh Beigi, and Huan Liu. 2021. Privacy Preserving Text Representation Learning Using BERT. In *Social, Cultural, and Behavioral Modeling*, Lecture Notes in Computer Science, pages 91–100, Cham. Springer International Publishing.

Sergey Bartunov, Anton Korshunov, Seung-taek Park, Wonho Ryu, and Hyungdong Lee. 2012. Joint Link-Attribute User Identity Resolution in Online Social Networks Categories and Subject Descriptors. In *The Sixth SNA-KDD Workshop Proceedings*, volume 12.

Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. 2019. Privacy preserving text representation learning. In *HT 2019 - Proceedings of the 30th ACM Conference on Hypertext and Social Media*, pages 275–276. ArXiv: 1907.03189.

Lera Boroditsky and Lauren A. Schmidt. 2000. Sex, Syntax, and Semantics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 22(22).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret Sharer: Evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Security Symposium*, pages 267–284. USENIX Association. ArXiv: 1802.08232.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. *CoRR*, abs/2012.07805.

Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. 2018. Privacy-preserving Neural Representations of Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Brussels, Belgium. Association for Computational Linguistics.

Erenay Dayanik and Sebastian Padó. 2021. Disentangling Document Topic and Author Gender in Multiple Languages: Lessons for Adversarial Debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–61, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Christopher J. Shallue, Mohammad Norouzi, Andrew M. Dai, and George E. Dahl. 2018. Embedding text in hyperbolic spaces. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Student Research Workshop*, pages 59–69. Association for Computational Linguistics (ACL).

Cynthia Dwork and Aaron Roth. 2013. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–487.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised Differential Privacy for Text Document Processing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11426 LNCS, pages 123–148. Springer Verlag.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy- And utility-preserving textual analysis via calibrated multivariate perturbations. In *WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 178–186. Association for Computing Machinery, Inc.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the ACM Conference on Computer and Communications Security*, volume 2015-Octob, pages 1322–1333, New York, New York, USA. Association for Computing Machinery.

Max Friedrich, Arne Köhn, Gregor Wiedemann, and Chris Biemann. 2019. Adversarial Learning of

Privacy-Preserving Text Representations for De-Identification of Medical Records. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 5829–5839. ArXiv: 1906.05000 Publisher: Association for Computational Linguistics (ACL).

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2017. Domain-adversarial training of neural networks. In *Advances in Computer Vision and Pattern Recognition*, volume 17, pages 189–209. Springer London.

Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. 2013. Exploiting innocuous activity for correlating users across sites. In *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, pages 447–457.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User Review Sites as a Resource for Large-Scale Sociolinguistic Studies. In *WWW '15: Proceedings of the 24th International Conference on World Wide Web*, pages 452–461.

Z. Jorgensen, T. Yu, and G. Cormode. 2015. Conservative or liberal? Personalized differential privacy. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1023–1034. ISSN: 2375-026X.

Parameswaran Kamalaruban, Victor Perrier, Hassan Jameel Asghar, and Mohamed Ali Kaafar. 2020. Not All Attributes are Created Equal: dX -Private Mechanisms for Linear Queries. *Proceedings on Privacy Enhancing Technologies*, 2020(1):103–125. Publisher: Sciendo Section: Proceedings on Privacy Enhancing Technologies.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing Pre-trained Contextualised Embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Klas Leino and Matt Fredrikson. 2020. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *Proceedings of the 29th USENIX Security Symposium*, pages 1605–1622. USENIX Association. ArXiv: 1906.11798.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 2, pages 25–30. Association for Computational Linguistics (ACL). ArXiv: 1805.06093.

Lingjuan Lyu, Xuanli He, and Yitong Li. 2020. 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*, pages 2355–2365.

Gaurav Maheshwari, Pascal Denis, Mikaela Keller, and Aurélien Bellet. 2022. Fair NLP Models with Differentially Private Text Encoders. Technical Report arXiv:2205.06135, arXiv. ArXiv:2205.06135 [cs] type: article.

Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *Proceedings - IEEE Symposium on Security and Privacy*, pages 111–125.

Hai Phan, My T. Thai, Han Hu, Ruoming Jin, Tong Sun, and Dejing Dou. 2020. Scalable differential privacy with certified robustness in adversarial learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7683–7694. PMLR.

Tatiana Pontes, Gabriel Magno, Marisa Vasconcelos, Aditi Gupta, Jussara Almeida, Ponnurangam Kumaraguru, and Virgilio Almeida. 2012. Beware of what you share: Inferring home location in social networks. In *Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*, pages 571–578.

Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *Proceedings of the ACM Conference on Computer and Communications Security*, volume 2015-Octob, pages 1310–1321.

Congzheng Song and Ananth Raghunathan. 2020. Information Leakage in Embedding Models. In *Proceedings of the ACM Conference on Computer and Communications Security*, pages 377–390.

Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. 2017. Machine learning models that remember too much. In *Proceedings of the ACM Conference on Computer and Communications Security*, pages 587–601. ArXiv: 1709.07886 ISSN: 15437221.

Xiaoxun Sun, Hua Wang, and Yanchun Zhang. 2012. On the identity anonymization of high-dimensional rating data. In *Concurrency Computation Practice and Experience*, volume 24, pages 1108–1122. John Wiley & Sons, Ltd. Issue: 10 ISSN: 15320626.

Leanne Townsend and Claire Wallace. 2016. Social Media Research: A Guide to Ethics. Technical report, University of Aberdeen.

Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2019. Demystifying Membership Inference Attacks in Machine Learning as a Service. *IEEE Transactions on Services Computing*. ArXiv: 1807.09173.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Long Beach, California, USA. Curran Associates Inc.

Xuan Son Vu, Son N. Tran, and Lili Jiang. 2019. dpUGC: Learn Differentially Private Representation for User Generated Contents. *arXiv*. ArXiv: 1903.10453 Publisher: arXiv.

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4076–4089, Online. Association for Computational Linguistics.

Yabo Xu, Ke Wang, Ada Wai Chee Fu, and Philip S. Yu. 2008. Anonymizing transaction databases for publication. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–775, New York, New York, USA. ACM Press.

Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. 2020. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 250–258. Institute of Electrical and Electronics Engineers (IEEE).

A Model Setup

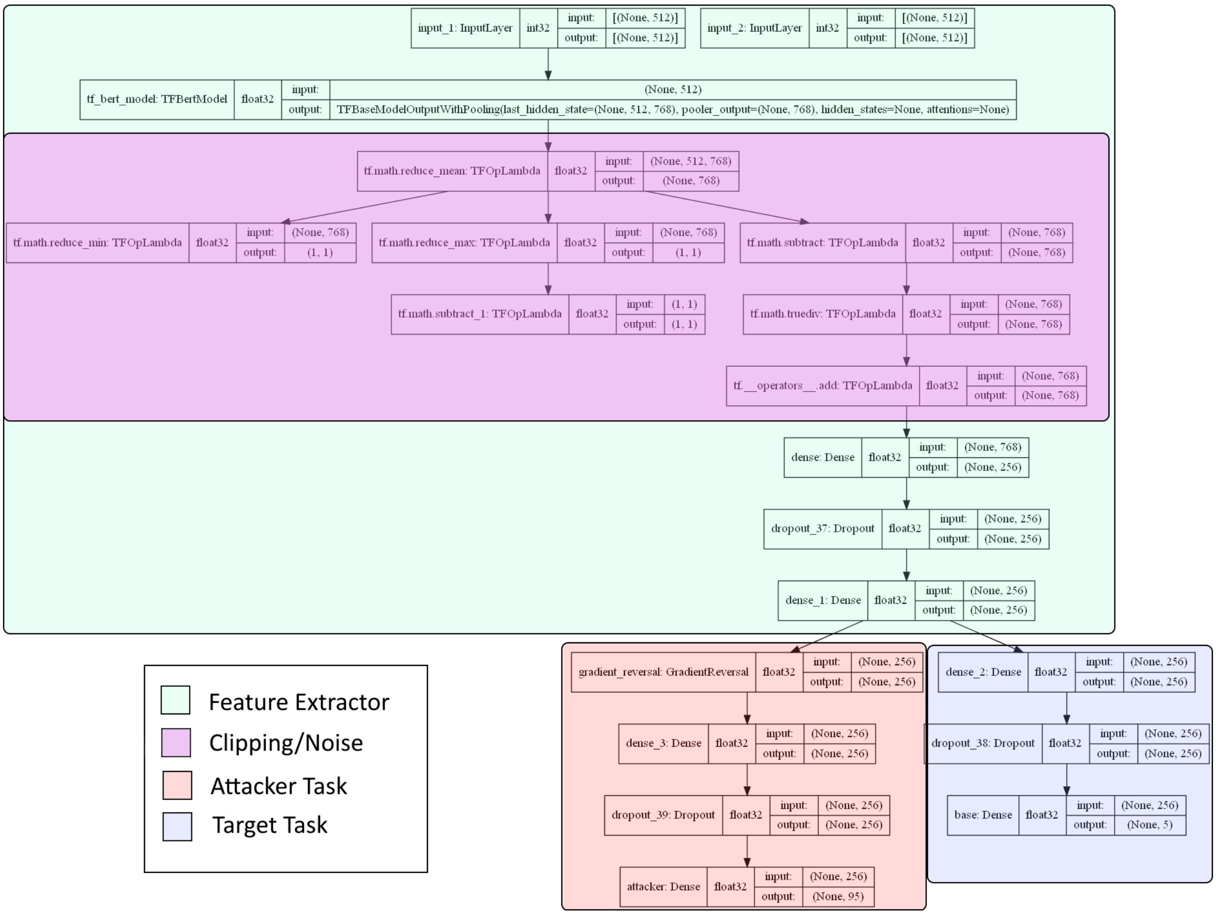

Figure 2: Complete Model Diagram

B Original results

Preserved here are the original results obtained for experiments using the erroneous methodology of Lyu et al. (2020), which fails to constrain the L1 norm of the embedding representation to 1. Results in the main body of the paper instead use the corrected methodology proposed in Maheshwari et al. (2022).

	Target			Attacker		
Approach	Acc.	F1	SD	Acc.	F1	SD
Location						
Base	**0.818**	**0.776**	0.006	0.853	0.785	0.001
Adv.	0.817	0.762	0.005	0.846	0.783	0.009
DP	0.745	0.637	0.000	0.847	0.781	0.004
CAPE	0.748	0.639	0.001	**0.833**	**0.756**	0.009
Gender						
Base	**0.820**	**0.772**	0.003	0.659	0.762	0.006
Adv.	0.818	0.772	0.009	0.632	0.752	0.007
DP	0.746	0.637	0.000	**0.610**	0.755	0.004
CAPE	0.749	0.642	0.005	0.620	**0.733**	0.008
Age Range						
Base	0.816	0.773	0.008	0.234	0.210	0.018
Adv.	**0.824**	**0.779**	0.007	0.188	0.098	0.004
DP	0.744	0.637	0.001	0.183	0.053	0.003
CAPE	0.748	0.634	0.008	**0.171**	**0.052**	0.002

Table 5: Results for the target task and the simulated attacker task. SD = Standard Deviation of F1 score over four cross-validation runs.

		Target		Attacker	
	Value	Acc.	F1	Acc.	F1
Baseline		0.873	0.816	0.723	0.691
ϵ	0.01	**0.847**	**0.777**	**0.498**	0.662
	0.1	0.846	0.774	0.498	**0.655**
	0.5	0.844	0.772	0.496	0.669
	1.0	0.843	0.772	0.511	0.678
λ	0.1	**0.848**	**0.778**	0.512	0.676
	0.5	0.841	0.768	0.506	0.672
	1.0	0.839	0.766	**0.498**	0.674
	1.5	0.847	0.776	0.499	**0.669**

Table 6: Impact of privacy parameters.