Privacy-Preserving Federated Learning for Hate Speech Detection

Ivo de Souza Bueno Júnior[1*] **Haotian Ye**[12] **Axel Wisiorek**[12] **Hinrich Schütze**[12]

[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning (MCML)
[*]*ivosb.junior@gmail.com*

Abstract

This paper presents a federated learning system with differential privacy for hate speech detection, tailored to low-resource languages. By fine-tuning pre-trained language models, ALBERT emerged as the most effective option for balancing performance and privacy. Experiments demonstrated that federated learning with differential privacy performs adequately in low-resource settings, though datasets with fewer than 20 sentences per client struggled due to excessive noise. Balanced datasets and augmenting hateful data with non-hateful examples proved critical for improving model utility. These findings offer a scalable and privacy-conscious framework for integrating hate speech detection into social media platforms and browsers, safeguarding user privacy while addressing online harm.

1 Introduction

Protecting personal data while enabling effective machine learning is a critical challenge, especially in low-resource languages where data scarcity compounds the difficulty of detecting hate speech. Traditional models primarily focus on high-resource languages, leaving underrepresented languages unsupported. Federated learning (FL) with differential privacy (DP) offers a solution by enabling collaborative model training without sharing sensitive data. However, the trade-off between privacy and performance in low-resource settings remains a significant concern. This paper investigates the use of privacy-preserving FL for hate speech detection in low-resource languages, specifically Afrikaans and Russian, which are considered low-resource with regard to labeled hate speech resources, addressing three research questions:

- **(RQ1)** Can privacy-preserving methods effectively support federated hate speech detection models in low-resource languages?
- **(RQ2)** What is the trade-off between privacy and model accuracy in this context?
- **(RQ3)** How minimal can low-resource data be while still ensuring user privacy?

The main contribution of this work is the adaptation of differential privacy within a federated learning framework for hate speech detection in a low-resource environment, and the understanding of the challenges imposed by such systems.

2 Related Work

Hate speech detection has primarily focused on high-resource languages like English. Efforts to address low-resource languages include Ranasinghe and Zampieri (2021), who applied transfer learning to fine-tune transformer models for Arabic, Bengali, and Hindi, showing that pre-trained BERT-based models, like ALBERT, work well in these contexts. Fine-tuning pre-trained models remains a dominant approach, with studies like Geet d'Sa et al. (2020) and Wullach et al. (2021) demonstrating its effectiveness. However, BERT fine-tuning can be unstable, particularly with small datasets, as noted by Mosbach et al. (2021).

Privacy concerns, driven by regulations like the EU's GDPR (of the European Union, 2016), have led to federated learning adoption for decentralized data processing. While early work like Zampieri et al. (2024) showed FL's promise, vulnerabilities in shared model weights have been identified, as seen in Geiping et al. (2020). Differential privacy, introduced by Dwork (2006), mitigates such risks by adding noise to gradients, ensuring privacy while enabling collaborative learning. Both global (Wei et al., 2020) and local (Truex et al., 2020) DP methods in federated learning have shown effectiveness and limitations, as reviewed by Ouadrhiri and Abdelhadi (2022). While recent approaches, such as Ye et al. (2024), leverage FL for few-shot hate speech detection in low-resource languages, this

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 129–141

paper adapts DP to further enhance model security and evaluate its impact on performance.

3 Methods

Dataset. For our experiments, we used hate speech data from two low-resource languages: Afrikaans and Russian. The Afrikaans dataset includes statements targeting black people and LGBTQ+ individuals, while the Russian dataset focuses on hate speech directed at war-affected groups and LGBTQ+ individuals. The datasets were created by native speakers between June 2023 and March 2024 as part of the Respond2Hate research project (Ye et al., 2024). Hate speech examples were inspired by anonymized content from social media and news outlets and were carefully adapted to ensure privacy and cultural relevance. The merged dataset consisted of 1,543 sentences, with 865 (56%) labeled as hateful and 678 (44%) as non-hateful.

Of the 1,543 sentences, 309 were randomly selected as a test set, and the rest were used for fine-tuning. Each client in the federated system received a distinct set of sentences, ensuring non-overlapping data.

Models. Multiple `BERT`-based models were used for various experiments conducted in this work. They are: `BERT Base uncased`, `BERT Large uncased` (Devlin et al., 2019), `HateBERT` (Caselli et al., 2021), `ALBERT Base`, `ALBERT Large`, `ALBERT XLarge`, `ALBERT XXLarge` (Lan et al., 2020), `BERT Base Multilingual uncased` (Devlin et al., 2019), `XLM-RoBERTa Base`, `XLM-RoBERTa Large` (Conneau et al., 2020), and `DistilBERT Base Multilingual cased` (Sanh et al., 2020) More information on the selected models can be seen in Appendix A.

Federated Learning and Differential Privacy Implementation. Federated learning was implemented using the Flower framework (Beutel et al., 2020), which facilitates communication and aggregation between the server and clients. Flower was selected for its support of manual client training steps. Differential privacy was implemented using Opacus (Yousefpour et al., 2021), a PyTorch (Paszke et al., 2019) library that enables DP by adding noise to model gradients. Opacus automatically calculates the noise scale σ based on (ϵ, δ)-DP and the $\mathcal{L}_2$ norm clipping threshold C. PyTorch was used for model fine-tuning, and pre-trained models were sourced from HuggingFace (Wolf et al., 2020).

4 Experiments and Results

4.1 Experimental Setup

The `ALBERT Base` model from Hugging Face was selected for fine-tuning due to its strong performance, as explored in the Model Comparison experiment described below, and seen in Table 1, and efficient fine-tuning times. Privacy parameters were set to $\epsilon = 5$ and $\delta = 10^{-5}$, with a clipping threshold C of 0.5, clipping 1% of the highest gradient values.

The training setup involved one server and eight clients, each receiving 50 balanced sentences (25 hateful, 25 non-hateful). Fine-tuning used a batch size of 1, cross-entropy loss, and the Adam optimizer with a learning rate of 10^{-4} to maintain stability with DP. Baseline experiments included versions without DP ("No DP") and without fine-tuning ("No FT"). For "No DP," the learning rate was reduced to 2×10^{-5} to prevent divergence. All experiments ran for 10 FL rounds.

The weighted F1-score, which is calculated separately for each class, and returned as the weighted sum, was used as the primary evaluation metric due to slight dataset imbalance. Each experiment was run five times, with metrics averaged across clients to minimize variability and account for fine-tuning instabilities. The following experiments were conducted:

Model Comparison. This experiment evaluated the performance of various models fine-tuned with FL and DP for low-resource hate speech detection. Several pre-trained models were tested, but `BERT Large uncased` and `XLM-RoBERTa Large` were excluded due to communication timeouts in FL, likely caused by their large number of parameters. The Flower framework could not handle the computational overhead for these models. No other hyperparameter modifications were made.

Level of Privacy Comparison. The privacy-utility trade-off was tested by fine-tuning the model with various values of ϵ, δ, and clipping threshold C. ϵ values tested ranged from 100 (weak privacy) to 0.1 (strong privacy), with corresponding C values chosen to clip gradients at various percentages: $C = 100$ (no clipping), $C = 0.5$ (1%), $C = 0.1$ (10%), $C = 0.05$ (25%), and $C = 0.01$ (50%). These C values were selected based on observed gradient ranges after initial training rounds. The default $\delta = 10^{-5}$ was used, and `ALBERT Base` and

BERT Base Multilingual models were compared, keeping all other hyperparameters unchanged.

Additionally, different δ values (10^{-3}, 10^{-5}, 10^{-7}) were tested on ALBERT Base with $\epsilon = 5$ to assess their impact on the privacy-utility trade-off. For each δ value, various C values were also tested, with all other hyperparameters kept at their defaults.

Dataset Size Comparison. This test evaluated how the model responded to FL with DP fine-tuning using varying dataset sizes per client. Each client fine-tuned the model with datasets starting at 10 sentences (5 hateful, 5 non-hateful), increasing in increments of 10 up to 130 sentences (65 hateful, 65 non-hateful). All other hyperparameters were kept at their default values and the ALBERT Base model was used.

Dataset Composition Comparison. This experiment tested how different data compositions affected model performance. Three compositions were tested: an "unchanged" composition with the natural imbalance of 56% hateful and 44% non-hateful sentences, a "balanced" composition with 50% hateful and 50% non-hateful sentences, and a "hate-only" composition with only hateful sentences. The "hate-only" composition was tested to simulate a federated system where users report only hateful sentences, and the data is not augmented with negative samples. All other hyperparameters were kept at their default values and the ALBERT Base model was used.

4.2 Results and Analysis

Model	No Diff. Priv.		Diff. Priv.	
	ACC	F1	ACC	F1
BERT Base	0.762	0.762	0.511 (-0.251)	0.395 (-0.367)
HateBERT	0.770	0.770	0.532 (-0.238)	0.415 (-0.355)
ALBERT Base	0.728	0.725	**0.602 (-0.126)**	**0.542 (-0.183)**
ALBERT Large	0.710	0.707	0.513 (-0.197)	0.385 (-0.322)
ALBERT XLarge	0.668	0.663	0.510 (-0.158)	0.353 (-0.310)
ALBERT XXLarge	0.714	0.710	**0.587 (-0.127)**	**0.551 (-0.159)**
BERT Base Multilingual	**0.819**	**0.819**	0.490 (-0.329)	0.403 (-0.416)
XLM-RoBERTa Base	**0.847**	**0.847**	0.489 (-0.358)	0.327 (-0.520)
DistilBERT Base	0.807	0.807	0.524 (-0.283)	0.405 (-0.402)

Table 1: Model comparison between different models fine-tuned by using FL with and without DP. The utility loss between the private and the non-private fine-tuning is shown in red.

Model Comparison. Table 1 shows the results of the model comparison, with best scores marked in bold and utility loss with DP highlighted in red. Multilingual models (BERT Base Multilingual and XLM-RoBERTa Base) performed best in accuracy and F1-score without DP, even in low-resource settings, as they were pre-trained on data containing the low-resource languages used. However, these models suffered the greatest utility loss with DP.

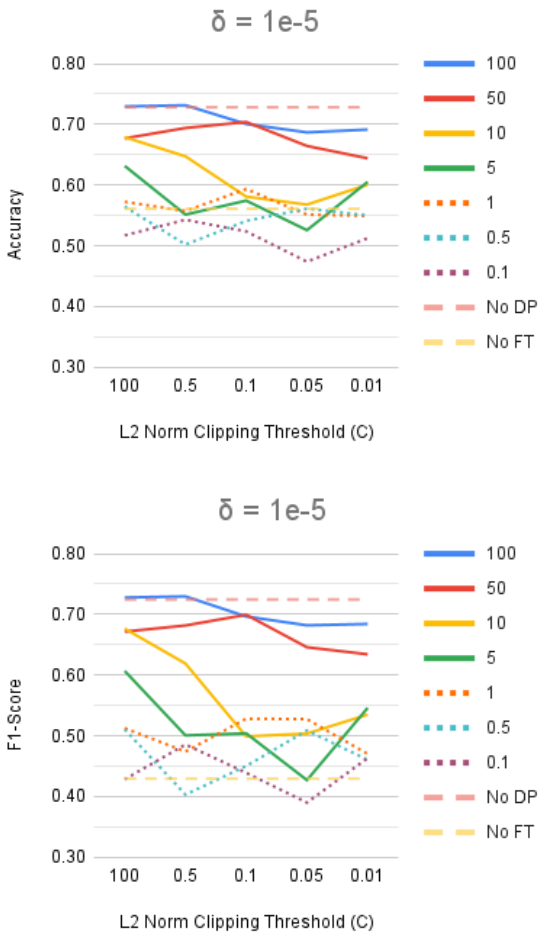

Figure 1: Accuracy and F1-score comparison of different values of ϵ for $\delta = 10^{-5}$.

In contrast, ALBERT models maintained high utility under DP, with ALBERT Base and ALBERT XXLarge showing the lowest utility loss. Their fewer layers (12) compared to the other two ALBERT models (24 layers) likely contributed to this performance. Notably, model size did not significantly affect the privacy-utility trade-off, as ALBERT XXLarge exhibited the lowest utility loss, while XLM-RoBERTa Base showed the highest.

Level of Privacy Comparison. Two experiments assessed the impact of privacy levels on model performance. The first experiment evaluated different ϵ values with $\delta = 10^{-5}$ (Figure 1). As ϵ decreased, indicating stronger privacy, accuracy and F1-scores degraded compared to non-private fine-tuning (No DP). For $\epsilon = 100$ and $\epsilon = 50$,

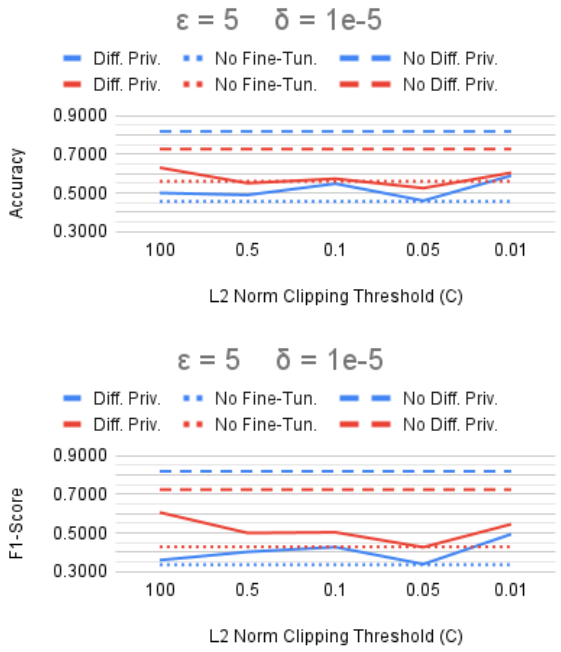

Figure 2: Accuracy and F1-score comparison of BERT (blue) and ALBERT (red) at the same level of privacy ($\epsilon = 5$, $\delta = 10^{-5}$).

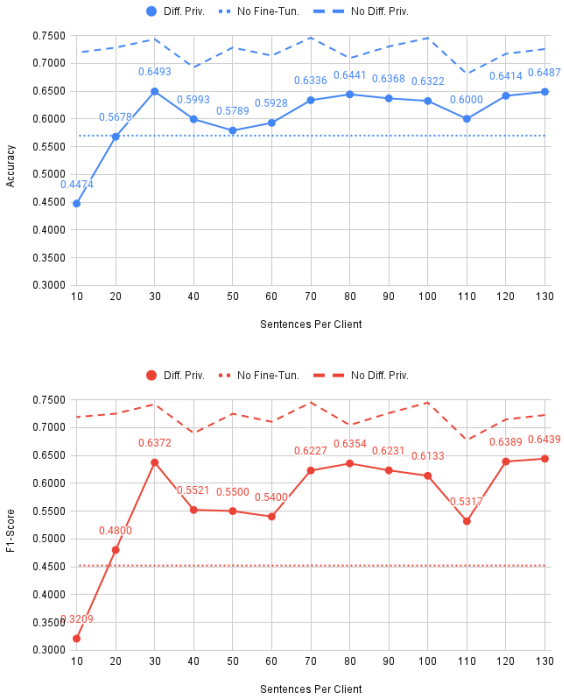

Figure 3: Accuracy (blue, above), and F1-score (red, below), for models fine-tuned with FL clients with different sizes of datasets.

utility loss was moderate but represented weak privacy. Real-world applications typically use $\epsilon < 10$, where performance steeply declined, especially with gradient clipping ($C = 0.5$). At $\epsilon \leqslant 1$, accuracy fell below non-fine-tuned (No FT) levels, and results became noisier. Higher clipping thresholds did not consistently improve scores, particularly at lower ϵ. Similar experiments for $\delta = 10^{-3}$ and $\delta = 10^{-7}$ are shown in Appendix B.

The second experiment evaluated the impact of privacy on fine-tuning ALBERT and BERT for $\epsilon = 5$ and $\delta = 10^{-5}$ (Figure 2). Additional comparisons for other ϵ values are in Appendix C. Without privacy, BERT outperformed ALBERT, but the opposite was true for models without fine-tuning. Both models exhibited similar trends under privacy constraints, hovering near non-fine-tuned levels, with ALBERT achieving higher accuracy and F1-scores than BERT. Notably, BERT showed greater fine-tuning instability, with 29% of runs (51/175) failing to improve after the first FL round, compared to 11% (19/175) for ALBERT.

Varying δ values for a fixed ϵ value offered no relevant insights. These results are in Appendix D.

Dataset Size Comparison. Figure 3 shows the results of the dataset size comparison, with accuracy (blue) and F1-scores (red). As a baseline, we evaluated on the test set by using a model fine-tuned without DP and a model without fine-tuning. The x-axis represents the number of sentences per client during FL.

The model fine-tuned without DP outperforms the one fine-tuned with it, as expected due to the noise introduced by DP. When fine-tuning with very small datasets (10–20 sentences per client), the model performs slightly worse than the non-fine-tuned baseline. This occurs because the noise added by DP is not proportional to the dataset size, leading to parameter updates dominated by noise rather than data.

In this experiment, model performance peaks at 30 sentences per client, achieving an accuracy of 0.64 and an F1-score of 0.63. A similar peak is observed in the non-private fine-tuning version. Figure 4 highlights the difference in accuracy and F1-scores between models fine-tuned with and without DP. A logarithmic interpolation was applied, yielding the best fit with $R^2 = 0.683$ for accuracy and $R^2 = 0.701$ for F1-score, compared to other interpolation methods. The results indicate that as the dataset size increases, the performance of the private model approaches that of the non-private model. However, this trend is not linear and stabilizes eventually, demonstrating that while larger

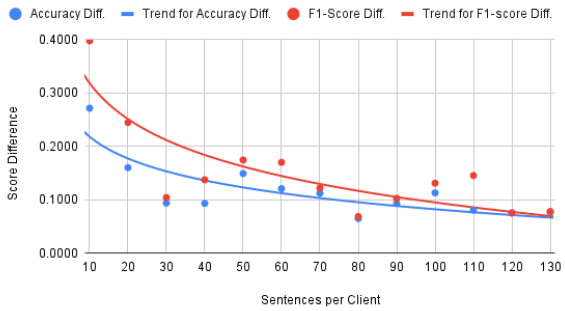

Figure 4: Accuracy (blue) and F1-score (red) difference between models fine-tuned with and without differential privacy, at different dataset sizes.

datasets mitigate the effects of DP noise, they cannot fully eliminate its impact.

Dataset Composition Comparison. Table 2 presents accuracy and F1-scores for different dataset compositions. Results are provided for models evaluated without fine-tuning (No Fine-Tun.) and fine-tuned with (Diff. Priv.) or without differential privacy (No Diff. Priv.). The best metric in each category is highlighted in bold, with utility loss and gain compared to DP fine-tuning marked in red and green, respectively.

The table reveals minimal differences in accuracy between the unchanged and balanced dataset compositions. While the balanced dataset yields higher F1-scores without DP, this advantage disappears under DP fine-tuning. The unchanged dataset composition delivers the best scores and privacy-utility trade-off when fine-tuning with DP, which could point out that having a slight imbalance towards hateful sentences might be advantageous.

As expected, fine-tuning exclusively on hateful sentences, regardless of DP, performs worse in both accuracy and F1-scores than skipping fine-tuning altogether.

Data Comp.	Accuracy		
	Diff. Priv.	No Fine-Tun.	No Diff. Priv.
Unchanged	**0.608**	**0.561** (-0.047)	0.721 (0.113)
Balanced	0.604	**0.561** (-0.043)	**0.742** (0.138)
Hate-Only	0.553	**0.561** (0.008)	0.553 (0.000)
	F1-Score		
	Diff. Priv.	No Fine-Tun.	No Diff. Priv.
Unchanged	**0.565**	**0.429** (-0.136)	0.719 (0.154)
Balanced	0.558	**0.429** (-0.129)	**0.741** (0.183)
Hate-Only	0.406	**0.429** (0.023)	0.396 (-0.010)

Table 2: Dataset composition comparison.

5 Discussion

This paper investigates federated learning with differential privacy for hate speech detection in low-resource environments. Results show that this approach is feasible for fine-tuning models, even with limited data, but models react differently to added noise. `ALBERT` models (`Base` and `XXLarge`) performed the best due to parameter sharing, which might have mitigated the noise. Deeper and multilingual models experienced greater utility loss, though further research is needed to confirm these findings.

Achieving strong privacy guarantees remains challenging. At $\epsilon \leqslant 1$, performance dropped below the non-fine-tuned baseline, highlighting the difficulty of selecting optimal ϵ values, which depend on the model, dataset, and parameter interactions.

More local data per client improved results, with 50 sentences per client showing consistent gains. However, limited data hampers effective learning under differential privacy. Balanced datasets are critical, but a slight imbalance towards hateful sentences helped overcome the noise added by differential privacy. Sampling non-hateful examples is crucial for effective training. Despite challenges, federated learning with differential privacy remains advantageous where privacy is paramount.

Addressing the research questions:

- **(RQ1)** Privacy-preserving federated learning for hate speech detection in low-resource languages is feasible, but may not meet strong privacy standards without sufficient data.
- **(RQ2)** The privacy-utility trade-off is significant, with better results achievable at lower privacy levels.
- **(RQ3)** For minimal data, 50 sentences per client suffice for moderate privacy, though more data reduces degradation and stabilizes training.

6 Conclusion

This paper explored federated learning with differential privacy for hate speech detection in low-resource settings. Fine-tuning a pre-trained `ALBERT` model showed improved performance at moderate privacy levels. Key findings included the importance of nearly-balanced datasets and the impact of differential privacy parameters (ϵ, δ, and C), with `ALBERT` outperforming other `BERT`-based models.

The results addressed the research questions, highlighting both strengths and areas for improvement.

In conclusion, despite challenges in low-resource environments, federated learning with differential privacy can effectively detect hate speech while ensuring user privacy.

7 Limitations

This paper has several limitations. Training required each client to store a local model, limiting experiments to eight clients, and the use of smaller, BERT-based models, instead of LLMs, due to memory constraints. Future work could explore varying client numbers and adaptive clipping thresholds, which were untested due to fixed C values in Opacus. Adaptive methods, as proposed by Andrew et al. (2021), could improve performance. Additionally, non-BERT models like GPT or LLaMA were not evaluated. Finally, the number of federated learning rounds and epochs was not varied, but exploring these hyperparameters may impact model performance.

Acknowledgements

The data used in this research was created within the ERC Proof of Concept project Respond2Hate, funded by the European Research Council (ERC) under Grant No. 101100870.

References

Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. 2021. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, volume 34, pages 17455–17466. Curran Associates, Inc.

Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D. Lane. 2020. Flower: A friendly federated learning research framework. *arXiv*, abs/2007.14390.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 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.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, ICALP'06, pages 1–12, Berlin, Heidelberg. Springer.

Ashwin Geet d'Sa, Irina Illina, and Dominique Fohr. 2020. Classification of Hate Speech Using Deep Neural Networks. *Revue d'Information Scientifique & Technique*, 25(01).

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. 2020. Inverting gradients - how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 16937–16947. Curran Associates, Inc.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, ICLR2020.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*, ICLR2021.

Official Journal of the European Union. 2016. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance). https://eur-lex.europa.eu/eli/reg/2016/679/oj. [Accessed 27-06-2024].

Ahmed El Ouadrhiri and Ahmed Abdelhadi. 2022. Differential privacy for deep and federated learning: A survey. *IEEE Access*, 10:22359–22380.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang,

Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual offensive language identification for low-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(1).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv*, abs/1910.01108.

Stacey Truex, Ling Liu, Ka-Ho Chow, Mehmet Emre Gursoy, and Wenqi Wei. 2020. LDP-Fed: federated learning with local differential privacy. In *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, EdgeSys '20, page 61–66, New York, NY, USA. Association for Computing Machinery.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony Q. S. Quek, and H. Vincent Poor. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Ye, Axel Wisiorek, Antonis Maronikolakis, Özge Alaçam, and Hinrich Schütze. 2024. A federated approach to few-shot hate speech detection for marginalized communities. *arXiv*, abs/2412.04942.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. 2021. Opacus: User-friendly differential privacy library in PyTorch. *arXiv*, abs/2109.12298.

Marcos Zampieri, Damith Premasiri, and Tharindu Ranasinghe. 2024. A federated learning approach to privacy preserving offensive language identification. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 12–20, Torino, Italia. ELRA and ICCL.

A Model Information

- BERT Base uncased (Devlin et al., 2019) (110M parameters), and BERT Large uncased (336M parameters), both trained with monolingual English data.
- A BERT-based model trained on hate speech data: HateBERT (Caselli et al., 2021) (110M parameters, monolingual English).
- ALBERT Base (Lan et al., 2020) (11M parameters), Large (17M parameters), XLarge (58M parameters), and XXLarge (223M parameters), all trained with monolingual English data.
- BERT Base Multilingual uncased (Devlin et al., 2019) (110M parameters), pre-trained using multilingual data from Wikipedia in 102 languages, including Afrikaans and Russian.
- XLM-RoBERTa Base (Conneau et al., 2020) (270M parameters), and XLM-RoBERTa Large (550M parameters), both pre-trained using multilingual data from CommonCrawl in 100 languages, including Afrikaans and Russian.
- DistilBERT Base Multilingual cased (Sanh et al., 2020) (134M parameters), which is a distilled version of BERT Base Multilingual Cased, which was pre-trained using multilingual data from Wikipedia in 104 languages, including Afrikaans and Russian.

B Privacy comparison with different values of ϵ, for the model `ALBERT Base`.

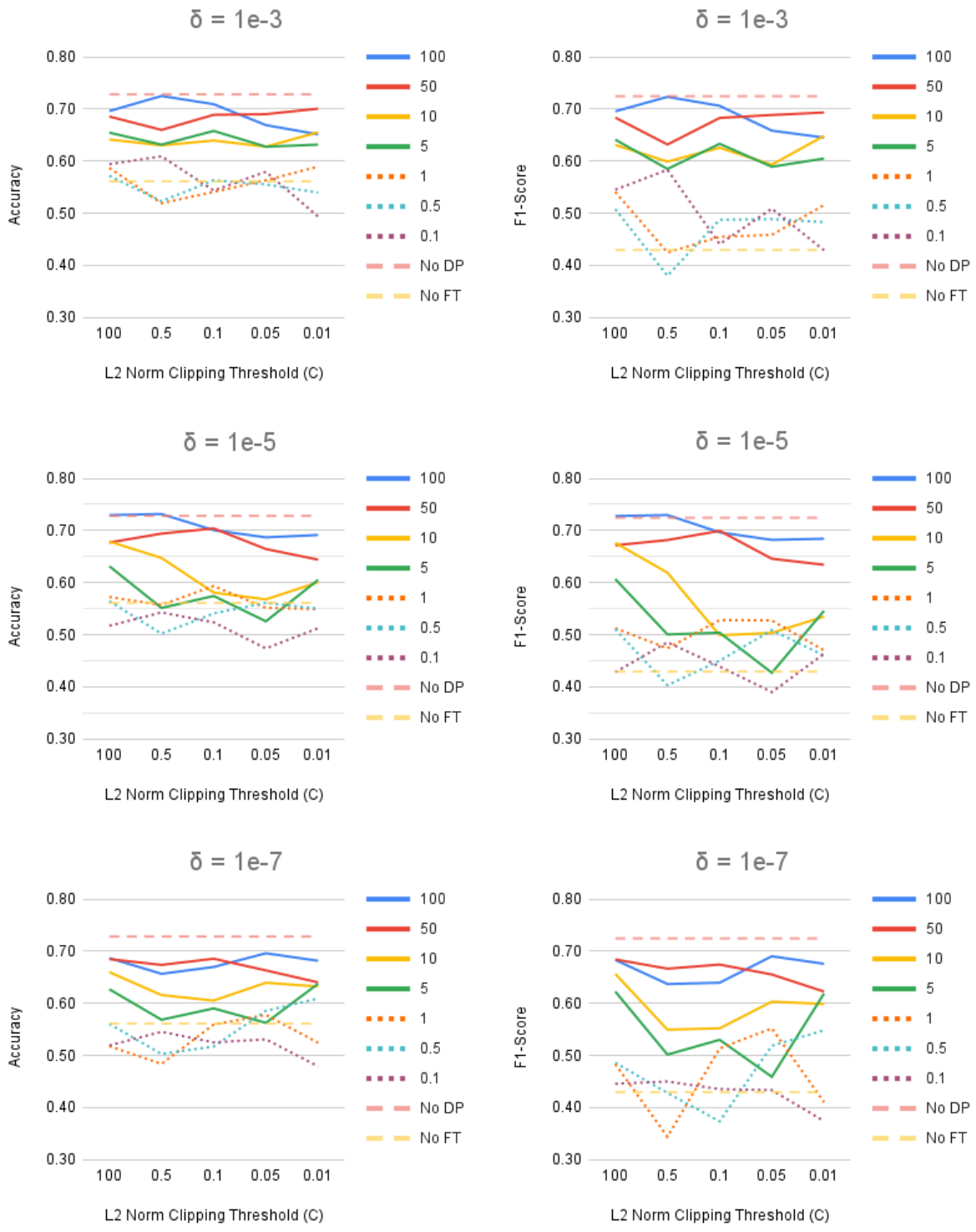

Figure 5: Accuracy (left) and F1-score (right) comparison of different values of ϵ for $\delta \in \{10^{-3}, 10^{-5}, 10^{-7}\}$.

C BERT and ALBERT comparison with different levels of privacy.

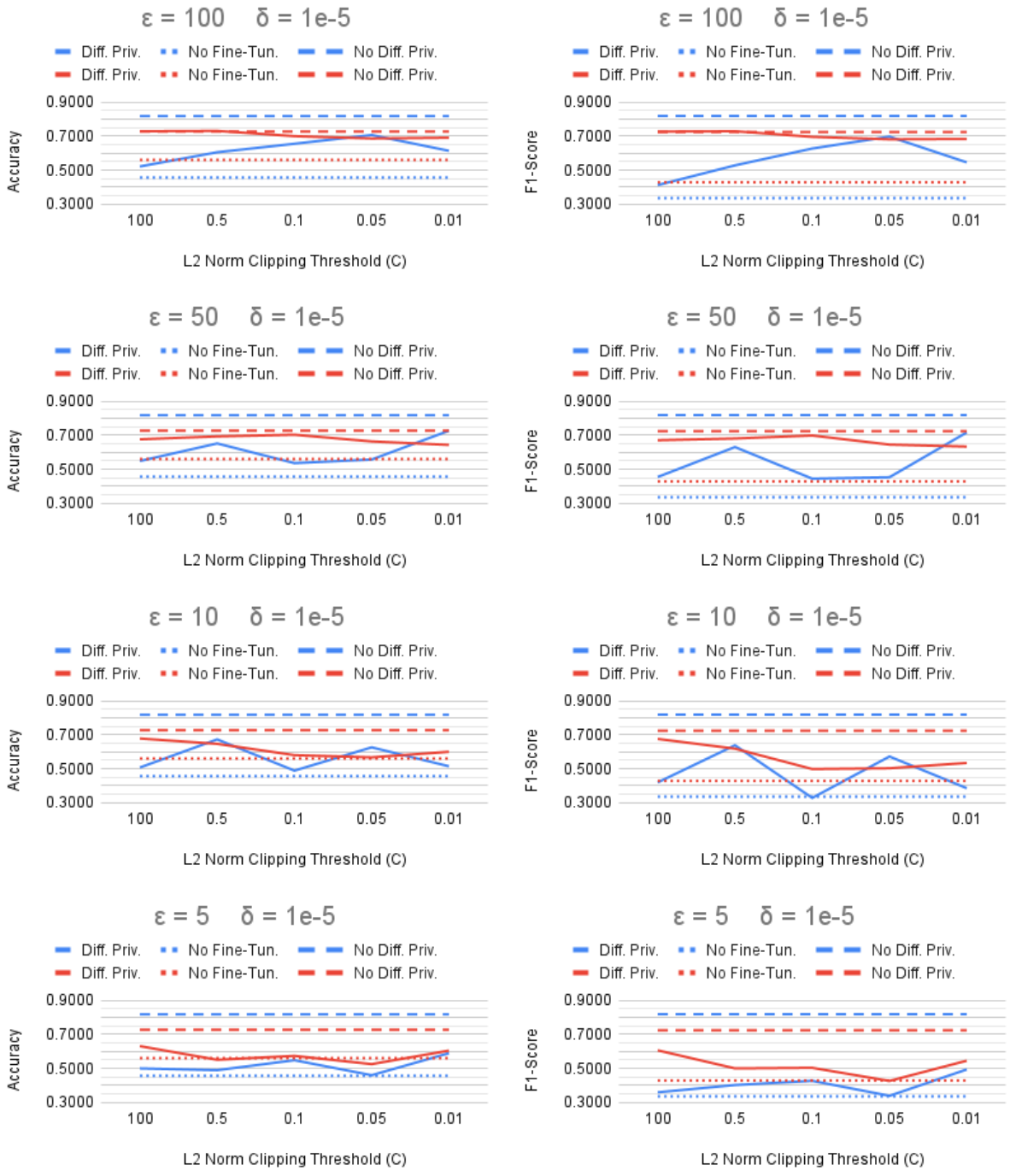

Figure 6: Accuracy (left) and F1-score (right) comparison of BERT (blue) and ALBERT (red) at the different levels of privacy ($\epsilon \in \{100, 50, 10, 5\}$, $\delta = 10^{-5}$).

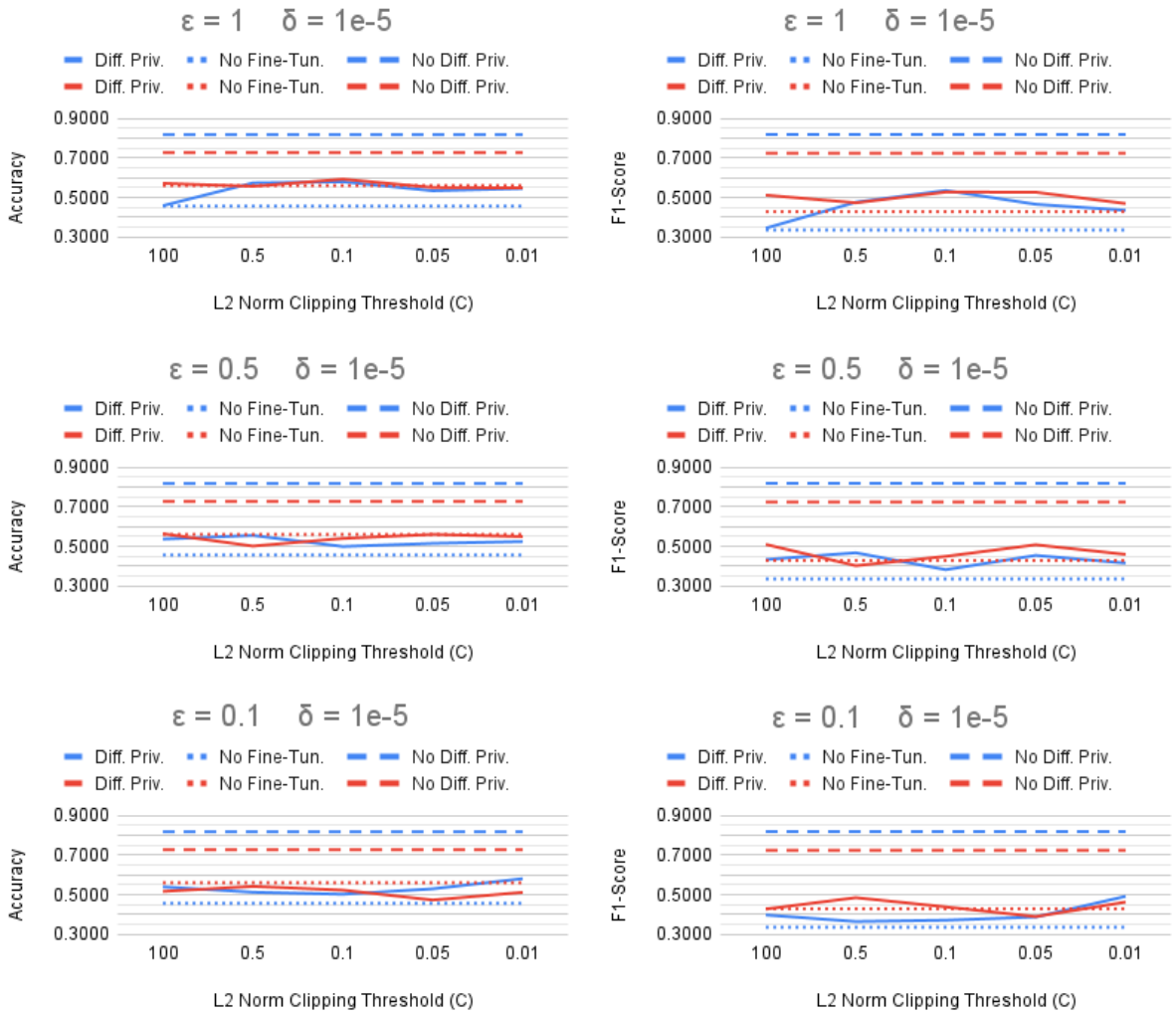

Figure 7: Accuracy (left) and F1-score (right) comparison of BERT (blue) and ALBERT (red) at the different levels of privacy ($\epsilon \in \{1, 0.5, 0.1\}$, $\delta = 10^{-5}$).

D Privacy comparison with different values of δ, for the model `ALBERT Base`.

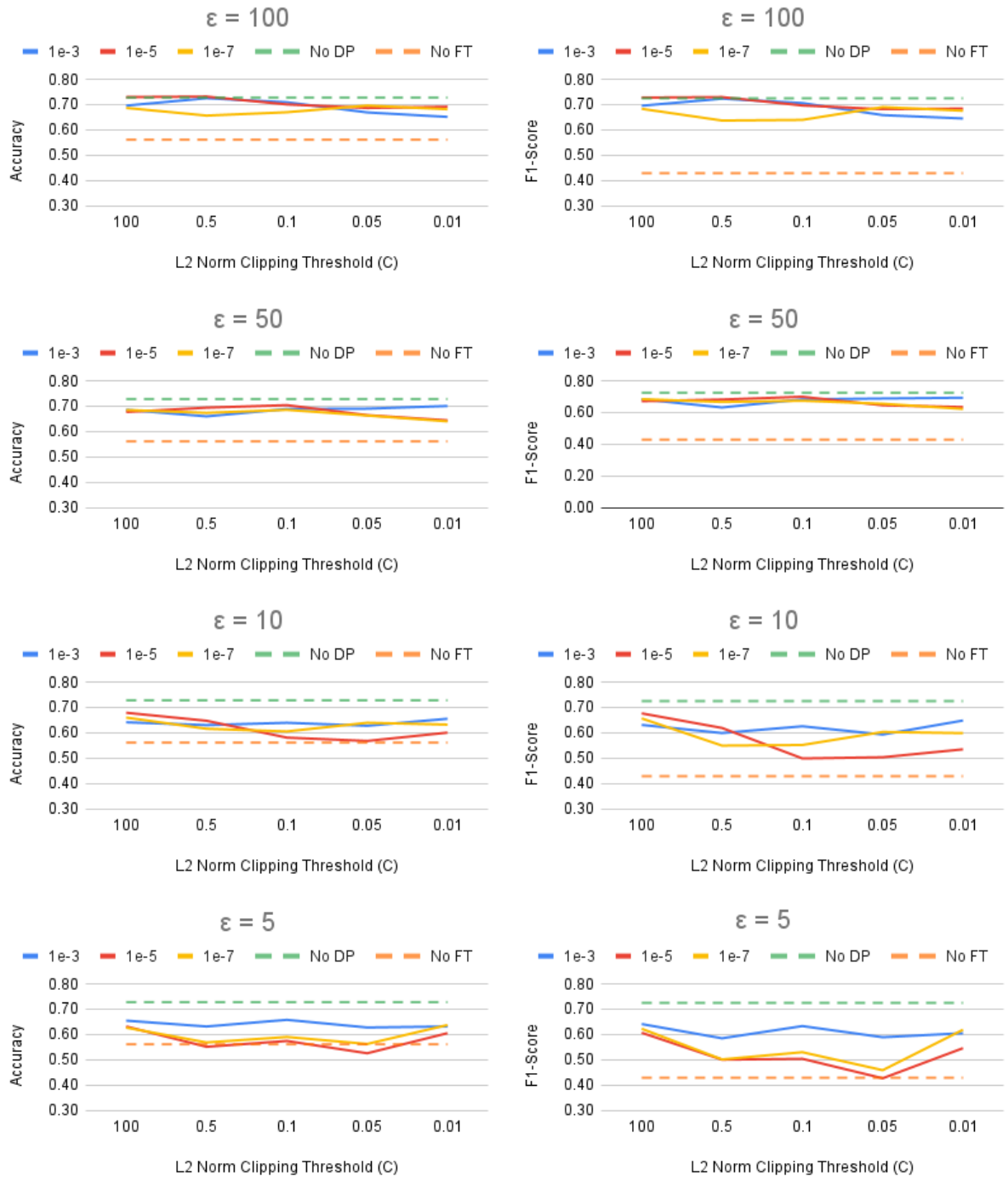

Figure 8: Accuracy (left) and F1-score (right) comparison of different values of δ for $\epsilon \in \{100, 50, 10, 5\}$.

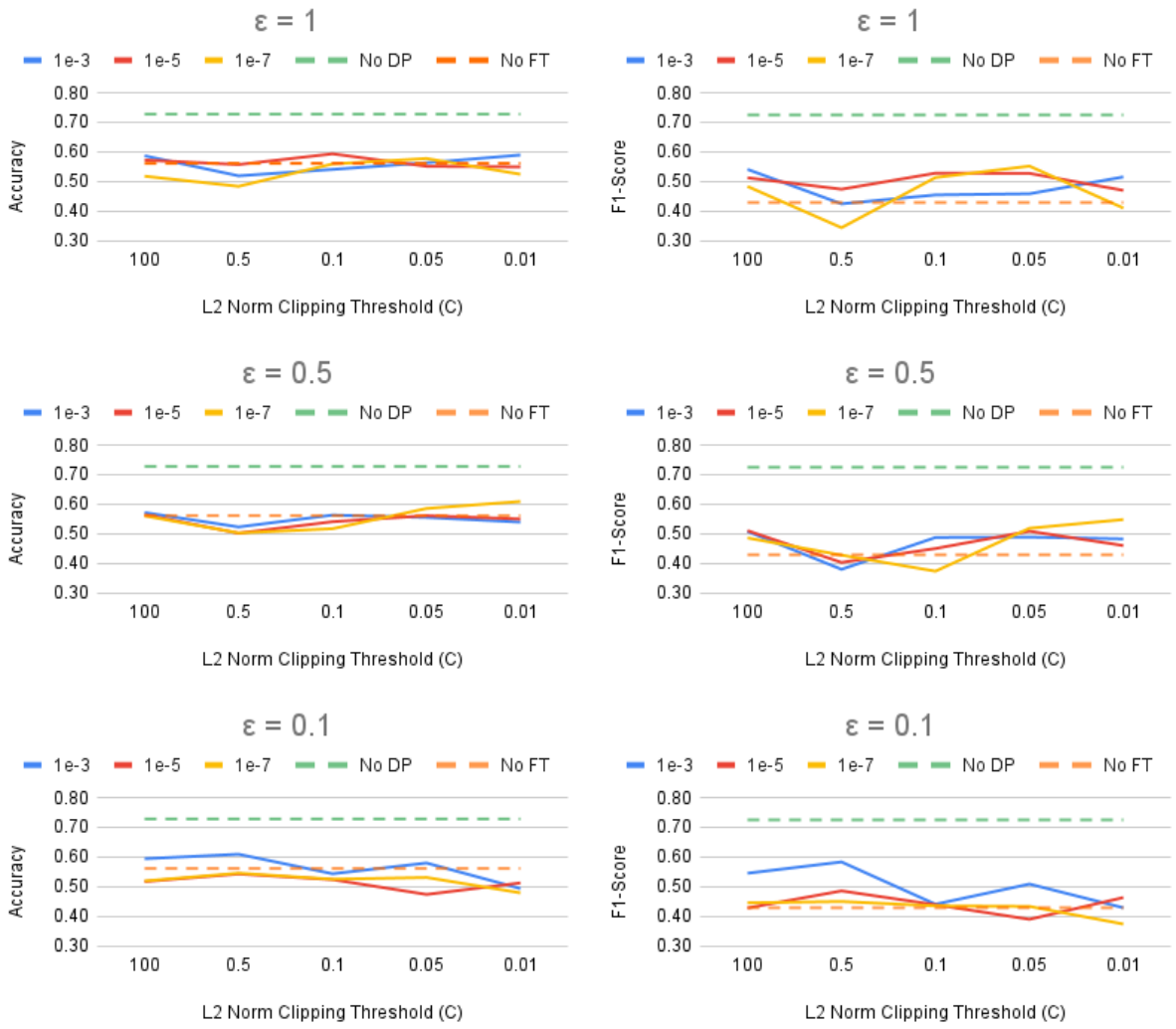

Figure 9: Accuracy (left) and F1-score (right) comparison of different values of δ for $\epsilon \in \{1, 0.5, 0.1\}$.