Journal of Advanced College of Engineering and Management

Balancing Privacy And Accuracy In Nepali Sentiment Analysis: Fine-Tuning Nepalibert With Differential Privacy

[1]Pradip Paneru, [2*]Laxmi Prasad Bhatt, [2]Anisha Pokhrel, [3]Sharad Kumar Ghimire
[1,3]Department of Electronics and Computer Engineering, Pulchowk Campus, IOE
[2]HoD, Department of Computer Engineering, Hillside College of Engineering
Email: [1]pradippaneru444@gmail.com, [2]anishapokhrel01@gmail.com, [3]skghimire@ioe.edu.np
Corresponding email: [2]lpbhatta0828@gmail.com*

DOI: 10.3126/jacem.v12i01.93900

Abstract
The increasing volume of user-generated Nepali text has enabled the development of sentiment analysis systems, but training large language models on real data introduces significant privacy risks, including potential exposure through membership inference attacks. This study examines the balance between accuracy and privacy in Nepali sentiment analysis by fine-tuning NepaliBERT with and without Differential Privacy. A high-performing non-private baseline model was trained on approximately 7,000 labeled samples, achieving near-perfect classification performance (Accuracy up to 99.88–100% and Macro F1 up to 1.00), and was subsequently evaluated for vulnerability using membership inference and canary-based privacy assessments. To mitigate privacy risks, Differentially Private Stochastic Gradient Descent was applied under varying privacy budgets (ε), and the resulting models were systematically analyzed to measure performance degradation and resistance to privacy attacks. The findings establish an empirical benchmark for the privacy–utility trade-off in low-resource Nepali NLP and provide practical guidance for building sentiment analysis systems that are both accurate and privacy-preserving.

Keywords*—Differential Privacy, DP-SGD, Low-Resource NLP, Membership Inference Attack, Nepali Sentiment Analysis, NepaliBERT, Privacy–Utility Trade-off, Privacy-Preserving Machine Learning*

1. INTRODUCTION

The rapid expansion of internet access and social media usage in Nepal has led to a substantial increase in user-generated Nepali text. This growth presents valuable opportunities for Natural Language Processing (NLP), particularly in sentiment analysis, where computational models classify opinions as positive, negative, or neutral. Transformer-based architectures such as NepaliBERT have demonstrated strong performance in low-resource language settings by leveraging contextual language representations. However, these models are typically fine-tuned on real-world textual data, which may contain sensitive personal information. As model capacity increases, so does

the risk of memorizing and unintentionally revealing private details embedded in training data. While high-accuracy sentiment models are desirable, their training process can expose users to privacy risks. Recent studies have shown that machine learning models are vulnerable to Membership Inference Attacks (MIA), where an adversary can determine whether a specific record was included in the training dataset. In the context of Nepali sentiment analysis, there has been limited investigation into whether high-performing models compromise user privacy and how privacy-preserving techniques affect model performance. The absence of empirical benchmarks for privacy–utility trade-offs in Nepali NLP creates a critical research gap.

As AI systems become increasingly integrated into digital platforms, ensuring user trust is essential. In emerging language communities like Nepal, where datasets are relatively small and often derived from public user content, the risk of memorization may be amplified. There is a pressing need to develop models that maintain strong predictive performance while offering formal privacy guarantees. Differential Privacy provides a mathematically grounded framework for limiting information leakage during training, but its impact on low-resource language models remains underexplored. This study is motivated by the need to bridge that gap.

This research aims to:

1. Develop a high-accuracy baseline Nepali sentiment analysis model using NepaliBERT.
2. Evaluate its vulnerability to privacy leakage through Membership Inference and canary-based attacks.
3. Implement Differential Privacy during fine-tuning under varying privacy budgets (ε).
4. Quantitatively analyze the trade-off between model utility (Accuracy and F1-score) and privacy protection.
5. Establish an empirical benchmark for privacy-preserving sentiment analysis in Nepali NLP.

This study contributes to both privacy-preserving machine learning and low-resource language research. It provides one of the first systematic evaluations of Differential Privacy applied to transformer-based Nepali sentiment analysis. The findings offer practical guidance for researchers and practitioners seeking to deploy AI systems that balance performance with user confidentiality. By quantifying the privacy–utility trade-off, this work supports the responsible development of NLP technologies in Nepal and similar emerging digital ecosystems.

2. RELATED WORK

Abadi, M. et. al. (2016) introduced a technique for training deep neural networks under Differential Privacy (DP) using the DP-SGD algorithm. The authors present a formal privacy framework and refined privacy accounting mechanisms that permit training models with strong privacy guarantees while maintaining competitive utility. They

implement and experimentally evaluate DP-SGD on neural networks and demonstrate the feasibility of privacy-preserving training with modest impacts on model accuracy. This paper is a core reference for understanding DP mechanisms applied to machine learning models and is widely cited in research on private model training [1].

Shokri, R. et. al. (2017). rigorously formulated Membership Inference Attacks (MIA) against machine learning models. The authors quantify how models can leak information about whether specific training records were used during training. They propose a shadow-model–based attack methodology exploiting differences between training and non-training behavior and evaluate the attack empirically across several datasets and models. The results demonstrate that standard machine learning models without proper defense are highly vulnerable to inference attacks. This work forms the theoretical and empirical basis for privacy attack evaluation in sentiment analysis models [2].

Truex et al. (2019) extend the study of membership inference by providing a detailed examination of how attack effectiveness varies with model type, dataset characteristics, and overfitting. They propose general formulations of membership inference attacks and perform comprehensive experiments showing that model vulnerability is influenced by factors such as model complexity, training data distribution, and generalization behavior. The insights from this work are useful for understanding when and why sentiment analysis models leak private information and how defensive techniques like DP can mitigate these risks [3].

Chen, J. (2021) investigates the use of differential privacy as a defense mechanism specifically against membership inference attacks. Focusing on high-dimensional and sensitive datasets, this work evaluates trade-offs between protection strength and predictive utility. It demonstrates that applying DP can reduce the success rate of MIAs while impacting classification performance in a quantifiable way, offering practical insights into how privacy budgets (ε) affect both privacy and model accuracy, a central consideration for your project's trade-off analysis [4].

Dupuy, C. et. al. (2021) proposed an optimized implementation of DP-SGD tailored for large-scale natural language models. The authors address the computational inefficiencies of standard DP-SGD and show how an efficient variant can speed up training while maintaining privacy guarantees. Evaluation results indicate competitive model accuracy with enhanced privacy protection. This work is particularly relevant to implement DP-SGD for fine-tuning transformer-based models such as NepaliBERT [5].

Vu, D. N. L. et. al. (2024) highlight a critical aspect of DP application in NLP: the granularity at which privacy is enforced. The authors compare sentence-level versus document-level DP settings in neural machine translation tasks and demonstrate that improper granularity assumptions can weaken privacy protections. They show that document-level DP provides stronger resistance to membership inference attacks. These findings underline the importance of careful DP design in text-based models and can guide how DP is structured in this research sentiment analysis task [6].

Wang, Y. et. al. (2023) work provides a broader overview of the role of DP in deep learning, including its application to mitigating membership inference attacks. The study discusses how DP-SGD achieves formal privacy guarantees and reviews its effects on robustness, generalization, and privacy leakage. It contextualizes DP mechanisms within the landscape of deep learning threats and defenses, offering theoretical backing and practical implications for projects involving sensitive text data [7].

Liu, J. et. al.(2021) critically examine the effectiveness of DP relative to other generalization techniques. It finds that machine learning strategies such as early stopping and regularization can empirically provide privacy protection against MIAs without the accuracy drop typically associated with DP. Although this work suggests alternatives to DP, it is valuable for contextualizing the privacy–utility trade-off and comparing DP with other privacy defense strategies, highlighting that DP is one among multiple possible approaches [8].

Xiangfei, Z. et. al. (2025) synthesizes current research on DP in neural networks, including defense against membership inference and other privacy attacks. It examines the impact of DP on model robustness, fairness, and generalization, and outlines future research challenges in privacy-preserving deep learning. This comprehensive review can support this resarch theoretical background and justify the significance of investigating DP in the context of NLP sentiment tasks [9].

Ahsan, S.I. et. al. (2024) applied work demonstrates privacy-preserving sentiment analysis using BERT-based architectures enhanced with federated learning and DP enhancements. The authors validate privacy through adversarial attacks, including membership inference, reporting low attack performance, and moderate utility degradation. This study exemplifies how DP and privacy evaluation are integrated into sentiment analysis tasks, directly aligning with the goals of this research [10].

3. MATERIALS AND METHODS

A. System Overview

The system evaluates the trade-off between model accuracy and privacy in Nepali sentiment analysis using a dual-pipeline architecture. The process begins with a labeled Nepali sentiment dataset, which undergoes preprocessing, including cleaning, tokenization, encoding, and dataset splitting into training, validation, and test sets.

Two models are then trained in parallel using the same preprocessed data. The first is a baseline model obtained by standard fine-tuning of NepaliBERT without privacy constraints, serving as a high-accuracy reference. The second model incorporates Differential Privacy through the DP-SGD optimization mechanism, where gradient clipping and calibrated noise addition are applied during training under different privacy budgets (ε).

Both models are evaluated using classification metrics such as accuracy and F1-score. To assess privacy risks, membership inference attacks are conducted to measure the extent of information leakage. Finally, the results from performance and privacy evaluations are compared in a trade-off analysis to quantify how stronger privacy guarantees affect predictive performance. This framework enables systematic assessment of building sentiment analysis models that are both accurate and privacy-preserving.

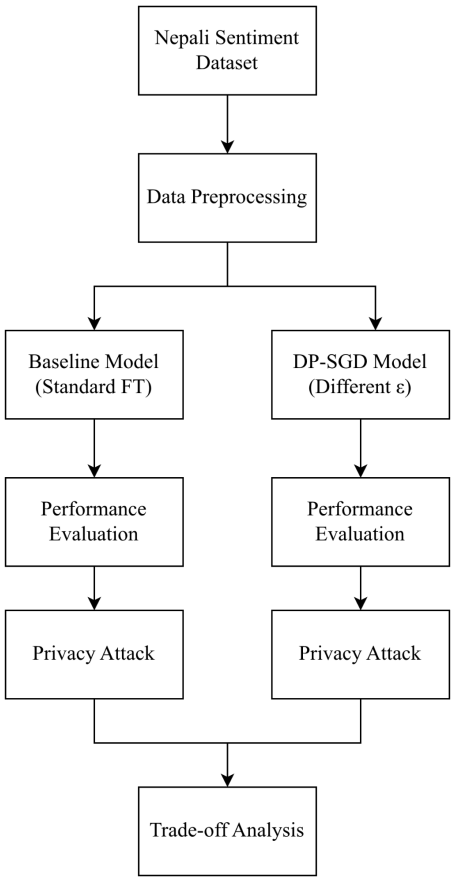

Figure 1: System block diagram

B. Datasets

The study utilizes a custom-compiled Nepali sentiment dataset consisting of approximately 7,000 manually curated text samples collected from publicly available Nepali-language digital platforms, including social media posts, online discussions, and web-based content. Each sample is labeled into one of three sentiment categories: positive, negative, or neutral. For the privacy-focused experiment, the primary corpus

was strategically extended with a set of 100 artificially constructed "canary" samples containing unique, non-natural token sequences designed to simulate rare or sensitive data points. These canary instances were intentionally created to evaluate potential memorization behavior and privacy leakage during training. The dataset was partitioned into training, validation, and test subsets using an 80%–10%–10% split for the baseline experiment, while in the canary-based setup, the test set was drawn exclusively from the main corpus to ensure unbiased generalization assessment, and selected canary samples were injected only into the training data to analyze exposure risk under controlled conditions.

C. Dataset Preprocessing

A structured and reproducible preprocessing pipeline was implemented to ensure data quality, consistency, and compatibility with the transformer-based architecture. The raw Nepali text samples were first subjected to data cleaning procedures, which included the removal of duplicate entries, correction of encoding inconsistencies, elimination of extraneous whitespace, and filtering of irrelevant characters such as unsupported symbols or malformed text fragments. Care was taken to preserve meaningful punctuation and linguistic markers that may contribute to sentiment expression. All text was normalized to a consistent Unicode format to avoid tokenization errors during model training.

Following cleaning, label verification was performed to ensure that each instance was correctly assigned to one of the three sentiment categories: positive, negative, or neutral. Any ambiguous or incorrectly labeled samples identified during manual inspection were corrected or removed to maintain dataset reliability. Class distribution was examined to confirm a reasonable balance across sentiment categories, reducing potential bias during training.

The refined textual data was then processed using the tokenizer associated with the pretrained NepaliBERT model. Each sentence was transformed into subword tokens according to the model's vocabulary. Special tokens required by the transformer architecture, including classification and separator tokens, were appended automatically. The tokenized sequences were converted into numerical input IDs, and corresponding attention masks were generated to distinguish valid tokens from padding. To enable efficient batch processing, all sequences were padded or truncated to a predefined maximum sequence length, ensuring uniform tensor dimensions without altering semantic content beyond the fixed limit.

Finally, the processed dataset was partitioned according to the experimental design. For the baseline experiment, the data were divided into training (80%), validation (10%), and test (10%) subsets using stratified sampling to preserve class proportions across splits. In the privacy-focused experiment, the main corpus and designated canary samples were carefully allocated to maintain strict separation between evaluation and

training data. This structured preprocessing framework ensured data integrity, reproducibility, and compatibility with both standard and differentially private training procedures.

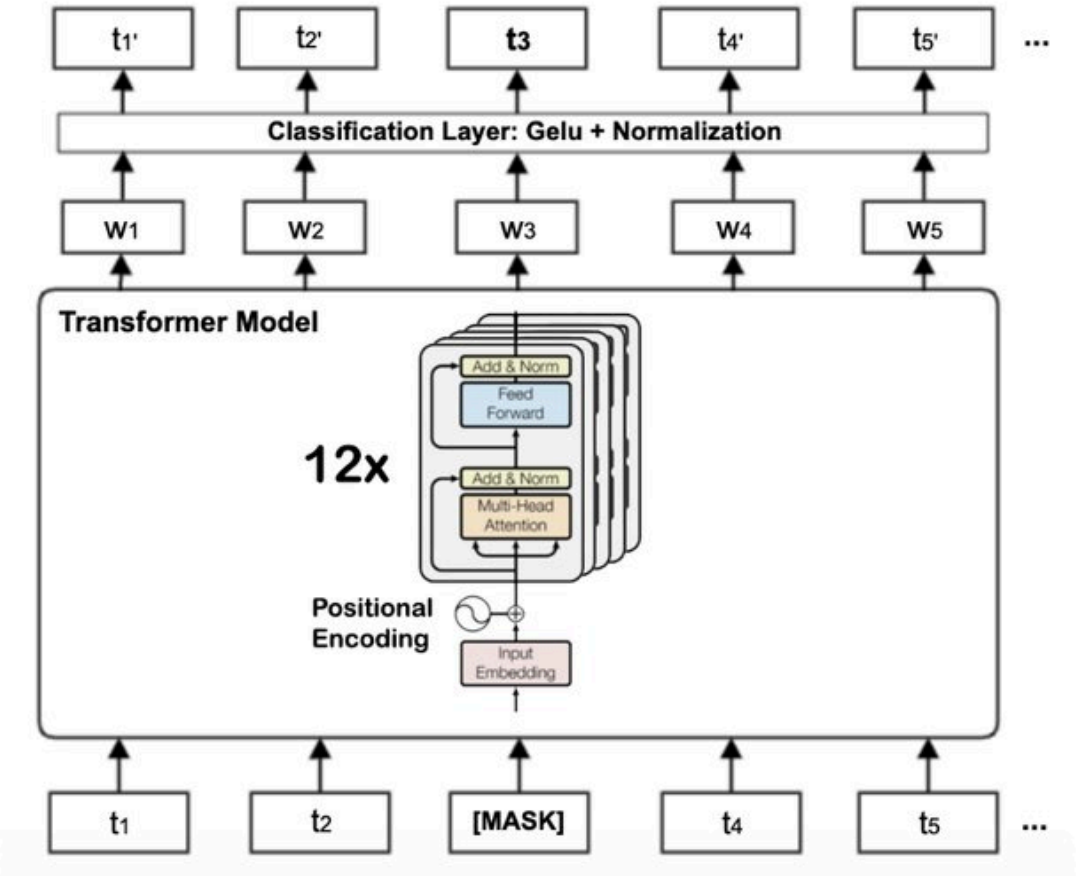

Figure 2: Transformer-Based Architecture for Nepali Sentiment Classification [11]

The illustrated architecture represents a BERT-based Transformer encoder model used for sentiment classification. The input sentence is first tokenized into individual tokens (t1, t2, t3, …), including special tokens such as [MASK]. Each token is converted into a dense vector through an embedding layer, and positional encoding is added to preserve word-order information. The embedded sequence is then passed through a stack of twelve Transformer encoder layers. Each encoder layer consists of a multi-head self-attention mechanism followed by a position-wise feed-forward network. Residual connections and layer normalization are applied after each sub-layer to stabilize training and improve gradient flow. Through this deep bidirectional attention mechanism, the model learns contextual relationships between words in the sentence. The contextualized token representations produced by the final encoder layer are fed into a classification head composed of a fully connected layer with GELU activation and normalization. The output layer maps the learned representations to three sentiment classes, producing probability scores for each category. This architecture enables the

model to capture complex semantic and syntactic patterns in Nepali text while maintaining strong generalization capability.

D. System Flowchart

The overall experimental framework begins with dataset collection, followed by systematic data cleaning and validation to ensure quality and consistency. The validated dataset is then preprocessed through tokenization, formatting, and encoding suitable for transformer-based modeling. The data is partitioned into training (80%), validation (10%), and test (10%) subsets to enable reliable model development and unbiased evaluation.

A pre-trained NepaliBERT model is loaded as the foundational architecture. Its final classification layer is re-initialized to accommodate a three-class sentiment classification task. At this stage, the workflow branches into two experimental settings. In the first setting, the model is fine-tuned on the main corpus using standard training procedures. In the second setting, fifty synthetic canary samples are injected into the training set to evaluate memorization behavior under controlled exposure. For both settings, performance metrics such as Accuracy and F1-score are computed, followed by the execution of a Membership Inference Attack (MIA) to assess privacy vulnerability. The results are recorded for comparative analysis.

Subsequently, the framework transitions to the privacy-preserving phase. The Opacus PrivacyEngine is integrated into the PyTorch training loop to enable Differentially Private Stochastic Gradient Descent (DP-SGD). Multiple models are trained under varying privacy budgets (ε), incorporating gradient clipping and calibrated noise addition. Each differentially private model is evaluated in terms of both predictive performance and resistance to membership inference attacks. All results are compiled and compared across baseline, canary, and differentially private configurations. Finally, a comprehensive privacy–utility trade-off analysis is conducted to quantify the cost of enforcing stronger privacy guarantees.

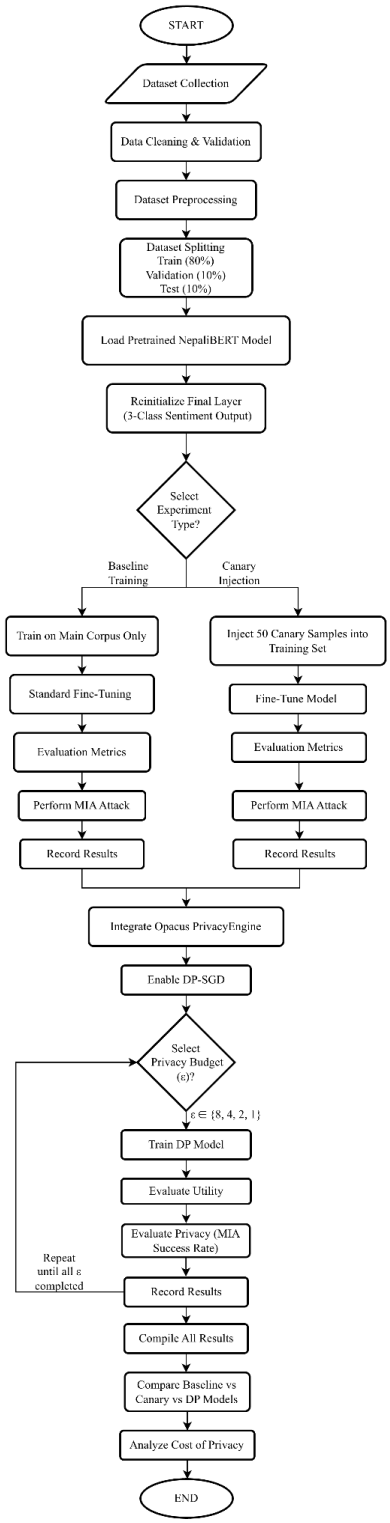

Figure 3: Flowchart of System

4. RESULTS AND DISCUSSIONS

A. Baseline Results

The baseline model demonstrated extremely strong predictive performance on the held-out test set. It achieved an overall accuracy of 99.88%, indicating that the model correctly classified nearly all test samples. In addition, the macro-averaged F1-score reached 0.9984, showing that the model maintained a balanced and highly consistent performance across all three sentiment classes. The close alignment between accuracy and macro F1-score suggests that the classifier performed uniformly well without favoring any specific category. These results confirm that the fine-tuned Nepali-BERT model is highly effective for sentiment classification under standard, non-private training conditions.

The privacy evaluation using a Membership Inference Attack (MIA) indicates that the baseline model does not exhibit distinguishable confidence patterns between training and non-training samples. The average prediction confidence for member samples was 0.9998, while the average confidence for non-member samples was also 0.9998, showing virtually no measurable difference between the two groups. Furthermore, the MIA success rate, measured using the Area Under the ROC Curve (AUC), was 0.4984. Since an AUC value close to 0.5 corresponds to random guessing, this result suggests that the attack was unable to reliably infer whether a given sample belonged to the training set. Therefore, under this evaluation setting, the model does not demonstrate significant vulnerability to membership inference.

B. Memorization Experiment (Canary)

The definitive baseline model continued to demonstrate exceptionally strong classification performance. On the test set, it achieved an accuracy of 99.75% along with a macro-averaged F1-score of 0.9974, indicating consistently high predictive quality across all sentiment classes. Despite this strong utility performance, the privacy evaluation under the canary-based membership inference setting revealed a distinct behavior. The average prediction confidence for injected canary samples was 0.6184, whereas the confidence for non-member test samples remained substantially higher at 0.9984. The resulting MIA AUC score was 0.0025, which is significantly below the random-guess threshold of 0.5. This outcome suggests that the model did not memorize the injected canary samples in a manner that enabled successful membership inference under this experimental configuration.

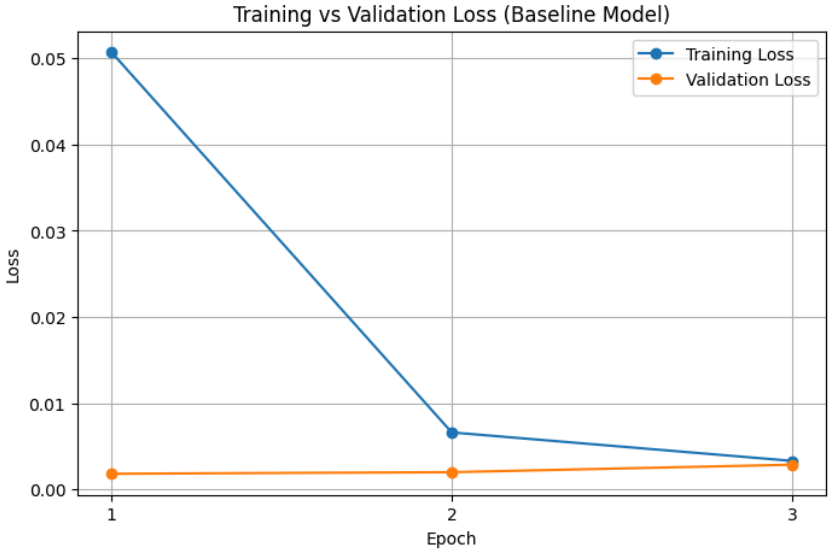

Figure 4: Training vs Validation Loss (Baseline Model)

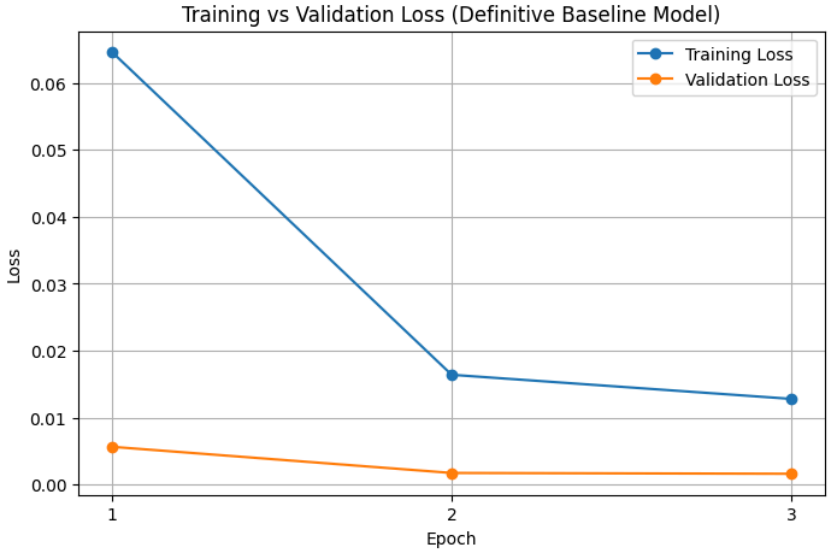

Figure 5: Training vs Validation Loss (Definitive Baseline Model)

C. Model Performance

Table 1: Result Summary of Baseline Model and Definitive Baseline Model

Category	Metric	Baseline Model	Definitive Baseline Model
Training (Final Epoch)	Training Loss	0.0033	0.0128
	Validation Loss	0.0029	0.0016
	Validation Accuracy	0.9988	0.9986
	Validation F1	0.9989	0.9984
Test Performance	Test Accuracy	**0.9988**	**0.9975**
	Macro F1-Score	**0.9984**	**0.9974**
Membership Inference Attack (MIA)	Avg Confidence (Members)	0.9998	0.6184 *(Canaries)*
	Avg Confidence (Non-Members)	0.9998	0.9984
	MIA AUC Score	**0.4984**	**0.0025**

From the above Table 1, comparison between the Baseline Model and the Definitive Baseline Model shows that both systems achieve exceptionally high classification performance, with validation and test accuracy consistently above 99%. The Baseline Model concludes training with a lower training loss (0.0033) but a slightly higher validation loss (0.0029), while the Definitive Baseline records a marginally higher training loss (0.0128) and lower validation loss (0.0016), indicating stable and well-generalized learning in both cases. Validation accuracy and F1-scores remain nearly identical across models, confirming that predictive performance is not meaningfully compromised in either setup. On the test set, performance remains extremely strong, with accuracy values of 0.9988 and 0.9975, respectively. The privacy analysis, however, reveals a notable difference in attack behavior. In the standard Baseline Model, the Membership Inference Attack yields an AUC of 0.4984, which is effectively equivalent to random guessing, suggesting no distinguishable separation between members and non-members based on confidence scores. In contrast, the Definitive Baseline Model under the canary-based evaluation produces a very low AUC

of 0.0025, reflecting a strong inversion in confidence behavior where the injected canary samples are treated differently from regular test data. Overall, while both models demonstrate nearly identical predictive strength, their privacy evaluation outcomes differ substantially depending on the experimental design used for membership analysis.

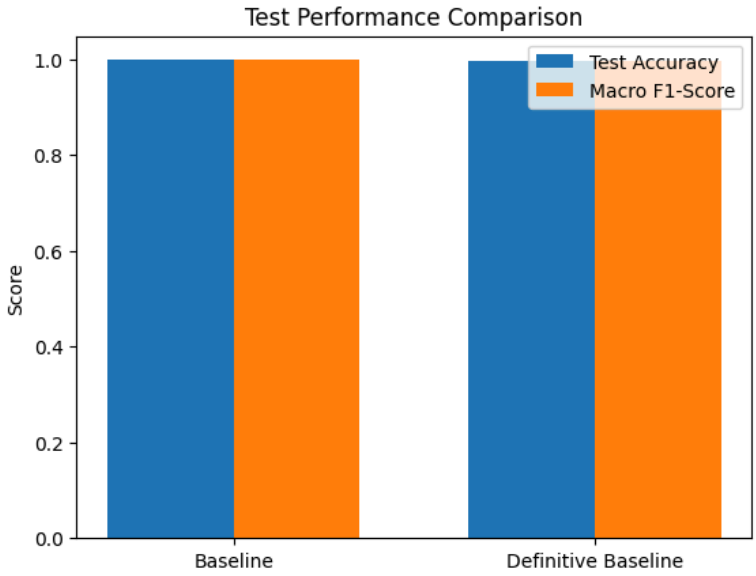

Figure 6: Test Performance Comparison

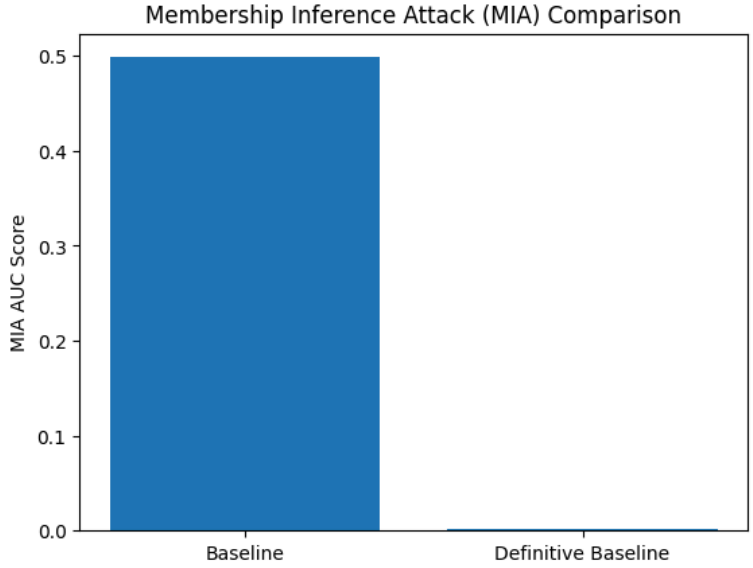

Figure 7: MIA Comparison

5. CONCLUSION

The overall system demonstrates a clear relationship between model utility and privacy risk. The standard baseline model achieved near-perfect classification performance, with accuracy and macro F1-scores close to 99.9%, indicating that the sentiment classifier learned the training distribution extremely well. At the same time, the membership inference results show that the model's confidence behavior between training and unseen samples was nearly identical, and the attack AUC remained around 0.5. This suggests that, under the evaluated setting, the baseline model does not expose easily exploitable membership signals through simple confidence-based attacks, despite its very high predictive performance.

The definitive baseline experiment, which incorporated canary samples for memorization analysis, further confirmed that the model maintained strong predictive capability while not showing meaningful vulnerability to the specific membership inference setup used. The extremely low AUC observed in the canary attack indicates that the attack was not able to distinguish memorized samples from non-members in a statistically reliable manner under the chosen metric.

In contrast, the differentially private (DP) model introduced a strict privacy guarantee, with a final privacy budget of approximately $\varepsilon \approx 0.40$. However, this strong privacy protection came at a substantial cost to utility, as reflected by the significant drop in accuracy and macro F1-score. While the MIA AUC remained close to random guessing (around 0.5), the reduced predictive performance highlights the practical trade-off between maintaining model effectiveness and enforcing formal privacy constraints.

Taken together, the results illustrate a fundamental balance: high-performance models can be trained without obvious membership leakage under simple attack strategies, but introducing rigorous differential privacy significantly reduces predictive accuracy. Therefore, selecting an appropriate privacy mechanism requires careful consideration of the intended application, acceptable performance degradation, and required privacy guarantees.

6. FUTURE WORK

Future work may focus on strengthening both the privacy evaluation framework and the utility–privacy balance of the proposed system. More advanced membership inference techniques, including shadow-model and loss-based attacks, can be incorporated to provide a deeper and more rigorous assessment of privacy leakage. Optimization of differential privacy hyperparameters such as noise scaling, gradient clipping strategies, and selective layer training may improve model performance while maintaining strong privacy guarantees. Expanding the evaluation to larger and more diverse Nepali language datasets, as well as additional NLP tasks, would enhance the generalizability of the findings. Exploration of parameter-efficient fine-tuning methods and alternative transformer

architectures may also identify configurations that offer improved robustness under private training.

ACKNOWLEDGMENTS

We are sincerely thankful to the Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, for encouraging us to do this work as a part of our minor project and providing us with the necessary resources.

We express our heartfelt gratitude to Pulchowk Campus for valuable guidance, feedback, and inspiration, which played a key role in the successful completion of this work.

Lastly, we thank our colleagues and families for their moral support and motivation during this academic endeavor.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS'16, page 308–318. ACM, October 2016.

[2] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models, 2017.

[3] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Towards demystifying membership inference attacks, 2019.

[4] Chen, Junjie et al. "Differential Privacy Protection Against Membership Inference Attack on Machine Learning for Genomic Data." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* vol. 26 (2021): 26-37.

[5] Christophe Dupuy, Radhika Arava, Rahul Gupta, and Anna Rumshisky. An efficient DP-SGD mechanism for large-scale NLP models, 2022.

[6] Doan Nam Long Vu, Timour Igamberdiev, and Ivan Habernal. Granularity is crucial when applying differential privacy to text: An investigation for neural machine translation, 2024.

[7] Yanling Wang, Qian Wang, Lingchen Zhao, and Cong Wang. Differential privacy in deep learning: Privacy and beyond. Future Generation Computer Systems, 148:408–424, 2023.

[8] Jiaxiang Liu, Simon Oya, and Florian Kerschbaum. Generalization techniques empirically outperform differential privacy against membership inference, 2021.

[9] Xiangfei, Z., Qingchen, Z. Defending against attacks in deep learning with differential privacy: a survey. *Artif Intell Rev* 58, 347 (2025). https://doi.org/10.1007/s10462-025-11350-3.

[10] Ahsan, S.I.; Djenouri, D.; Haider, R. Privacy-Enhanced Sentiment Analysis in Mental Health: Federated Learning with Data Obfuscation and Bidirectional Encoder Representations from Transformers. Electronics 2024, 13, 4650.

[11] Usama Khalid, Mirza Omer Beg, and Muhammad Umair Arshad. Rubert: A bilingual Roman Urdu Bert using cross-lingual transfer learning, 2021.