

Ethical AI for Unlearning the Biased and Sensitive Information

Arslan Ali Raza

arslan@cuivehari.edu.pk

Department of Computer Sciences, COMSATS University Islamabad, Vehari Campus

Ghulam Fatima

Department of Computer Sciences, COMSATS University Islamabad, Vehari Campus

Rubab Sohail

Department of Computer Sciences, COMSATS University Islamabad, Vehari Campus

Corresponding Author: Arslan Ali Raza arslan@cuivehari.edu.pk

Received: 08-02-2026

Revised: 23-02-2026

Accepted: 09-03-2026

Published: 24-03-2026

ABSTRACT

The rapid advancement of artificial intelligence systems across sensitive domains has elevated serious concerns about the retention of biased, private, and ethically problematic information within trained models. This research presents an ethical AI framework for unlearning biased and sensitive information from machine learning models. An initial BERT model was trained to achieve a good reference performance in the classification task. Gradient ascent-based unlearning was used to unlearn a certain amount of targeted information to eliminate it. Findings revealed that unlearning in itself resulted in the apparent reduction of accuracy and the inability to separate classes. In order to regain model stability, retraining was done with fine-tuning on the retained dataset. The overall performance of the fine-tuning was the best, as it enhanced precision and minimized the prediction errors. An intermediate measure to harmonize forgetting and performance recovery was also tested as hybrid method of unlearning and retraining. The effectiveness of retraining-based recovery was proved by comparative analysis with accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices. The results show that to perform ethical machine unlearning, effective forgetting and performance restoration are needed. This piece of work facilitates building reliable AI systems in accordance with the requirements of fairness and privacy.

Keywords: Big Data, Deep Learning, Federated Learning, LLM Behaviour, Transformers, Unlearning

INTRODUCTION

Artificial Intelligence (AI) has changed the reality of contemporary society fast through giving machines capabilities to carry out the tasks that used to be executed by human intelligence. Artificial intelligence is currently applied in the most important fields, including the diagnosis treatment industry, finance risk management, automated hiring, police surveillance, learning platforms, and social media recommendation systems. Such AI systems usually affect physical choices that directly affect people and populations. Due to this increased power, it has become the responsibility of all countries to make sure that AI models are responsible, fair, and safe. This has led to an idea of Ethical AI, which aims at developing and implementing AI systems that uphold human values, maintain user privacy, minimize discrimination, and remain transparent and accountable. Nonetheless, bias and sensitive data within machine learning models is one of the greatest ethical AI development challenges. The AI models acquire patterns based on the data involved in the training process. When there are biased behaviors, stereotypes, or historical inequalities in the training data, the AI model will practice them unconsciously [1]. This is generally referred to as algorithmic bias. An illustration is that a biased AI-based recruitment system might favor one gender over the other or a loan approval model might not consider some socio-economic groups because it has been observed in the previous records. These consequences undermine the confidence in AI systems and may result in severe detriments as they support inequality and

unfairness in the decision-making process in the real world. Besides bias, AI machines may also have sensitive or personal information carried and stored in training data [2]. Such sensitive information can be personal identifiers, medical history, financial data, personal or private conversations, and demographic information such as religion and ethnicity. In other instances, even when large-scale AI models are asked to reproduce certain sections of their training data, they can do so, and do so perfectly. This poses significant threats of privacy breach, unauthorized access, and misconduct of confidential information [3]. The matter is even more severe when the models of AI are trained on a vast amount of data gathered through the Internet, organizations or a user, and where sensitive data might be hidden. With the increase in the power and the wide adoption of AI models, the capacity to identify and eliminate such information is gaining significance to secure people and adhere to both ethics and legal standards.

In order to counter such challenges, scholars have unveiled a potentially effective solution called Machine Unlearning. Machine unlearning is a concept that allows the AI models to forget particular data, patterns or knowledge which is no longer relevant to the model prediction or output. As opposed to the classical deletion algorithms in which the data is only deleted in storage, unlearning is meant to eliminate the impact of the data on the trained model. It is particularly convenient when the model is already deployed and retraining it at once is costly, time-consuming or unfeasible. Machine unlearning can create the possibility to remove some of the most detrimental bias, some more sensitive information, or even satisfy user demands to have their data removed without retraining the entire AI system [4]. Machine unlearning also closely relates to contemporary privacy policies, particularly the existence of laws like the General Data Protection Regulation (GDPR), which propose a concept of the Right to be Forgotten. The right gives people the right to request that their personal information be removed out of systems. But with the case of the AI models, the problem is even more complicated due to the fact that the personal data might be injected into the learned parameters of the model during the training. It is not automatically true that by merely removing the data in the dataset the learner is also removed by the model. Thus, ethical AI should be provided with superior solutions, including unlearning, to make sure that the AI models do not keep relying on the information that is unlawful, harmful, sensitive, or unjust.

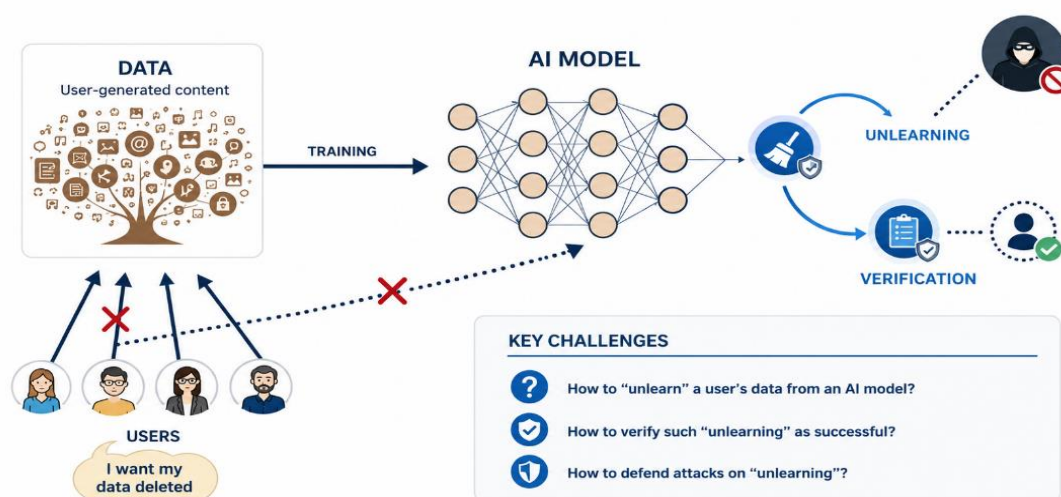


Figure1: Machine unlearning workflow illustrating the removal of a user's sensitive data from a trained AI model, along with unlearning verification and protection against potential attacks

With the growing significance of machine unlearning, we strive to conduct comprehensive research on the current landscape in this field, as illustrated in the following figure 1. Our goal is to offer a systematic perspective on the ongoing advancements in this area and to highlight its critical role in the

constantly evolving machine learning ecosystem. Machine unlearning presents multiple research and operational issues although it has potential. Any removal of biased or sensitive information has to be done with care not to affect the accuracy, stability and reliability of the model. In case the unlearning is not conducted effectively, the model can lose valuable general knowledge or can give unreliable outcomes [5]. Also, there needs to be transparency, verifiability, and invulnerability to attacks regarding the processes of unlearning. One of the questions still persists: in what way can we be sure that the AI model has really lost the information that was desired, and that the knowledge that was deleted cannot be reawakened in some other way? This brings up the problems of model integrity, fairness testing, privacy leakage testing and robust verification. Hence, the direction of Ethical AI of Unlearning Biased and Sensitive Information has turned out to be a critical and important research field in responsible machine learning. To begin, we present a detailed literature review on the unlearning of sensitive information in Section 2, which summarizes the findings and gaps in the related ongoing studies. The methodology and work flow used in this study is described in Section 3. Section 4 presents the results and discusses their implications. The study ends in the recommendations for future research and practice in Section 5.

LITERATURE REVIEW

To unravel the literature and related research in the field 'Ethical AI for Unlearning Biased and Sensitive Information', more focus was needed on developing techniques that can detect and remove undesirable information from trained machine learning models without compromising their performance and reliability. The fast implementation of AI systems to conduct sensitive processes like hiring, healthcare, credit scoring, and surveillance have presented a serious question of fairness, privacy and accountability. It is therefore important to note that the AI models are not just learning to become good predictors, but they can also internalize (learning) biased correlations and sensitive information from the training data, which will then appear in the output and decisions. This has spurred the research community to think about what to do about this problem; adding solutions, which can forget special data points, user history, or malicious trends without necessarily re-training anew [6]. One of the key ethical issues covered by a significant part of the available literature is the issue of algorithmic bias. When some groups of people are overrepresented or underrepresented, prejudice is typically based on unequal backgrounds and data sets from the past. There are many studies that emphasize that the training data used to develop AI systems can result in demographic-specific inaccuracies, such as poorer accuracy in minority groups, or discriminatory favoritism to the majority groups. The many bias reduction strategies have been suggested by researchers, such as pre-processing (cleanup or balancing data), in-processing (changing model training goals), and post-processing (altering post-training predictions) [7]. However, these solutions are incomplete and might be inadequate when discriminatory or sensitive patterns are strongly represented in the parameters of the model.

Meanwhile, there is one more important research direction: the problem of privacy leakage and sensitive data exposure in machine learning models. A number of studies also show that models, in particular, deep neural networks and large language models, can memorize training samples unintentionally. Such memorization can enable attackers to steal confidential data in either model inversion or membership inference attacks or querying [8]. These dangers are even greater in case the models are trained on real-world data sets with personal identifiers, with medical records, with financial transactions, or with personal user-generated content. In its turn, this is why researchers believe that privacy should not be limited solely to the practices of dataset anonymization, since even anonymous data can be re-identified in certain cases due to indirect correlations. In order to address privacy-related and bias-related issues, the research community has proposed the notion of machine unlearning, the goal of which is to eliminate the impact of a particular data point on a trained model [9] [10]. Machine unlearning can be said to be the reverse of learning, in which the goal is not to enhance knowledge, but to remove specific information. The initial investigations into the field of unlearning looked at retraining-based approaches, in which the model is retrained using a blank slate following the erasure of the undesired data. Though effective, full retraining is computationally costly and not practically feasible to large-scale AI systems.

This constraint compelled scholars to come up with better unlearning algorithms that do not need full retraining but still permit that the deleted data will not influence model decisions.

However, according to the recent literature, such categories of unlearning techniques have been suggested as exact unlearning, approximate unlearning, and certified removal techniques. Exact unlearning seeks to attain the same model behavior as that would have been attained with the deleted data in training never being present. Genuine unlearning is however only often practical on simpler models or particular architectures [11]. Rather, approximate unlearning is interested in fast unlearning with little computation but not necessarily with optimal forgetting. Certified removal methods strive to establish mathematical or even experimental evidence that the unlearning has been put into practice. When there's a high stakes system, these are important methods where unlearning must be measurable, transparent, and defensible, and this is precisely why they are so important. The second one which is related to the other comes from the use of ethical and legal considerations, specifically the Right to be Forgotten, which has been put forward in privacy laws like GDPR. Compliance to such rules needs not only to strip away raw user data but learned body of that data too, scholars note, as part of the model itself. This has inspired the development of unlearned pipelines, which combine data governance policies and technical elimination plans [12]. Moreover, AI scientists working on ethics make it clear that unlearning needs to be not privacy-centered only but also aims at assisting in the elimination of discriminatory trends, biased characteristics, and destructive stereotypes acquired during training. Moreover, unlearning of verification and robustness are also pointed out in the literature. Verification means ascertaining that a model has indeed forgotten the information requested out of it [13] [14] [15]. A few of the proposed methods that researchers suggest could be part of the evaluation process are: comparison of model behavior before and after unlearning, parameter changes, evaluation without membership inference, and audit of privacy leakage [16]. While others caution that unlearning techniques could be the targets of adversarial manipulations and attacks in which the adversary attempts to reverse or introduce errors in the forgetting system. Therefore, the unlearning should be efficient with strong algorithms, and should also be verified with strong verification procedures and be protected with strong precautions.

Recent research places unlearning as a tool for governance to rectify AI systems that have learned something that is harmful [17]. The arguments currently circulating are not just on the need to unlearn, but also on how to measure its effectiveness by deletion success, retained utility, scalability, and auditability and not just accuracy by itself. Unlearning is being looked at more in sensitive AI systems, as a post-deployment measure to eliminate memorized entities, unsafe associations, and unwanted behavioral patterns [18][19]. Unlearning has been introduced as a realistic approach for aligning harmful responses, copyrighted content, and knowledge related to hallucination suppression for large language models. Recent LLM research has demonstrated that unlearning may be more computationally efficient than complete retraining, however there remain challenges with selecting data and tuning hyperparameters to ensure unlearning is successful [20]. The risk of over-forgetting is another aspect of Knowledge un-learning research, which indicates that eradicating sensitive knowledge can also lead to loss of unrelated useful knowledge. Fair machine unlearning studies caution that group-level fairness may be inadvertently affected by the deletion process and that the bias needs to be assessed prior and after the unlearning process [21]. In addition, adversarial studies argue for the importance of testing unlearned models under the adversarial setting, as forgetting can alter boundaries. Finally, robustness-based analysis has shown that adversarial testing is also important for unlearned models due to the potential for forgetting to change boundaries. Verification research shows that while model providers seem to obey the requests to unlearn, they still have information influence that is hidden. Table 1 shows the literature summary of proposed study.

Table 1: A Perspective of its Past and Future: Ethical AI for Unlearning Biased and Sensitive Information

Author / Study	Objective	Key Findings	Research Gap / Limitations
Cao & Yang et al., [11]	Introduce the concept of machine unlearning for removing training data influence from ML models	Proposed the foundation of unlearning by updating models without full retraining	Mostly theoretical and limited scalability for deep learning and large models
Bourtole et al., [12]	Develop efficient unlearning through partitioning (Sharding, Slicing, Aggregation)	Achieved faster unlearning compared to full retraining while maintaining performance	High storage overhead and complex training pipeline for real-world large datasets
Ginart et al. [13]	Study unlearning mechanisms for simpler ML models (e.g., convex/linear learning)	Efficient removal possible for specific algorithms with provable guarantees	Not generalizable to deep neural networks and complex architectures
Guo et al., [14]	Provide certified guarantees that deleted data is removed from model influence	Introduced formal verification concepts for unlearning success	Certification methods can be computationally expensive and difficult for large models
Shokri et al., [15]	Demonstrate privacy leakage by predicting if a record was part of training data	Showed models can leak training membership information	Focus is on privacy attack detection, not direct unlearning solution
Fredrikson et al., [16]	Investigate how models leak sensitive information through inversion	Proved attackers can reconstruct sensitive features using model outputs	Does not provide methods for removing leaked information from deployed models
Papernot et al., [3]	Explore privacy-preserving training methods and security of ML systems	Highlighted privacy risks in ML and proposed secure learning approaches	Preventive methods only; does not address “forgetting” already learned data
Dwork et al., [4]	Ensure privacy by limiting information leakage during training	Differential privacy reduces memorization and leakage in models	Often decreases model accuracy; cannot fully remove already learned sensitive knowledge
Mehrabi et al., [6]	Review sources of bias in AI systems and bias mitigation techniques	Bias is common across data, models, and deployment environments	Most methods reduce bias partially but do not remove bias from trained model memory

Barocas & Selbst et al., [7]	Study ethical and legal aspects of discrimination caused by algorithms	Algorithmic decisions can reinforce inequality in society	Focus is policy/legal; lacks practical machine-unlearning techniques
Guo, and Liu et al., [8]	To study how machine unlearning can help large language models forget undesirable or harmful learned behaviors, including harmful responses, copyrighted content, and hallucination-related knowledge.	The study found that unlearning can support ethical AI by removing unwanted model behavior using mainly negative examples. It is also more computationally efficient than full retraining and useful when the problematic training samples are known.	The approach still requires stronger verification methods to prove that the model has truly forgotten the targeted information. It may also risk reducing useful knowledge if the unlearning process is not carefully controlled.

To conclude this literature and associated work demonstrate that Ethical AI unlearning has turned into a critical field of research in order to develop fair, privacy-preserving, and trustworthy AI systems. The available literature attests to the in compositeness of bias mitigation and privacy protection by conventional dataset filtering or retraining models only. Machine unlearning represents a good line of work to eliminate biased and sensitive learning in the deployed models to become ethically compliant and gain better trust in society. Nevertheless, there are still some problems with optimizing unlearning, thus turning it into an efficient, verifiable, scalable, and attack-resistant process, which leaves plenty of research and innovation prospects in the domain of responsible AI creation.

METHODOLOGY

This paper presents an empirical approach to the learning process of an ethical machine unlearning to learn information that is not biased or sensitive and still achieve good performance. The general workflow will consist of the steps of training on the baseline with the use of BERT, unlearning and retraining/fine-tuning, during which various methods (fine-tuning and hybrid approach) are tested. Figure 2 shows the flow of methodology.

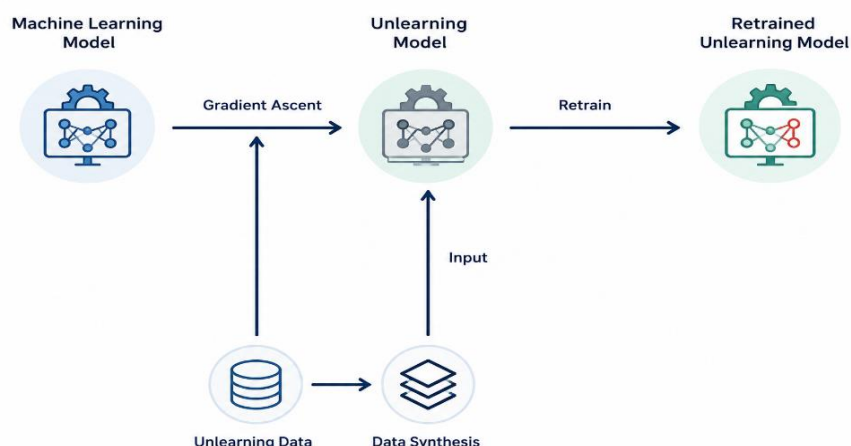


Figure 2: Proposed Methodology

Baseline Model Selection BERT

The initial step was a baseline pre-trained BERT model since it has shown good performance in text classification tasks and has the capacity to learn the contextual representations based on the language data. The original dataset was used to train the model (fine-tune) in order to set a performance benchmark. This is the baseline model acting as the before unlearning model that gives results of comparison in terms of accuracy, stability, and fairness.

Identification of Target Data for Unlearning

After baseline training, a specific subset of data was marked as sensitive or biased information, referred to as unlearning data. This unlearning subset may include user-related records or samples containing biased associations that should not remain in the model. The purpose of this step is to define what knowledge must be removed from the trained model in an ethical and controlled way.

Unlearning Process (Gradient Ascent-Based Forgetting)

The explicit unlearning method was used to unlearn the desired samples by employing a gradient ascent. The parameters of the model in this approach are updated using the unlearning data in the opposite direction of the learning direction. This move compels the model to be a little less assured and less reliant on the data which has been removed, allowing it to actually delete its effects and not just cover them up. Unlearning might cause a decrease in accuracy as a whole, or destabilize the model, so the unlearned model was retrained and tuned to use the remaining clean datas (retain set). The fine-tuning process can be used to regain the generalization of the model and at the same time making sure that the sensitive knowledge that was removed does not re-occur. The retraining is useful in enhancing performance and minimizing the adverse effects of direct unlearning. A hybrid model approach was also adopted to complete the compromise between strong forgetting and stable accuracy. This is a hybrid strategy that integrates two steps: 1. Explicit deletion Gradient ascent (forgetting phase) 2. Fine-tuning (recovery phase) and stabilization. The hybrid method was applied as it offers greater overall outcomes compared to the application of a single technique and thus, it is applicable to real-world application of ethical AI.

RESULTS AND ITS IMPLICATIONS

Baseline BERT model yielded good performance that was reliable on the dataset. It had an accuracy of 85.78 with precisions of 87.54, recalls of 94.72 and an F1-score of 90.99 indicating that the model was capable of classifying the data. Good class separation is also established by ROC-AUC score (0.8817).

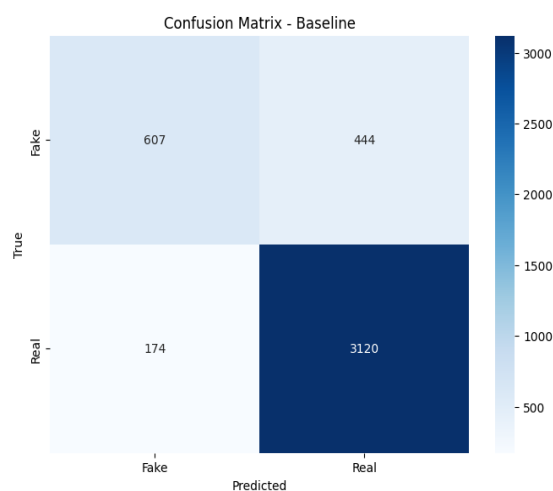


Figure 3: Confusion Matri for Baseline BERT

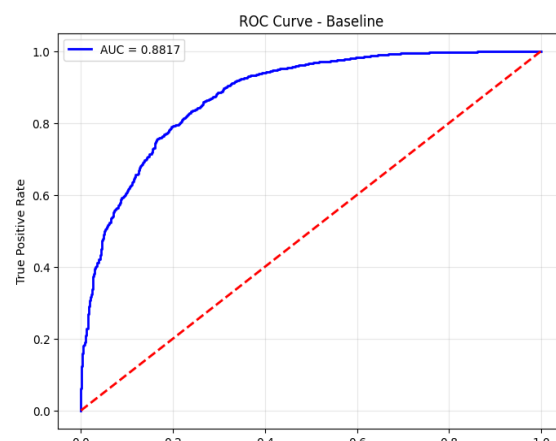


Figure 4: ROC-Curve for Baseline BERT

The confusion table indicates that there are 3120 correct (Real) and 607 correctly (Fake) predictions and also some misclassifications were evident (FP = 444, FN = 174). These findings demonstrate that BERT is a good baseline model and offers a solid point of reference in the comparison of unlearning experiments. Nonetheless, false positives also indicate that sometimes the model can be used to erroneously identify Fake samples as Real which is potentially dangerous in practice. Consequently, such a baseline performance can be used to assess the impact of unlearning and retraining on both the accuracy and ethical compliance.

Table 2: Model Performance Comparison Table

Model / Technique	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Baseline (BERT)	0.8578	0.8754	0.9472	0.9099	0.8817
Gradient Ascent (Unlearning)	0.7581	0.7581	1.0000	0.8624	0.5000
Fine-tuning (Retraining)	0.9328	0.9326	0.9824	0.9568	0.9612
Combined (Hybrid)	0.8741	0.8964	0.9429	0.9199	0.9089

The results of the comparison indicate clearly the effects of various strategies of unlearning and retraining on the performance of the model. The Baseline BERT model had a balance between the accuracy of the model (0.8578) and the recall (0.9472) and the ROC-AUC (0.8817) that were good hence making the model a valid reference before implementing any unlearning technique. Once the Gradient Ascent Unlearning was implemented, the precision declined dramatically to reach 0.7581. Despite the recall reaching to 1.0000 the ROC-AUC dropped to 0.5000 meaning poor class separability and unstable predictions. This implies that gradient ascent causes the model to forget too fast, but it harms the learning generalization and hence poor discrimination between the classes. Table 2 shows the model comparison performance. Hence, although it offers a better forgetting capability, it impacts adversely on the practical use of the model. Conversely, the strategy that yielded the best results in terms of virtually all measures is the Fine-tuning (Retraining) approach. It had the largest accuracy (=0.9328), precision (=0.9326), F1-score (=0.9568) and ROC-AUC (=0.9612). This signifies that retraining assists the model to regain stability and increase generalization without interfering with high classification performance.

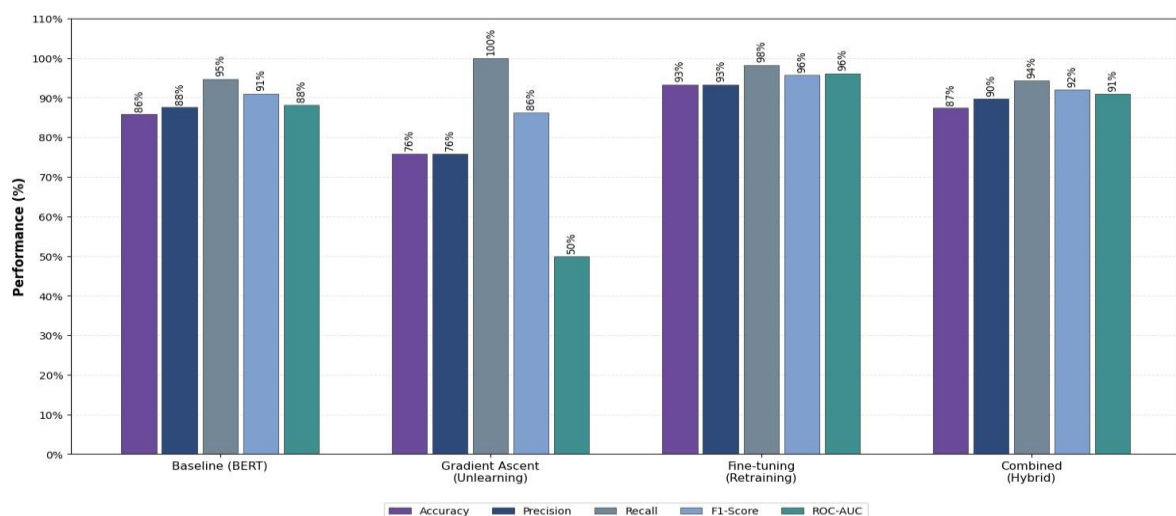


Figure 5: Model Performance Comparison

After any damage caused by unlearning, fine-tuning is the best way of restoring accuracy. Lastly, the Combined (Hybrid) performed in a way that was balanced with accuracy increasing to 0.8741 and ROC-AUC to the better of 0.9089 which is not better than the baseline but still worse than the fine-tuning. Figure 5 shows the bar graph of this comparison performance. This technique illustrates a trade off in practice, in which unlearning eliminates vulnerable patterns, and retraining recovers model performance. Nevertheless, it is not the most accurate when compared to fine-tuning. The discussion has validated the fact that unlearning by itself negatively affects performance, but re training and hybrid strategies can restore performance stability. Although combining the two is more ethically sound, fine-tuning is the most powerful of all the strategies, and maintaining accuracy and forgetting are in conflict with each other. Figure 5 depicts the comparison output of this study in the form of confusion matrix.

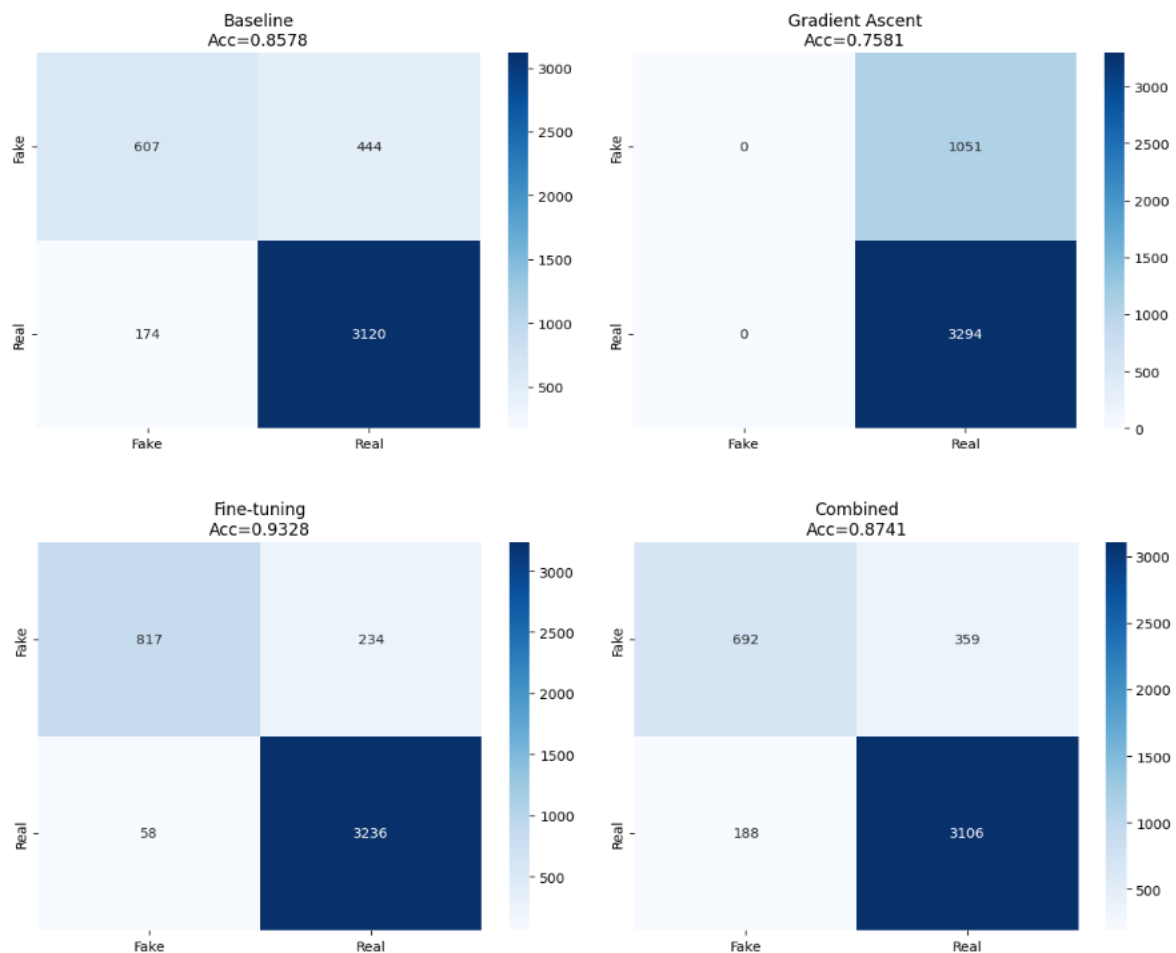


Figure 5: Model Performance Comparison Confusion Matrices

The confusion matrices show exactly the effects that each technique has on the classification performance. The overall prediction of the Baseline BERT model is good (607 True Negatives and 3120 True Positives); however, it still has quite a good number of errors (444 False Positives and 174 False Negatives). With the use of Gradient Ascent unlearning the model is highly imbalanced and the class of fake is not found at all as TN = 0 and all fake samples are misclassified as real (FP = 1051), which proves that aggressive unlearning harms the decision boundaries of the model. Conversely, Fine-tuning (retraining) provides the most significant improvement by both incrementing right classifications (TN = 817, TP = 3236) and decreasing errors (FP = 234, FN = 58), thus the most consistent and the most precise method. The Combined (hybrid) approach offers moderate performance (TN = 692, TP = 3106) yet at a higher rate of misclassification (FP = 359, FN = 188) than fine-tuning, which means that,

despite the ability of the hybrid unlearning approach to balance forgetting and recovery, the former remains the most reliable approach overall.

CONCLUSION

This paper has examined Ethical AI based on machine unlearning to eliminate the biased and sensitive data of a BERT-based model. The initial performance results proved that BERT has powerful classification performance prior to unlearning. The trade-off of gradient ascent unlearning was, however, more evident as it lowered the accuracy and harmed the separation of classes. Unlearning followed by fine-tuning showed the highest overall performance, which recovered the stability and better prediction qualities. The hybrid (combined) method offered a compromise solution by incorporating forgetting and recovery and was, however, computationally more difficult. In general, the findings confirm that unlearning is not enough, and it should be accompanied by retraining methods. More sophisticated techniques of unlearning can be adopted in future work to cause stronger forgetting with little loss in accuracy. Fairness and bias measurement metrics can also be introduced into future studies to get a more profound ethical analysis. Moreover, such success of unlearning can be proven by privacy audits and attack-resistance tests to enhance credibility. Lastly, the introduction of the techniques to more extensive datasets and on practical sensitive applications in the real world will reinforce a practical implementation of ethical unlearning systems.

REFERENCES

- 1) J. Cao and J. Yang, "Towards Making Systems Forget with Machine Unlearning," in Proc. IEEE Symp. Security and Privacy (SP), 2015, pp. 463–480.
- 2) L. Bourtole et al., "Machine Unlearning," arXiv preprint arXiv:1912.03817, 2019.
- 3) C. Guo, T. Goldstein, A. Hannun, and L. van der Maaten, "Certified Data Removal from Machine Learning Models," in Proc. Int. Conf. Machine Learning (ICML), 2020.
- 4) R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in Proc. IEEE Symp. Security and Privacy (SP), 2017, pp. 3–18.
- 5) S. Mehrabi et al., "A Survey on Bias and Fairness in Machine Learning," ACM Computing Surveys, vol. 54, no. 6, 2021.
- 6) J. Cao and J. Yang, "Towards Making Systems Forget with Machine Unlearning," in Proc. IEEE Symp. Security and Privacy (SP), 2015, pp. 463–480.
- 7) L. Bourtole et al., "Machine Unlearning," arXiv preprint arXiv:1912.03817, 2019.
- 8) C. Guo, T. Goldstein, A. Hannun, and L. van der Maaten, "Certified Data Removal from Machine Learning Models," in Proc. Int. Conf. Machine Learning (ICML), 2020.
- 9) R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in Proc. IEEE Symp. Security and Privacy (SP), 2017, pp. 3–18.
- 10) M. Fredrikson, S. Jha, and T. Ristenpart, "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures," in Proc. ACM CCS, 2015.
- 11) C. Dwork, "Differential Privacy," in Proc. ICALP, 2006, pp. 1–12.
- 12) N. Papernot et al., "Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data," in Proc. ICLR, 2017.

- 13) S. Mehrabi et al., “A Survey on Bias and Fairness in Machine Learning,” *ACM Computing Surveys*, vol. 54, no. 6, 2021.
- 14) S. Barocas and A. D. Selbst, “Big Data’s Disparate Impact,” *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016.
- 15) A. Gohilkar, A. Achille, and S. Soatto, “Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks,” in *Proc. CVPR*, 2020.
- 16) R. Neel, A. Roth, and S. Sharifi-Malvajerdi, “Descent-to-Delete: Gradient-Based Methods for Machine Unlearning,” in *Proc. NeurIPS*, 2021.
- 17) T. T. Nguyen et al., “A Survey of Machine Unlearning,” *ACM Computing Surveys*, vol. 57, no. 1, pp. 1–38, 2024.
- 18) [2] D. Choi and D. Na, “Towards Machine Unlearning Benchmarks: Forgetting the Personal Identities in Facial Recognition Systems,” *arXiv:2311.02240*, 2023.
- 19) [3] Y. Yao, X. Xu, and Y. Liu, “Large Language Model Unlearning,” *Advances in Neural Information Processing Systems*, *NeurIPS* 2024.
- 20) [4] J. Yao, E. Chien, M. Du, X. Niu, T. Wang, Z. Cheng, and X. Yue, “Machine Unlearning of Pre-trained Large Language Models,” *Proceedings of ACL* 2024, pp. 8403–8419, 2024.
- 21) [5] B. Tian et al., “To Forget or Not? Towards Practical Knowledge Unlearning for Large Language Models,” *Findings of EMNLP* 2024, pp. 1524–1537, 2024.
- 22) [6] K. Tran and S. S. Woo, “Fairness and Robustness in Machine Unlearning,” *Companion Proceedings of the ACM Web Conference* 2025, pp. 1336–1340, 2025.