Kong, S. C., Lan, Y. J., Zhao, J. H., Song, Y. J., & Mitrovic, T. (Eds.). (2025). *Proceedings of the 3rd International Conference on Metaverse and Artificial Companions in Education and Society*. Hong Kong: The Asia-Pacific Society for Computers in Education.

The Art of Unlearning: An Exploration of the Affective Privacy Unlearning Model in Online Education

Yaowen KUANG [1], Junjie Gavin WU [2], Yiyu WU [1], Tao WANG [1*]
[1] Hubei Key Laboratory of Digital Education, Faculty of
Artificial Intelligence in Education, Central China Normal University, Wuhan
[2] Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR
kuangyaowen@mails.ccnu.edu.cn, gavinjunjiewu@gmail.com, wuyy@mails.ccnu.edu.cn, tmac@ccnu.edu.cn,

Abstract: *With the rapid advancement of online education, the phenomenon of emotional deficiency has become increasingly pronounced. In recent years, there has been a growing emphasis within academia on researching the application of emotional interaction and affective computing technologies in online learning environments, with the objective of effectively mitigating the shortcomings in emotional communication prevalent in this domain. Affective computing aims to explore learners' emotional states by utilizing technological means to collect various state data during the learning process, thereby identifying and analyzing the emotional characteristics of learners. However, the collection and use of learner data raise a series of ethical and legal issues, including privacy and data security concerns. These privacy-related information and characteristic patterns are interrelated and highly sensitive. The leakage could pose numerous security risks, thereby restricting the full application and realization of the value of learning data in educational transformation. With the growing maturity of machine unlearning technology, we integrate it into the affective computing model within online learning environments. This study innovatively proposes the concept and technical framework of the Affective Unlearning Model (AUM), and demonstrates the implementation of the AUM within a facial affective computing model in online learning environments as an example. The research results indicate that this affective unlearning model can effectively achieve the unlearning of private data without occupying significant computational resources, thus safeguarding the students' right to be forgotten. Additionally, it ensures the subsequent usability of the model.*

Keywords: affective computing, machine unlearning, online learning, privacy preservation

1. Introduction

In recent years, AI has deeply integrated into education, especially with the rise of online learning platforms accelerated by the COVID-19 pandemic. While online classrooms play a vital role in teaching, the challenge of "affective-cognitive dissociation" can lead to learner isolation and fatigue, hindering online education's progress. Affective computing is a broad technology on education that utilizes artificial intelligence to learn and perceive student emotions., facilitating the delivery of personalized learning services. Personalized services require extensive collection of learners' data, such as facial expression. However, collecting and processing personal data poses significant ethical and legal risks, especially concerning privacy and data security. It is worth noting that even when private data have been removed from the database, the machine learning model can still retain and "remember" much of the underlying information(Nguyen et al., 2022). The "right to be forgotten" has been codified in privacy laws like the European Union's General Data Protection Regulation (GDPR)(2016) and China's Personal Information Protection Law (2021). In online education, this right ensures students' private information—including raw data in databases and traces in models like affective computing systems—is erased upon graduation or account termination. By minimizing unnecessary data retention and dissemination, such measures mitigate privacy breaches and safeguard user security.

To fulfill this requirement for forgetting, the simplest way is to retrain a new model using the retained data. However, this approach is often impractical because it consumes huge computing resources. Therefore, to effectively remove the specified data and its influence on the affective computing model, we propose an Affective Unlearning Model (AUM) from a technical perspective, utilizing the technical, "machine unlearning" (Aman et al., 2021). With our proposed the AUM, we can remove the specific student's privacy data from the model without the need for retraining while maintaining the usability. This approach not only safeguards intellectual property and privacy, but also ensure that the model retains its original predictive power even after the specified privacy information is removed.

2. Affective Unlearning Model

2.1. Definition of the Affective Unlearning Model

The AUM refers to the process by which a trained Affective Computing Model (ACM) intentionally forgets specific privacy data points from its training dataset, effectively rendering the model as though it had never encountered the forgotten data. Consider a cluster of data that we want to remove the training dataset from the trained ACM, denoted as x^*. An unlearning process U is defined as a trained Affective Computing Model (M_{ACM}), a training dataset D, and an remaining dataset $D/x*$ to a unlearned Affective Unlearning Model M'_{AUM}, which ensures that M'_{AUM} performs as though it had never seen the unlearning dataset $x*$. In this manner, the unlearning process is defined as:

$$M'_{AUM} = U\left(M_{ACM}, D, D/x^*\right)$$

Based on the above definition of the Affective Unlearning Model, the core objectives of the affective unlearning model include the following three main aspects:

Data deletion and privacy protection. To ensure students' exercise of the right to be forgotten, institutions must implement data deletion protocols to eliminate unnecessary personal information, preventing misuse. This requires balanced mechanisms that maintain model accuracy and functional integrity while removing sensitive data, achieving optimal equilibrium between privacy protection and system performance.

Data poisoning response and harmful data deletion. When the ACM is subject to data poisoning attacks, it may misinterpret students' emotional states, resulting in incorrect classification results. We can maintain the security and integrity of the data by using unlearning to delete the harmful data. This prevents it from spreading further or causing additional damage. The model can more accurately assess students' emotional states, allowing it to offer resources and recommendations that better meet their needs and enhance the overall user experience.

Remove biases to ensure fairness. In the context of affective computing for online learning, models can be adversely affected by inherent biases or errors present in historical training data, resulting in biased predictions and unfair decisions. To address these potential sources of bias, we implement a forgetting mechanism. This approach not only significantly enhances the predictive accuracy of the models but also serves as a powerful catalyst for advancing educational decision-making toward a more just and equitable framework.

2.2. The technical framework of the Affective Unlearning Model

Based on the aforementioned definition of the AUM, this study constructed its technical framework, as illustrated in Figure 1. Specifically, the AUM comprises three main components: model training, model unlearning and model evaluation.

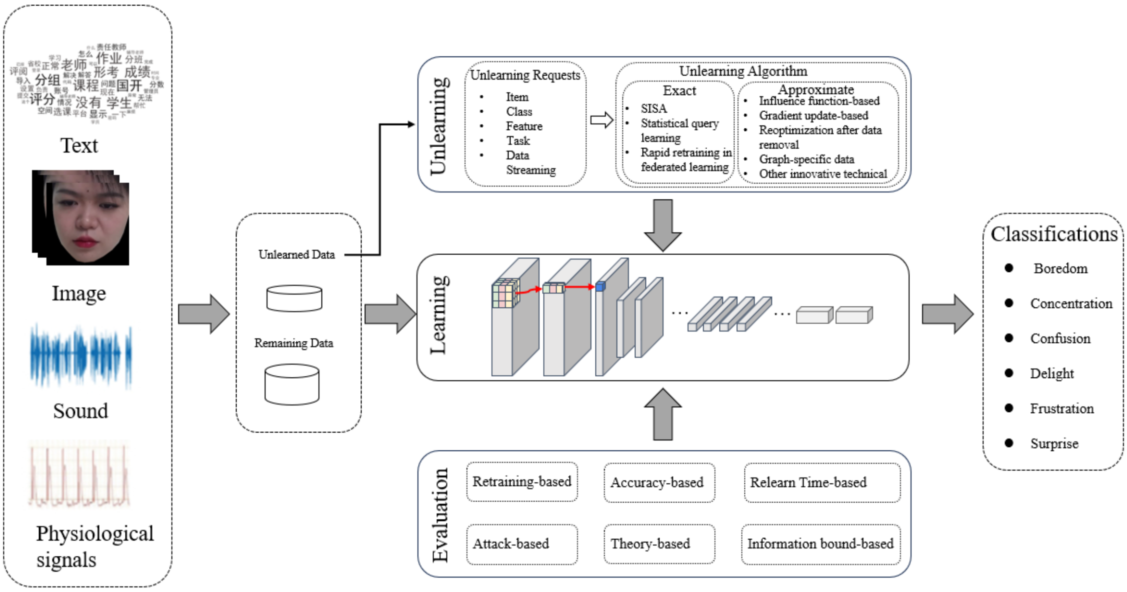

Figure 1. The framework of affective unlearning model.

①Model training: The meticulous extraction of facial expression features is the construction of an online affective computing model. First, capture facial images through the device. Next, it is essential to carefully extract key and relevant feature information from facial images. Good features are a vital part of an emotion recognition. Deep learning has proved to be a very effective feature extraction method. Last training the affective computing model (ACM) utilizing the extracted features. The trained ACM can be deployed in online learning scenarios to support personalized learning, facilitate educational assessment, and provide feedback. For example, Mehta et al.(2022) introduced a three-dimensional DenseNet self-attention neural network (3D DenseAttNet) for the automatic detection of students' engagement in E-learning platforms.

②Model unlearning: These unlearning algorithm include: item-level forgetting(Bourtoule et al., 2021); feature-level forgetting(Warnecke et al., 2021); class-level forgetting(Tanha et al., 2020); task-level forgetting(Liu et al.,2022); and data streaming forgetting(Nguyen et al., 2017). Within the framework of AUM in online learning, tailored forgetting algorithms can be effectively chosen to address the various types of forgetting requests from individual students. Specifically, in cases where a student requests the deletion of all their facial data for privacy protection, this data can be categorized as a class, allowing class-level forgetting algorithms to be used to completely erase all facial data associated with that student. Furthermore, when managing textual datasets that contain sensitive information, such as personal names and genders, these sensitive terms can be classified as features forgetting. By employing feature-level machine forgetting algorithms, these confidential details can be effectively removed from the textual data. Additionally, in light of the potential for racial bias in facial recognition software(Rhue, 2018), where the interpretation of emotions may vary according to a person's race, feature-level forgetting algorithms can be utilized to mitigate the influence of racial bias within such software systems.

③Model evaluation: A thorough evaluation of the AUM is crucial. Accuracy: as the primary measure of model performance, accuracy is crucial in this context. The AUM should consistently demonstrate efficient affective computing to ensure that data unlearning is accomplished without compromising predictive accuracy. Verifiability: to ensure effective protection of private data, a series of attack tests, including backdoor attacks, membership inference attacks, and model inversion attacks, are conducted. These tests aim to verify whether the AUM successfully defends against potential data leakage risks, thus demonstrating its effectiveness in safeguarding private information.

3. Application of the Affective Unlearning Model

This study explores facial expression-based affective computing models, which rely on extensive student facial image datasets. These images contain highly sensitive personally identifiable information which is not only unique and irreversible but also includes additional privacy details like the individual's age and gender, posing severe privacy risks if leaked. To mitigate such risks, we focus on implementing effective unlearning mechanisms to ensure thorough deletion and protection of specific facial image data from the model .

Figure 2. Example images that are sampled from our dataset.

In this study, we employed a self-constructed educational dataset. This dataset was meticulously developed from actual online learning environments. This dataset includes 11,442 samples, as shown in the Figure 2. We use the UNSIR algorithms (Choi et al., 2023) for unlearning. The method works as follows:

Error-Maximizing Noise: Introduce an error-maximizing noise. During the error maximization process, we maintain the weights of the pre-trained model in a fixed state. Given a noise matrix randomly initialized using a normal distribution N (0, 1), we aim to determine the noise that maximizes the model's classification loss function on unlearning data. We prepare for the subsequent erasure of information by determining the noise that is the inverse of the unlearning data.

Impair and Repair: For impair, we train the model on a small subset of data from the original distribution which also contains generated noise. This step effectively destroys the weight parameters of the initial model to remove the memory of the unlearning data from the initial trained model. For Repair, since the impair step may compromise the identification of the remaining data, we implemented a repair step. We retrain the impaired model on the remaining data for a single epoch to repair the model weights, ensuring that the model retains high performance.

We adopted top-2 accuracy as the primary metric. Additionally, we employed Membership Inference Attack (MIA)(Shokri et al. 2017) and established the forgetting score, where a lower score indicates better forgetting performance. Furthermore, We introduced the Normalized Machine Unlearning Score (NoMUS). This score effectively quantifies the overall performance of affective unlearning models by balancing the model's utility with the forgetting score.

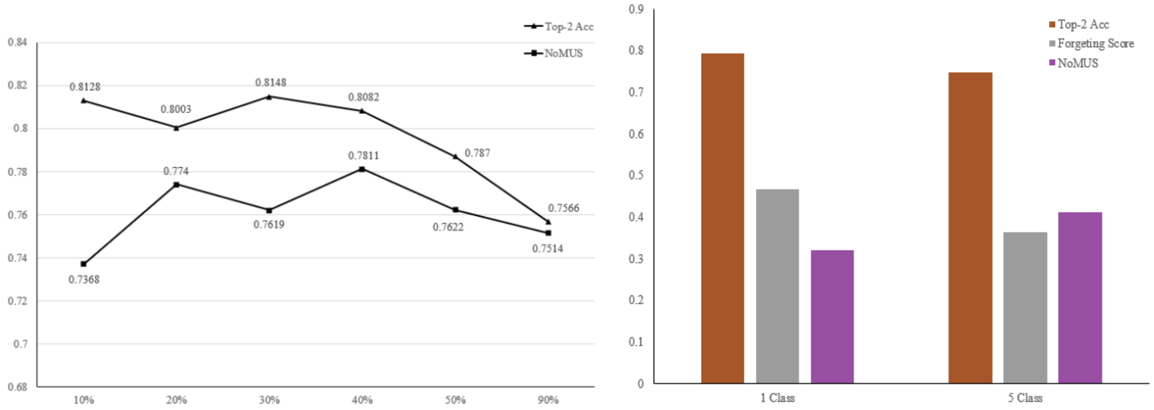

Figure 3. The performance of the AUM on forget class of data in the scenario of randomly forgetting face data and forgetting face data with specific identity.

(a) randomly forgetting face data (b) forgetting face data with specific identity.

We show the performance of our proposed method in the scenario of randomly forgetting face data and forgetting face data with specific identity. Results demonstrate the wide applicability of our method. Figure 3(a) depicts the performance of the AUM in the context of its forgetting mechanism following the random dropout of data at varying proportions. The observed results indicate a declining trend in accuracy as the proportion of forgotten data escalates. Nevertheless, amidst the pursuit of a balance between data privacy preservation and data availability, even post the elimination of 90% of data, the model's accuracy persists within acceptable bounds. Further scrutiny into the performance of NoMUS reveals that irrespective of whether 10% or 90% of data is randomly discarded, favorable forgetting performance are obtained, strongly suggesting the model's robustness concerning variations in data forgetting rates.

Within the framework of the AUM, we define the complete set of an individual's facial images as a class. Figure 3(b) meticulously illustrates the performance of the AUM after the forgetting of facial image data of one student (i.e., single class, 1 class) and five students (i.e., multiple classes, 5 classes). The method proposed in this study demonstrates outstanding unlearning outcomes in the task of forgetting data across multiple classes. This signifies the effective removal of relevant data from the model, rendering it unrecoverable through membership inference attacks, thereby robustly establishing the efficacy and security of the AUM model in data forgetting.

Table 1. Overall performance of various machine unlearning algorithms on our dataset with λ = 1/2. We emphasize the best score using bold and the second best score using italics. In the forgetting score, the lower is better.

Method	Test Acc	Forgetting score	NoMUS
Original Model	82.40%	0.2666	0.7916
Retrain Model	72.94%	0.0213	0.5302
FineTuningModel	75.04%	0.1321	0.6751
AUM	75.57%	0.1267	0.6654

Our results are compared with three baseline unlearning methods in Table 1. The original model exhibits the highest top-2 accuracy of 0.8240. However, the AUM model exhibits favorable performance in terms of top-2 Acc compared to other forgetting methods. For forgetting scores, the AUM model's efficacy in data privacy protection is approaching that of the retrain model, demonstrating its ability to maintain high model performance while safeguarding data privacy. For the NoMUS, the result indicates that the AUM model is capable of effectively implementing unlearning strategies without consuming large amounts of computational resources comparable to those of the retrained model, while maintaining a high level of recognition accuracy, thus achieving a good balance between performance and efficiency.

Thus, the AUM model can efficiently achieve the forgetting of students’ private data, thereby effectively safeguarding the students’ right to be forgotten and genuinely ensures the security of educational data.

4. Conclusion

The AUM framework integrates machine unlearning to enhance trust and security in affective computing, while promoting student engagement in digital learning. It optimizes data efficiency by reducing redundancy and streamlining management/analysis processes, thereby improving learning data quality for informed instructional decisions and personalized education. Through algorithmic fairness enhancements, the framework mitigates model bias and improves emotional recognition accuracy, eliminating demographic barriers (e.g., race/economic status) to educational resources while reducing systemic discrimination risks. This innovation fosters inclusive learning ecosystems that nurture socio-emotional skills and empathy through unbiased educational interactions.

AUM plays a pivotal role in online learning, which not only effectively protects students’ privacy, but also optimizes the performance and reliability of the system, laying the foundation for a harmonious and efficient online learning environment. This study theoretically defines the core concepts of the AUM, constructs the framework, and explores specific practical application scenarios, thereby laying a solid foundation for further exploration of the affective unlearning. In the feature, we will continue to be dedicated to promoting the widespread application of this model in the field of education, accelerating the process of educational intelligence.

Acknowledgements

This study was funded by National Natural Science Foundation of China under Grants 62477017, 62372333, 62277021 and 62277029, and Fundamental Research Funds for the Central Universities, Central China Normal University, under Grants CCNUai017, CCNUai018, CCNU22JC011 and CCNU22QN013.

References

Choi, D., & Na, D. (2023). Towards machine unlearning benchmarks: Forgetting the personal identities in facial recognition systems. *arXiv preprint* arXiv:2311.02240.

Deng, H., Yang, Z., Hao, T., Li, Q., & Liu, W. (2022). Multimodal affective computing with dense fusion transformer for inter-and intra-modality interactions. *IEEE Transactions on Multimedia*, 25, 6575-6587.

Elatlassi, R. (2018). Modeling student engagement in online learning environments using real-time biometric measures: electroencephalography (EEG) and eye-tracking.

Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., & Marsic, I. (2018, July). Multimodal affective analysis using hierarchical attention strategy with word-level alignment. *In Proceedings of the conference. Association for Computational Linguistics. Meeting* (Vol. 2018, p. 2225).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Liu, B., Liu, Q., & Stone, P. (2022, November). Continual learning and private unlearning. In *Conference on Lifelong Learning Agents* (pp. 243-254). PMLR.

Mehta, N. K., Prasad, S. S., Saurav, S., Saini, R., & Singh, S. (2022). Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement. *Applied Intelligence, 52(12)*, 13803-13823.

Nguyen, T. T., Duong, C. T., Weidlich, M., Yin, H., & Nguyen, Q. V. H. (2017). Retaining data from streams of social platforms with minimal regret. In *Twenty-sixth International Joint Conference on Artificial Intelligence*.

Regulation, P. (2018). General data protection regulation. *Intouch*, 25, 1-5.

Rhue, L. (2018). Racial influence on automated perceptions of emotions. Available at SSRN 3281765.

Savchenko, A. V., Savchenko, L. V., & Makarov, I. (2022). Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing, 13(4)*, 2132-2143.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)* (pp. 3-18). IEEE.

Standing Committee of the National People's Congress. Personal Information Protection Law of the People's Republic of China. 2021.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big data, 7,* 1-47.

TS, A., & Guddeti, R. M. R. (2020). Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and information technologies, 25(2),* 1387-1415.

Warnecke, A., Pirch, L., Wressnegger, C., & Rieck, K. (2021). Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*.

Zhang, Y., Tao, X., Ai, H., Chen, T., & Gan, Y. (2024). Multimodal Emotion Recognition by Fusing Video Semantic in MOOC Learning Scenarios. *arXiv preprint arXiv:2404.07484*.