Complex & Intelligent Systems (2025) 11:463
https://doi.org/10.1007/s40747-025-02086-2

ORIGINAL ARTICLE

GPT-based lifelong learning and ANFIS-driven reply memory ratio prediction for aspect-based sentiment analysis

Huang Huang[1] · Xiaohong Ye[2] · Mumtaz Begum Mustafa[2] · Qiyuan Dong[2] · Yu Li[3] · Adeleh Asemi[2,4] · Asefeh Asemi[5]

Received: 10 January 2025 / Accepted: 29 August 2025

Abstract

GPT is among the most powerful large language models (LLMs), known for its versatility and strong performance across a wide range of tasks. However, its substantial computational demands and limited cross-domain generalization present challenges, particularly in resource-sensitive applications like Aspect-Based Sentiment Analysis (ABSA). Existing ABSA models are typically domain-specific and suffer from catastrophic forgetting—losing previously acquired knowledge when sequentially trained on new domains—resulting in poor scalability and knowledge retention. To address these issues, we proposed a novel framework that integrates GPT-2 with a replay-based lifelong learning mechanism to support incremental, multi-domain ABSA while mitigating forgetting. The model is sequentially fine-tuned using real-data replay across four diverse ABSA domains: Laptops, Restaurants, Tweets, and Finance. Experimental results show that proposed model achieves an average accuracy of 0.85 and a Backward Transfer (BWT) score of -0.09, significantly outperforming the baseline model without lifelong learning (accuracy of 0.70, BWT of -0.64). We further conducted a t-test to validate the statistical significance of the improvements. we design an ANFIS-based model trained on experimental results to predict the proper memory ratio for new datasets, enabling more effective and adaptive lifelong learning in GPT-based architectures. In addition, we applied a multi-level data augmentation pipeline, which significantly improved performance across domains ($p = 0.0049$), enhancing both retention and generalization under constrained memory. Additionally, we introduced a domain sequence permutation study to test robustness against task-order sensitivity, while another part evaluates generalization under a simulated fifth domain constructed from a mixture of all original domains. These components validate the model's scalability and ability to generalize beyond trained distributions. An ablation study was conducted to isolate the contributions of each module, including replay, ANFIS prediction, and data augmentation. The results demonstrate that removing any single component leads to a measurable drop in performance, confirming that each part of the framework contributes meaningfully to overall effectiveness. A post-hoc comparison with a distillation-based lifelong learning baseline was conducted, showing improved performance of our replay+ANFIS approach. We acknowledge that additional comparisons with advanced baselines such as GEM or A-GEM are needed to further strengthen this claim, which we leave for future work. Overall, this study demonstrates the effectiveness of combining GPT-2 with replay-based lifelong learning and adaptive memory control, offering a scalable and robust solution for continuous, multi-domain sentiment analysis. Our proposed model provides a practical foundation for future research in dynamic, cross-domain ABSA systems.

Keywords Aspect-Based sentiment analysis · GPT · Lifelong learning · ANFIS · Data augmentation · Replay memory ratio prediction

✉ Adeleh Asemi
adeleh@um.edu.my

[1] School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou, Zhejiang 325000, China

[2] Department of Software Engineering, Faculty of Computer Science & Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia

[3] College of Computing, Illinois Institute of Technology, Chicago, IL, USA

[4] Institute of Artificial Intelligence and Social and Advanced Technologies, Isf.C., Islamic Azad University, Isfahan, Iran

[5] Corvinus University of Budapest, Budapest 1093, Hungary

Published online: 10 October 2025

Introduction

With the widespread use of smartphones, internet users have transformed into active content creators, posting reviews daily on platforms like TAOBAO.COM. These user-generated reviews often contain rich expressions of sentiment, which reflect human perceptions through senses like sight, sound, taste, smell, and touch—typically conveyed through text [1]. Sentiment analysis has evolved from general opinion classification to more fine-grained methods, such as Aspect-Based Sentiment Analysis (ABSA). ABSA is a subdomain of natural language processing (NLP) that focuses on identifying specific aspects or features mentioned in text and determining the sentiment toward each aspect [2]. This approach provides more detailed insights than general sentiment analysis, enabling businesses to understand opinions about specific product or service attributes.

Unlike humans, who can reuse and build upon prior knowledge with ease, AI systems often face a major challenge: catastrophic forgetting—where learning new information leads to the erosion of previously acquired knowledge. Lifelong learning aims to address this issue by enabling AI models to adapt continually while retaining essential information across tasks. While deep learning models perform well on narrowly defined tasks, they often lack human-like flexibility. Lifelong learning strengthens model generalizability by maintaining a balance between acquiring new knowledge and preserving earlier learning. In ABSA, this capability is crucial, as real-world applications involve analyzing reviews from diverse, dynamic, and previously unseen domains—far beyond the static, single-domain setups most current models are trained on.

Recent studies have explored various methods to improve ABSA using lifelong learning and generative models like GPT. One such method involves prompt tuning for multi-domain ABSA, offering a lightweight mechanism for adapting to different tasks [3]. However, this technique heavily relies on manually designed prompts, limiting its flexibility in novel or highly diverse domains. Other research has incorporated replay-augmented memory frameworks to simulate long-term memory for continual sentiment learning [4]. While effective, such approaches raise concerns about data privacy and scalability due to their dependence on storing large quantities of previous task data. Another approach involves adaptive parameter updating to reduce forgetting during continual learning [5]. Although this method maintains performance on prior tasks by controlling parameter drift, it still depends on manual rule-setting, reducing its applicability to dynamic, real-world settings.

For instance, Mohammad AL-Smadi and colleagues [6] proposed a deep learning model combining Gated Recurrent Units (GRUs) and features from the Multilingual Universal Sentence Encoder (MUSE). Their "Pooled-GRU" model, trained on Arabic hotel reviews, effectively identifies aspects and classifies their sentiment. Despite its domain-specific success, the model struggles to generalize across languages or technical domains. This highlights a common limitation across existing models: while they perform well within their trained scope, they falter when exposed to new or varied data. Multi-domain training has been attempted to bridge this gap, but such models are still susceptible to catastrophic forgetting, significantly degrading in performance when new domains are introduced.

Moreover, many rehearsal-based methods continue to suffer from catastrophic forgetting due to unpredictable shifts in domain characteristics when new tasks are introduced [6]. These observations reveal three persistent challenges in ABSA: (1) limited generalization across diverse domains, (2) difficulty assimilating new domain information without losing prior knowledge, and (3) dynamically adapting memory usage in lifelong learning scenarios. These gaps form the core motivation of this study.

In response to these challenges, this study introduces a model designed to incrementally acquire new domain-specific knowledge while retaining previously learned information. The foundation is a traditional generative model, enhanced with lifelong learning strategies to combat catastrophic forgetting. To address the challenge of dynamically adapting memory usage in lifelong learning scenarios, we introduce an Adaptive Neuro-Fuzzy Inference System (ANFIS) aimed at predicting the optimal memory ratio based on dataset-specific performance characteristics. While many lifelong learning methods rely on manual tuning or exhaustive search to determine memory size, our proposed ANFIS provides an intelligent and data-driven mechanism that leverages prior experimental results across domains. The system uses key performance indicators—F1-score, recall, precision, and METEOR—from preliminary learning phases as inputs to infer the most suitable memory ratio for new tasks. Although this ANFIS is not used during the current experiments, it is trained on our results and offers a promising direction for future work where the memory configuration can be adaptively tuned without manual intervention. This enhances the scalability and autonomy of lifelong learning frameworks, particularly in domain-agnostic GPT-based systems.

The structure of this paper is as follows: The Introduction section provides an overview of the research context, highlights the core problem, and outlines the objectives of the study. The Related Work section reviews relevant literature to identify existing knowledge gaps. The Proposed Model section describes the architecture of the model and the experimental methodology. The Experimental Design section outlines three groups of experiments conducted

to evaluate the proposed model. The Results and Discussion section presents and analyzes the experimental findings. Finally, the Conclusion section summarizes the key outcomes of the study and suggests directions for future research.

Related works

In the Related Works section, we review recent advancements across four key areas relevant to this study: Language Models for Aspect-Based Sentiment Analysis (ABSA), Lifelong Learning Methods, Data Augmentation Techniques, and the application of Adaptive Neuro-Fuzzy Inference Systems (ANFIS). These topics form the foundational landscape for the development of our proposed framework. To provide a structured overview, we present two comprehensive review tables that summarize and compare recent works on (1) language model-based approaches for ABSA and (2) lifelong learning strategies. These tables highlight key trends, model architectures, strengths, and limitations in the current literature, offering a clear context for our contributions. The section also discusses the role of data augmentation in improving model robustness, and how ANFIS has been applied in related predictive tasks, motivating its integration into our dynamic memory control mechanism.

Language models for ABSA

In aspect-based sentiment analysis (ABSA), a variety of deep learning models have been employed to improve performance, particularly in the tasks of aspect extraction and sentiment classification. Recurrent Neural Networks (RNNs) were among the first deep learning architectures used in this area, well-suited for processing sequential text and capturing short-term contextual dependencies. These models are effective when sentiment cues are located near the target aspect. However, RNNs are limited in their ability to capture long-range dependencies due to the vanishing gradient issue. For example, Wang et al. [7] introduced the RNCRF (Recursive Neural Conditional Random Fields) model, which integrates a Discrete-Time RNN (DT-RNN) with Conditional Random Fields (CRF) and was applied to domains such as laptops and restaurants. While this approach achieved promising results in those specific domains, it struggled with broader domain generalization and continued to be affected by the vanishing gradient problem when handling long-term contextual information.

Long Short-Term Memory (LSTM) networks were developed to overcome the limitations of traditional RNNs. By incorporating memory cells capable of maintaining information over extended sequences, LSTMs effectively capture long-range dependencies in text. This makes them particularly suitable for sentiment analysis tasks where emotional cues may appear far from the target aspect within a sentence. However, the enhanced memory capabilities come at a cost—LSTMs require significant computational resources due to their large number of parameters, making them less efficient for large-scale datasets. For example, Li et al. [8] implemented an LSTM-based model for the laptop and restaurant domains, but the use of multiple LSTMs and manual memory operations significantly increased computational overhead, reducing its scalability.

To address LSTM's computational complexity, Gated Recurrent Units (GRUs) were introduced as a more lightweight alternative. GRUs streamline the architecture while preserving performance, offering faster training and fewer parameters. This makes them well-suited for capturing key patterns in aspect-sentiment pairs in ABSA tasks [9]. However, GRUs may struggle with more complex data or large-scale tasks, where their performance may fall short compared to more sophisticated models like transformers. For instance, Ma et al. proposed HMBi-GRU for sentiment analysis in English and Spanish tweets. Although the model achieved promising results, it was hindered by its hierarchical structure, limited generalization capability, overfitting risks, and heavy reliance on domain-specific features.

The emergence of transformer-based models marked a major advancement in ABSA. BERT (Bidirectional Encoder Representations from Transformers) introduced the ability to capture both preceding and succeeding context within a sentence, significantly improving the performance of aspect extraction and sentiment classification. Despite these advantages, BERT is constrained by a fixed input length of 512 tokens and lacks generative capabilities, limiting its flexibility for tasks involving longer texts or dynamic generation. For example, Seker et al. [10] applied AlephBERT to Hebrew comment analysis, but the model inherited BERT's token limit. XLNet, a permutation-based transformer model, further advanced contextual understanding by modeling sentence permutations to capture richer dependencies. While XLNet improves aspect-sentiment modeling, it also demands domain-specific fine-tuning for best performance [11]. Su and Yu's XLNetCN model demonstrated these improvements in the laptop and restaurant domains, but still encountered difficulties when applied to longer documents or more specialized ABSA tasks that required further adaptation.

Although BERT offers strong bidirectional context modeling, its limitations in handling longer inputs and lack of generative capabilities remain key drawbacks. GPT-2 addresses these issues by supporting longer sequences (up to 1024 or 2048 tokens) and leveraging a generative, autoregressive architecture. This enables GPT-2 to excel in

nuanced sentiment detection and long-form text processing, providing improved cross-domain generalization. However, because GPT-2 processes text in a left-to-right manner, it lacks the bidirectional understanding of BERT, which can impact accuracy in certain tasks. For example, Jiang developed a GPT-2-based ABSA model for the laptop and restaurant domains. While it benefited from GPT-2's generative strengths, the model still faced challenges like catastrophic forgetting during continual training [12].

Table 1 provides a comparative summary of existing deep learning-based ABSA models and their domain-specific applications.

Lifelong learning methods

Lifelong Learning (LL) methods are designed to enable machine learning models to continuously learn from new data while preserving previously acquired knowledge,

Table 1 Summary table of existing deep learning models

Article	Model	DATASET	Strength	Limitations
[7]	RNCRF: DT-RNN+CRF	SemEval Challenge 2014 Task 4	• Bidirectional propagation between aspects and opinions • Accurate representation of contextual information	• Shows limited ability to generalize across different domains and needs domain-specific adjustments • Fails to capture long-range dependencies effectively due to vanishing gradients
[8]	A deep multi-task learning framework based on LSTM.	SemEval 2014 (Laptop domain) SemEval 2016 (Restaurant domain)	• Memory Interaction: Improves aspect-opinion extraction by modeling co-occurrence relationships. • Multi-task Learning: Three LSTMs enhance task management and accuracy.	• Complexity: Higher computational demands due to multiple LSTMs and manual memory operations, making it less efficient for very large datasets. • Domain-Specific: Needs customization and fine-tuning when applied to different domains.
[9]	Hierarchical Multi-layer Bidirectional Gated Recurrent Units (HMBi-GRU)	English Tweets: Spanish Tweets:	• Performs exceptionally well in joint learning. • Attains cutting-edge performance in target detection. • Shows strong effectiveness in sentiment analysis.	• Challenges arise from the complexity of hierarchical layers. • Limited ability to generalize across diverse datasets. • Risk of overfitting due to the complexity of the model. • Dependence on data-specific features for optimal performance.
[12]	GPTs.	SemEval V2 Dataset: MAMS Dataset:	• Strong generalization and enhanced performance in few-shot sentiment analysis. • Exceptional at text generation and managing longer input sequences.	• Reliance on backbone models for performance. • Limited GPT applicability across some tasks. • Dataset limitations impacting model generalization. • Challenges in prompt engineering for achieving optimal results. • Continues to face issues with catastrophic forgetting.
[11]	XLNetCN	SemEval 2014 Task 4	• Improves aspect-awareness through auxiliary sentences. Utilizes Capsule Networks to capture local features.	• Greater complexity in the model architecture. • Elevated risk of overfitting. • Reliance on the quality of auxiliary sentences. • Demands domain-specific fine-tuning for optimal performance.
[10]	AlephBERT	Oscar (Hebrew portion) Wikipedia (Hebrew) Twitter (Hebrew)	• Performs exceptionally well in a range of Hebrew NLP tasks. • Achieves high performance across different Hebrew NLP tasks.	• Limitation of fixed-length input (512 tokens).
[43]	BERT based model	SemEval-2014 Task 4 Twitter Dataset	• Captures context within sentences, • Effectively handles mixed sentiments • Improves the overall accuracy of sentiment analysis.	• Absence of generative ability
[44]	BERT based model	SemEval-2016 Task 5	• Offers a comprehensive ABSA model • Excels in aspect classification Effectively performs sentiment analysis • Demonstrates strong generalization across various domains	• Difficulties in accurately identifying multiple aspects in a single text • Challenges in analyzing sentiments for each individual aspect Potential for misidentifying aspects, resulting in inaccurate interpretations • Complexity grows with longer or more complex sentences

ensuring that the models maintain high performance in ever-changing environments. The advantages of LL methods include enhanced knowledge retention, the ability to adapt to new information, and improved efficiency. There are three main types of LL approaches: Architecture-Based Lifelong Learning (ABLL), which focuses on adjusting the model's structure to integrate new tasks without forgetting old knowledge; Regularization-Based Lifelong Learning (RBLL), which uses regularization techniques to prevent the model from forgetting by penalizing modifications to previously learned parameters; and Data-Based Lifelong Learning (DBLL), which prioritizes data selection and replay strategies to reinforce learning from past experiences. These approaches offer a comprehensive framework for creating models that can continually adapt and retain knowledge over time.

Architecture-Based Lifelong Learning (ABLL) methods modify neural network architectures to accommodate new tasks while retaining previous knowledge, often by adding new layers or modules without disrupting existing information. An example of this is "Progressive Neural Networks" [13], which expand dynamically as new tasks are introduced. However, as more tasks are added, the architecture can become increasingly complex, leading to higher computational costs for both training and inference. This added complexity can also pose challenges when deploying models in resource-limited environments [14]. Furthermore, continuously expanding the architecture to handle new tasks may result in diminishing returns in performance, especially in larger models, which can limit the scalability of ABLL methods across a wide range of tasks [15].

In contrast, Regularization-Based Lifelong Learning (RBLL) methods use regularization techniques to avoid catastrophic forgetting, ensuring that learning new tasks doesn't negatively impact performance on previously learned ones. One common RBLL method is Elastic Weight Consolidation (EWC), which adds a regularization term to the loss function to penalize large changes to crucial parameters, helping retain knowledge while learning new tasks [16]. However, selecting the right regularization parameters can be difficult and often requires careful tuning. Poor tuning may lead to either insufficient knowledge retention or overly strict constraints that hinder learning [14]. Additionally, adding regularization terms to the loss function can increase the computational load during training, as the model must balance learning new tasks with retaining old knowledge [16]. Their strong reliance on memory configuration. These methods often depend on empirical tuning and heuristics for setting buffer size, sample selection, and replay frequency, making performance highly sensitive to specific tasks and resources. Moreover, they typically require large memory capacities, and their effectiveness drops sharply under memory constraints, worsening catastrophic forgetting. This reliance limits their practical deployment [3, 54]. Compared to ABLL and RBLL, Data-Based Lifelong Learning (DBLL) excels in efficiently preserving knowledge through selective experience replay [17, 18] without needing complex architectural changes or parameter adjustments. This enables DBLL to maintain model performance across various tasks while minimizing computational complexity [15]. Furthermore, DBLL is highly compatible with GPT-2, as its focus on data replay allows GPT-2 to maximize computational resources without dedicating significant power to the replay process. The replay mechanism relies mainly on storage, rather than altering the model's architecture or tuning its parameters, thereby enhancing the model's computational efficiency.

Data-Based Lifelong Learning (DBLL) methods focus on managing and utilizing data effectively. These approaches involve techniques like maintaining and revisiting past data, applying data augmentation methods, or using selective memory mechanisms to preserve key information from previous tasks. The advantages of DBLL include improved model stability and performance over time, better handling of shifts in data distributions, and an enhanced ability to generalize from both new and old data [19]. By integrating and utilizing diverse datasets, DBLL methods enable models to adapt and improve continuously as new information becomes available [20]. Replay-based lifelong learning is a subset of DBLL methods that prevents catastrophic forgetting by storing and replaying data from previous tasks during training on new tasks. This can involve exact replay, where real data is stored and revisited [17, 18], or generative replay, where synthetic data mimicking past experiences is generated [21, 22], allowing the model to learn new information while maintaining performance on earlier tasks. Robins [19] introduced techniques like "sweep rehearsal" and "pseudorehearsal," which use generated data to mitigate forgetting in neural networks. These methods have been effective in maintaining task performance and training networks without requiring the original data. However, they face challenges when handling complex or noisy data, which increases computational demands. Rebuffi et al. [17] proposed an exemplar management technique that efficiently preserves prior knowledge through resource optimization, making it scalable as new class sets are added. Despite the resource efficiency of replay-based learning, balancing old and new knowledge remains a challenge, which can affect learning outcomes.

Under the continual learning framework, models must not only tackle the challenge of forgetting over time but also demonstrate strong robustness and generalization in data-scarce (few-shot) new domains and when faced with domain sequences that exceed the number and complexity

of those seen during training [56]. However, existing evaluation paradigms often fall short of accurately simulating such real-world complexities. For instance, standard benchmarks rarely assess model performance under limited supervision. Moreover, when testing generalization to truly unseen domains—those beyond the number and distribution of training domains—external datasets are often introduced, bringing issues such as inconsistent labeling that undermine the scientific rigor of the evaluation [55].

Recent studies have explored knowledge distillation for continual learning, where a student model incrementally learns from soft targets generated by a teacher trained on past data. Such approaches help mitigate forgetting by encouraging the student to preserve output distributions over old tasks [18]. However, they often require full access to pre-trained teachers and may suffer from underfitting when applied across heterogeneous domains. Similarly, meta-learning methods such as Reptile [52] and ANIL [53] optimize the model's initialization or adaptation rules to quickly generalize to new tasks. Although effective in some settings, they require episodic training, which increases computational overhead and limits scalability in generative models like GPT. In contrast, our replay+ANFIS-based method provides a lightweight, scalable solution without requiring model architecture changes or heavy pre-training.

More recently, Pellegrini et al. [20] introduced Latent Replay, which improves continual learning by storing intermediate activations rather than raw data. This method offers real-time learning on edge devices and is flexible, allowing adjustable layer selection for mobile deployment. However, its effectiveness is limited by the degradation of stored activations over time, and its success depends on selecting the optimal layer during the learning process (Table 2).

Recent research has explored the direct integration of GPT-style models with lifelong learning strategies, introducing memory-aware adaptation techniques to address catastrophic forgetting in sequential ABSA tasks [3]. Previous foundational work has utilized memory networks and parameter-isolation methods to enhance long-term knowledge retention across multiple aspect domains [23, 24].

To further improve knowledge preservation and cross-task generalization, contrastive learning has emerged as a powerful approach. Some frameworks combine continual learning with contrastive objectives to preserve inter-task feature discrimination in aspect sentiment classification [25]. While not specifically targeting ABSA, advancements in representation learning have demonstrated that incorporating contrastive objectives with self-distillation or feature replay strengthens model robustness and efficiency under non-stationary conditions [26, 27]. These findings inform the development of contrastive-inspired memory strategies and privacy-preserving model updates in our framework.

Additionally, the increasing use of generative methods and recent performance evaluations of large language models like GPT-3.5/4 in ABSA tasks emphasize their growing role in capturing nuanced sentiment representations [28, 29]. However, these approaches often lack structured mechanisms for retaining long-term knowledge across tasks. To overcome this challenge, our proposed framework combines continual learning with lightweight adaptation techniques and privacy-aware constraints, enabling more stable and scalable lifelong sentiment modeling.

Current deep-learning ABSA models face significant challenges in generalizing across different domains [30]. While advanced models, such as those using GRU [6] or Hierarchical Knowledge Enhancement [31], perform well within their specific trained domains, they often struggle to maintain effectiveness when applied to new or different domains. This limitation typically requires extensive adaptation and fine-tuning to accommodate the unique characteristics and data distributions of new domains, making it challenging for these models to deliver consistent performance across diverse contexts [8, 11].

Moreover, these models face issues with catastrophic forgetting when incorporating data from new domains [12]. As models are updated with new information, they frequently experience a decline in performance on previously learned tasks due to the loss of previously acquired knowledge. Although methods like Elastic Weight Consolidation (EWC) [16] and COgnitive REplay (CORE) [32] have been proposed to mitigate this issue, they still struggle to effectively balance retaining past knowledge with integrating new data. This ongoing challenge underscores the need for more robust solutions that can maintain or even improve performance while continually adapting to new learning tasks.

To address the two primary challenges faced by current ABSA models, a combined approach utilizing GPT-2 and replay-based lifelong learning methods is essential. GPT-2's advanced autoregressive architecture plays a key role in tackling domain generalization issues, as it excels at generating and understanding complex text patterns across a wide range of sequences. This enables GPT-2 to adapt more effectively to new domains, unlike traditional models that often struggle with domain shifts. By leveraging GPT-2's flexibility and strong text generation capabilities, the model can be fine-tuned for domain-specific tasks with minimal adjustments, thereby improving its performance across diverse contexts.

Replay-based lifelong learning methods help address catastrophic forgetting by maintaining a replay buffer of previously encountered data and revisiting it periodically during training. This ensures that crucial knowledge from earlier tasks is retained while incorporating new information.

Table 2 Summary review table of lifelong learningmethods

Article	LL Methods	Datasets	Results	Strength	Limitation
[20]	Latent Replay	CORe50, iCubWorld	Latent Replay reduces mem ory and maintains accuracy in object recognition.	- Efficient learning - Scalable - Real-time - Edge devices	- Aging effect - Trade-offs: accuracy, com putation, memory - Limited generalization in continual learning
[45]	Net2Net	ImageNet, CIFAR-10,MNIST	Faster training with high ac curacy via efficient knowledge transfer.	- Net2Net: accelerates neural network training - Enhances model exploration - Preserves function - Ensures iterative improve ment	- Requires specific architec tures - Adds computational de mands - Relies on teacher network quality
[46]	PackNet	ImageNet, Places365, CUBS Birds, Stan ford Cars, Flowers	Pruning increases error rates in ImageNet-trained VGG-16.	- PackNet: prevents forgetting - Maintains performance across tasks - Works with diverse architec tures - Low storage overhead	- Task order impact - Potential computational overhead - Inference limitations - Task type adaptation.
[47]	PathNet	Binary MNIST, CIFAR, SVHN	Positive transfer across tasks, faster training, robust to hyper parameter choices.	- PathNet: efficient multi-task learning - Knowledge retention - Uses genetic algorithms for pathway evolution - Excels in transfer learning - Scales well with tasks and network size	- Computational overhead - Uncertain generalization - Potential advantages of fixed paths - Reliance on simple evolu tionary algorithm
[13]	Progressive Neural Net works	Synthetic Pong Vari ants, Atari Games	Superior Transfer Perfor mance: Progressive networks demonstrate better transfer performance compared to traditional finetuning.	- Prevents catastrophic forget ting - Effective transfer - Rich compositionality - Flexibility	- Parameter growth - Task compatibility - Inference complexity - Optimization challenges
[19]	Rehearsal Mecha nisms	random bi nary vectors for simula tions	Sweep Rehearsal, Pseudore hearsal	- Minimizes catastrophic for-getting - Integrates new information - Effective without original training data	- Network architecture con straints - Performance decline with more items - Lack of real-world dataset application - Unexplored computational complexity
[17]	iCaRL	CIFAR-100, Im ageNet ILSVRC 2012	iCaRL excels in class incremental learning, outper forming finetuning and other methods. It maintains robust performance with increasing classes.	- iCaRL: mitigates forgetting - Scales with class additions - Uses memory efficiently	- Lower performance than batch-trained systems - Memory limits affect per formance - Computational costs vary
[48]	synapses mecha nism	MNIST and CIFAR	Synaptic Intelligence reduces forgetting, maintains accuracy across tasks.	- Reduces forgetting - Efficient computation - Improves generalization	- Synaptic complexity - Parameter sensitivity - Endpoint correlation im pact
[49]	A-GEM	MNIST, CI FAR, CUB, Split AWA	A-GEM excels in accuracy and efficiency among bench marks.	- A-GEM: efficient - Performs well - Scales effectively	- More worst-case forgetting compared to GEM - Regularization methods (e.g., EWC) excel with over parameterized models - Beneficial with multiple epochs
[50]	2-RCL	NA	The paper explores 2-RCL's trade-off between forgetting and intransigence with varying regularization parameters to manage dissimilar task learn ing.	- Theoretical insights into 2-RCL - Continual learning dynamics - Risk management	- Fixed design setting - Simplistic model (linear re gression) - Reliance on specific as sumptions

Table 2 (continued)

Article	LL Methods	Datasets	Results	Strength	Limitation
[2]	GEM	MNIST, MNIST, CIFAR	GEM excels in ACC, BWT, and FWT, outperforming EWC and iCARL.	- GEM: efficient - Manages memory well - Supports positive backward transfer	- Lacks structured task de scriptors - Lacks advanced memory strategies - Faces computational bottle necks
[51]	Bayesian online learning frame work	MNIST, MNIST, Split MNIST, CIFAR-100, Om niglot, and Ro boschool.	UCL shows superior perfor mance across diverse bench marks, including supervised and reinforcement learning tasks, achieving higher accu racy with fewer parameters than competing methods.	- UCL: efficient - High-performing - Adaptable across diverse benchmarks - Superior effectiveness in su pervised and reinforcement learning tasks	- Requires careful hyperpa rameter tuning - Complex for large networks or datasets

Integrating this replay mechanism with GPT-2 allows the model to continuously adapt to new data without significant performance degradation on previously learned tasks. This combination provides a more robust solution to the challenge of preserving past knowledge while accommodating new learning objectives, overcoming the limitations of current ABSA models.

Data augmentation

Data augmentation in NLP refers to the process of synthetically generating new training samples by modifying existing data, with the goal of improving model generalization. It addresses common challenges such as lack of semantic diversity and limited dataset size. Unlike computer vision, where image transformations are straightforward, textual data augmentation requires preserving both syntactic structure and semantic meaning, making it more complex. Effective augmentation techniques can improve model robustness, reduce overfitting, and enhance performance in low-resource scenarios and domain-specific tasks.

Recent studies in NLP have explored multi-level data augmentation strategies to improve model generalization. At the syntactic level, paraphrasing techniques based on text-to-text transformers like T5 have shown strong performance in generating semantically equivalent yet structurally diverse outputs [33]. These paraphrases are often evaluated using sentence-level semantic similarity measures, such as Sentence-BERT, to ensure meaning preservation [34]. At the semantic level, multilingual back-translation remains a widely adopted technique, where intermediate translation chains (e.g., EN→DE→FR→ES→EN) introduce lexical variation without distorting content [35]. For lexical-level augmentation, simple yet effective strategies such as random word swaps, deletions, or insertions—commonly referred to as Easy Data Augmentation (EDA)—have been shown to boost robustness in classification tasks [36].

More recently, scholars have proposed compositional data augmentation pipelines that integrate multiple augmentation layers (e.g., syntactic + semantic + lexical), producing higher-quality and more diverse training examples. Such multi-strategy designs help reduce overfitting and improve cross-domain generalization, especially in low-resource scenarios [37]. This layered approach reflects a growing trend toward combining complementary augmentation techniques to maximize both diversity and semantic fidelity in augmented datasets.

Adaptive neuro-fuzzy inference systems

Adaptive decision-making in dynamic environments requires models that are both interpretable and capable of handling uncertainty. One such approach that has gained increasing attention in recent years is the Adaptive Neuro-Fuzzy Inference System (ANFIS). ANFIS combines the learning capabilities of neural networks with the transparency of fuzzy logic, making it a powerful framework for modeling nonlinear systems and incorporating human expert knowledge. This hybrid nature has enabled its successful application in diverse fields, ranging from investment prediction to industrial process optimization.

In the domain of decision support, Asemi et al. [38] proposed an ANFIS-based model to customize investment types by considering demographic features and feedback from potential investors. Their approach demonstrated the ability of ANFIS to balance structured data learning with adaptive reasoning, outperforming rigid rule-based systems. Similarly, their subsequent work [39] explored the influence of managerial traits on investment decision-making, where ANFIS effectively captured fuzzy relationships between qualitative behavioral factors and investor responses.

Beyond the financial domain, ANFIS has shown strong modeling capability in environmental and engineering applications. In a comparative study, ANFIS was evaluated

alongside Response Surface Methodology (RSM) and Artificial Neural Networks (ANN) for predicting methylene blue dye adsorption using Oryza sativa straw biomass [40]. The results revealed that ANFIS not only provided competitive prediction accuracy but also better interpretability compared to purely black-box models. In fluid dynamics, ANFIS was applied to map computational data of water reservoir homogenization using air bubble flows, where it helped simplify complex physical simulations while preserving core system behavior [41]. Additionally, in industrial control systems, ANFIS enabled precise derating torque estimation for induction motors, outperforming conventional methods under fluctuating operational loads [42].

These studies underscore ANFIS's suitability for tasks that demand a balance between prediction performance and model explainability. Motivated by these strengths, this study proposes the adoption of ANFIS to estimate the optimal memory ratio in lifelong learning GPT-based architectures. The model leverages prior task performance indicators to infer appropriate memory allocation, offering a practical and interpretable approach for enhancing memory management in continual learning scenarios.

Proposed model

In the Proposed Model section, we present the overall architecture of our framework, which integrates GPT-2 with a replay-based lifelong learning strategy guided by an Adaptive Neuro-Fuzzy Inference System (ANFIS). This framework is designed to enhance cross-domain generalization in aspect-based sentiment analysis (ABSA) while minimizing catastrophic forgetting. Following the architectural overview, we outline the training procedure and describe how ANFIS dynamically supports optimal memory replay decisions during fine-tuning.

Architecture of model

The architecture is designed to incrementally acquire new domain-specific knowledge while preserving previously learned information, thereby addressing the issue of catastrophic forgetting typically encountered in sequential domain adaptation tasks.

The model operates in two primary stages: pretraining and fine-tuning. In the pretraining stage, GPT-2 is trained on a large-scale general corpus to acquire robust language representations. This pretrained model is then fine-tuned across a sequence of ABSA tasks from different domains (e.g., Domain1 to Domain4). Each domain presents unique aspect-sentiment pairs and linguistic characteristics, requiring adaptive learning while maintaining previously gained performance.

To mitigate forgetting, the model incorporates a replay-based learning strategy. When fine-tuning on a new domain (e.g., Domain3), the model also replays selected samples from earlier domains (e.g., Domain1 and Domain2). This technique consolidates prior knowledge while enabling adaptation to new tasks. The memory replay mechanism is visualized in Fig. 1 using backward-flowing arrows, representing the reuse of prior data during current domain training.

A novel addition to this architecture is the integration of ANFIS, which dynamically predicts the optimal replay memory ratio—that is, the percentage of data from previous domains to be replayed during fine-tuning. ANFIS receives four inputs (precision, recall, F1-score, and METEOR) and outputs a predicted memory ratio. It was trained on empirical data from memory replay tests conducted across multiple domains and ratios (0.1 to 0.9). As a result, ANFIS enhances adaptability by recommending replay configurations tailored to the learning context of each new domain.

For the ABSA task itself, GPT-2 is used in a question-answering (QA) format. For example, given the sentence "The food is good," the model is prompted with "Can you find out the aspect sentiment pairs?" and generates "Aspect: food, sentiment: 1." This generative design eliminates the need for task-specific architectural changes, leveraging GPT's flexibility for both aspect extraction and sentiment classification.

Overall, the proposed architecture is a robust framework for lifelong ABSA. By combining a pretrained GPT backbone, ANFIS-guided memory replay, and flexible prompt-based inference, it offers a scalable and intelligent solution for dynamic sentiment analysis in evolving multi-domain environments. While this study focuses on replay-based lifelong learning, we acknowledge alternative strategies such as distillation and meta-learning, which are discussed in the Related Works section and empirically contrasted in Result and Discussion section.

Research method

This research aims to assess the effectiveness of the proposed GPT-2 + ANFIS model in acquiring new domain-specific knowledge while minimizing catastrophic forgetting. To achieve this, we adopted a comprehensive methodology that includes: development of the hybrid model, data collection from diverse ABSA domains (laptop, restaurant, tweets, and finance), data preprocessing and formatting, training using dynamic memory replay, and evaluation through a series of experiments.

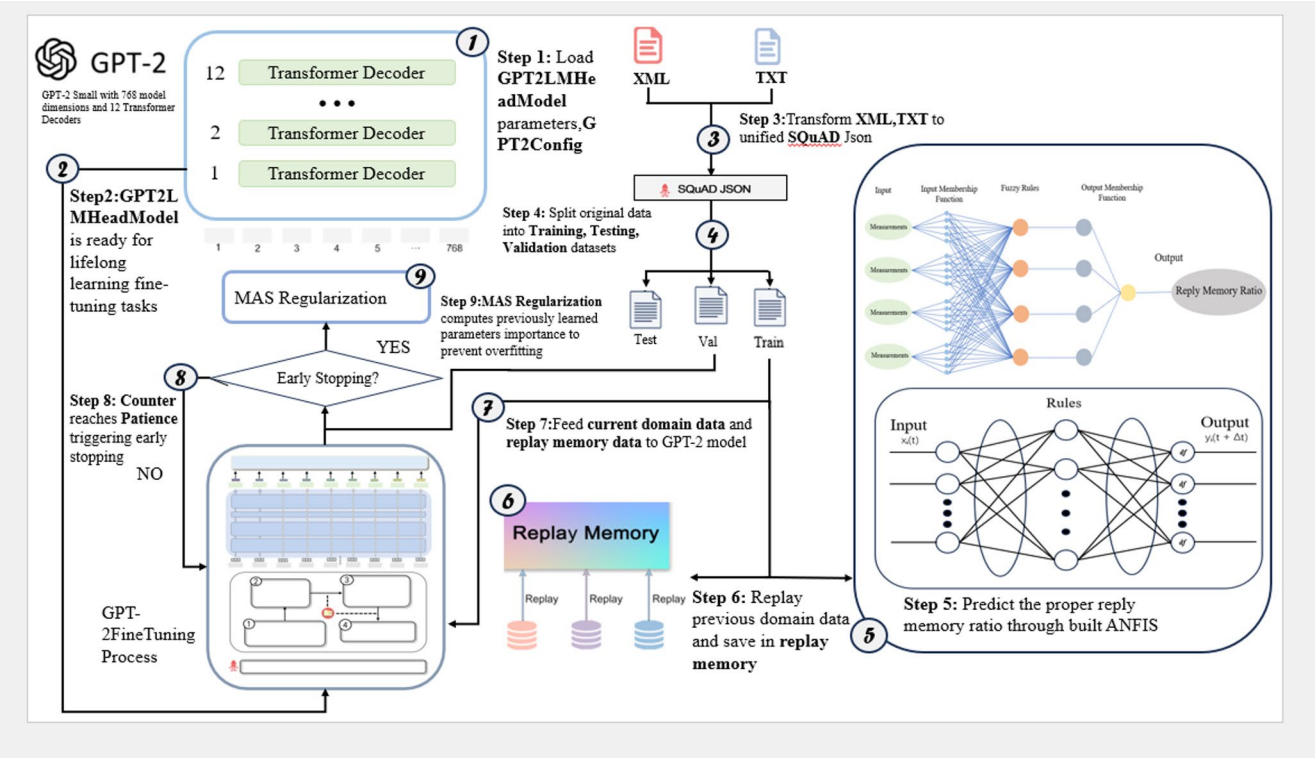

Fig. 1 Sequential domain adaptation using GPT-2 and ANFIS-guided memory replay

Table 3 ABSA datasets across four domains

Dataset	Format	Domain	Size
Tweets ABSA dataset	TXT	Tweets	6280
Laptop ABSA dataset	CSV	Laptop	1449
Restaurant ABSA dataset	XML	Restaurant	1708
Finance ABSA dataset	CSV	Finance	10,753

Note: TXT stands for 'Text File Document' CSV stands for comma-separated values) file. XML stands for extensible markup language

A key innovation in this study is the integration of an Adaptive Neuro-Fuzzy Inference System (ANFIS), trained to predict the optimal replay memory ratio during fine-tuning. By learning from performance metrics (Precision, Recall, F1-score, METEOR) across various memory configurations, ANFIS enhances the training strategy by dynamically selecting the most effective proportion of prior data to replay. This predictive memory management supports efficient knowledge retention and performance across evolving domains.

Datasets

Data collection

In this research, we collect four widely used ABSA datasets available on the internet: the Tweets ABSA dataset, the Laptop ABSA dataset, the Restaurant ABSA dataset, and the Finance ABSA dataset. The Tweets dataset was introduced in the paper "Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification" by Li Dong et al., presented at the 52nd Annual Meeting of the Association for Computational Linguistics (ACL) in 2014. Specifically, we use the SemEval-2014 Laptop ABSA dataset, the SemEval-2016 Restaurant ABSA dataset, and the SEntFiN-v1.1 Finance ABSA dataset. Table 3 provides the details of the original ABSA datasets across the four different domains.

Data preprocessing

Although the initial datasets were structured, both the Laptop and Restaurant datasets contain entries with single and multiple Aspect-Sentiment pairs. To improve clarity and facilitate analysis, we divided these datasets accordingly. The Laptop dataset was split into two separate datasets: one retaining entries with a single Aspect-Sentiment pair (Laptop Dataset) and another for entries with multiple Aspect-Sentiment pairs (Laptop-Multi Dataset). Similarly, the Restaurant dataset was divided into the Restaurant Dataset, which includes only entries with a single Aspect-Sentiment pair, and the Restaurant-Multi Dataset, designed for entries with multiple Aspect-Sentiment pairs. This study excludes the Laptop-Multi and Restaurant-Multi datasets, focusing on the primary four datasets—Laptop, Restaurant, Tweets,

and Finance—for model training and evaluation. The datasets are further split into a training set (70%), a testing set (20%), and a validation set (10%), with the dataset sizes shown in Table 4 below.

Data formatting

The original formats of these datasets vary, including TXT, CSV, and XML, and each contains the "original sentence, aspect, and corresponding sentiment." To ensure consistency and ease of use, we convert all datasets into a standardized JSON format (SQuAD Format). This format is commonly used in natural language processing (NLP) for training and testing models on their ability to understand and process text to provide accurate answers. Each JSON file encapsulates three key pairs of keys and values:

Context (C) This key stores the original sentence, which is an expression of a person's sentiment towards one or several specific aspects.

Question (Q) This key specifies the task assigned to the model.

Answer (A) This is where the output of the Aspect-Based Sentiment Analysis model is recorded.

Model training

The training process follows a multi-step strategy that integrates GPT-2's generative capabilities with ANFIS-guided replay-based lifelong learning:
Step 1: Load GPT2LMHeadModel parameters, GPT2Config.
Step 2: GPT2LMHeadModel is ready for lifelong learning fine-tuning tasks.
Step 3: Data Transformation to SQuAD JSON Format.
Step 4: Split original data into Training, Testing, Validation datasets.
Step 5: Predict the proper reply memory ratio through built ANFIS.
Step 6: Replay previous domain data and save in replay memory.
Step 7: Feed current domain data and replay memory data to GPT-2 model.
Step 8: Counter reaches Patience triggering early stopping.

Table 4 Summary of domains, data sizes

Domains	#Train	#Test	#Validation
Laptop	634	182	91
Restaurant	894	257	127
Tweets	4373	1256	619
Finance	7527	2161	1065

Step 9: MAS Regularization computes previously learned parameters importance to prevent overfitting.

By integrating ANFIS as a predictive component in the lifelong learning process, the training method becomes more adaptive and data-driven. This dynamic replay strategy improves both learning efficiency and memory retention across domains, providing a significant advancement over static replay configurations.

Experimental designs

In the Experimental Designs section, we present three comprehensive experiments designed to evaluate the effectiveness, adaptability, and scalability of the proposed model. Experiment I investigates the model's performance under static memory settings to demonstrate catastrophic forgetting, establish performance baselines, and evaluate lifelong learning across four ABSA domains. Experiment II introduces a dynamic memory control strategy using an ANFIS model to intelligently predict the optimal replay memory ratio per domain based on training performance. Experiment III explores the impact of structured data augmentation techniques on knowledge retention under fixed memory constraints. Together, these experiments validate the proposed model's ability to mitigate forgetting, generalize across domains, and maintain efficiency through adaptive memory and augmentation strategies.

Experiment I: static memory based experiment

This experiment investigates the performance of a GPT-2 based model under three settings: sequential domain learning without lifelong mechanisms (Part 1), domain-specific independent training (Part 2), and training with lifelong learning strategies (Part 3). The domains used in this experiment are denoted as a sequence $\{D_1, D_2, \ldots, D_i\}$, where D_1 represents the first domain and so on. Specifically, we refer to D_1, D_2, D_3, and D_4 as Dlaptop, Drestaurant, Dtweets, and Dfinance, respectively. Ti represents the test dataset corresponding to Di. We define E_{ij} as the model accuracy when trained on Di and evaluated on T_j without applying lifelong learning, while R_{ij} denotes the same when lifelong learning is incorporated.
Part 1: Demonstrating Catastrophic Forgetting in Sequential Domain Learning

To examine catastrophic forgetting, the model is sequentially fine-tuned across D_1 to D_4 without applying any lifelong learning strategies. At each stage, the model is evaluated on all previously encountered test sets $\{T_1, T_2, T_3, T_4\}$. By comparing E_{11}, E_{21}, E_{31}, and E_{41}, we assess how

performance degrades over time on earlier domains, providing evidence of catastrophic forgetting.

Part 2: Establishing Baseline Performance Metrics Across Individual Domains

This section aims to establish benchmark performance in each individual domain. The GPT-2 model is independently trained and tested on each D_i, yielding E_{11}, E_{22}, E_{33}, and E_{44} as baseline results. These values serve as reference points to evaluate how much knowledge is lost or retained during multi-domain training in other settings.

Part 3: Evaluating Lifelong Learning through Sequential Domain Adaptation

We then implement lifelong learning mechanisms (e.g., replay-based methods) and evaluate the model's ability to retain previously learned knowledge while acquiring new domain information. Performance under lifelong learning is denoted by R_{ij}, indicating that the model was trained on D_i and tested on T_j. This allows us to evaluate forgetting ($i>j$) and forward transfer ($i<j$), as summarized in the lifelong learning performance table (Table 5). The results help reveal whether the model benefits from continual learning in multi-domain environments.

Part 4: Few-Shot Domain Adaptation Setup

In addition to the standard full-data setting, we also considered a low-resource scenario to evaluate the model's robustness under limited supervision. Specifically, for each of the four training domains (Laptop, Restaurant, Tweets, and Finance), we randomly sample 30% of training data per domain under a fixed seed and repeat this process 3 times independently. This design helps mitigate the variance caused by a single random split and provides a more reliable estimate of performance under constrained data conditions. The average across runs is reported for consistency.

Part 5: Testing Generalization Under Extended Domain Sequences

To assess model behavior when trained on more than four domains, a simulated fifth domain is constructed by combining samples from the four domains. The sampling process is controlled to ensure semantic diversity and label balance, allowing the simulated domain to reflect a realistic but unseen distribution shift. This design enables evaluation of cross-domain generalization in a reproducible and cost-effective way without introducing new datasets.

Table 5 Lifelong learning performance evaluation table

After Training	Test on			
	D1	D2	…	Dd
D1	R1,1	R1,1	…	R1,d
D2	R2,1	R2,2	…	R2,d
…	…	…	…	…
Dd-1	Rd-1,1	Rd-1,2	…	Rd-1,d
Dd	Rd,1	Rd,2	…	Rd, d

While incorporating an entirely external domain might appear ideal for testing generalization, it often brings confounding factors such as annotation inconsistency, mismatched label distributions, and domain-specific terminology. These differences make fair comparison with the original benchmark difficult. In contrast, the simulation-based approach maintains the benchmark's controlled structure while still capturing the key characteristics of domain shift, thus providing a scientifically grounded proxy for evaluating generalization beyond the observed domains.

Statistical evaluation of model Performance(t-test):

Hypothesis Formulation: To determine whether the proposed GPT-2 model enhanced with lifelong learning provides a statistically significant improvement over the original GPT-2 model, the following hypotheses were formulated for each domain and for the combined analysis:

- Null Hypothesis (H_0): There is no significant difference in the mean accuracy between the proposed GPT-2 with lifelong learning and the original GPT-2 model.

$$H_0 : \mu_1 - \mu_2 \leq 0$$

- Alternative Hypothesis (H_1): The proposed GPT-2 with lifelong learning has a significantly higher mean accuracy than the original GPT-2 model.

$$H_1 : \mu_1 - \mu_2 > 0$$

A one-tailed paired t-test was selected to compare the models, as the same test sets were used for both models and the research goal was to demonstrate improvement (not just difference).

Computational complexity

In this section, we provide a detailed computational complexity analysis of our proposed lifelong learning framework for aspect-based sentiment analysis (ABSA) using a GPT-based architecture. The analysis considers both time and space complexity, which are essential to evaluate the scalability and efficiency of the model across multiple domains.

Let T denote the number of tasks (domains), E the number of training epochs per task, B the number of training samples in each domain, and bs the batch size. The input sequence length is denoted as L, the hidden size of the GPT model as d, and the number of Transformer layers as H. Each training step involves a forward and backward pass through the GPT model, whose core computational unit is the self-attention mechanism. The self-attention operation in each

Transformer layer has a time complexity of $\mathcal{O}\left(L^2 \cdot d\right)$, due to the necessity of computing pairwise interactions between all tokens in the input sequence.

Since the model contains H layers, the per-step computation becomes:

$$\mathcal{O}\left(H \cdot L^2 \cdot d\right)$$

For each domain, training occurs over E epochs with B samples, leading to (B/bs) steps per epoch. Therefore, the total time complexity across all T domains is given by:

$$\mathcal{O}\left(T \cdot E\left(B/bs\right) \cdot H \cdot L^2 \cdot d\right)$$

Regarding space complexity, several components contribute to memory consumption:

- $\mathcal{O}\left(P\right)$: Model parameter storage.
- $\mathcal{O}\left(bs \cdot L \cdot d\right)$($bs \cdot L \cdot d$): Runtime memory for activations.
- $\mathcal{O}\left(B_{\text{replay}} \cdot L\right)$: Replay buffer memory.

Thus, the total space complexity is summarized as:

$$\mathcal{O}\left(P + bs \cdot L \cdot d + B_{\text{replay}} \cdot L\right)$$

It is worth noting that although the time complexity scales quadratically with the input sequence length due to attention, this can be alleviated in practice via optimized attention variants or truncated inputs. Additionally, the use of mixed-precision training (e.g., FP16) and data parallelism across GPUs reduces practical resource usage without changing the theoretical bounds. It should be noted that our analysis assumes the conventional quadratic-cost attention mechanism, which provides a clear baseline for comparison. However, recent advances such as sparse attention (e.g., Longformer, BigBird), low-rank projections (Linformer), or kernel-based linear attention (Performer) can reduce the asymptotic complexity to nearly linear in sequence length. While these methods were not implemented in our current framework, they represent promising directions to further improve scalability in lifelong ABSA. We will highlight this limitation in the discussion and consider integrating such optimizations in future work.

In summary, the proposed framework demonstrates polynomial time and space complexity, making it suitable for scalable and efficient lifelong learning on aspect-based sentiment tasks across diverse domains.

Experiment II: dynamic memory based experiment

This experiment aims to investigate how the size of replay memory affects model performance in a lifelong learning setting, particularly within the context of Aspect-Based Sentiment Analysis (ABSA). In addition to evaluating static replay configurations, we analyze domain-specific sensitivity to different memory ratios and introduce an adaptive control mechanism using the Adaptive Neuro-Fuzzy Inference System (ANFIS) to dynamically determine the most suitable replay ratio for each domain. The goal is to assess whether intelligent, data-driven memory scheduling can preserve past knowledge and support cross-domain generalization while minimizing memory usage.

To explore this, we designed an experimental setup involving sequential domain training combined with varying levels of memory replay. A memory buffer was employed to store a subset of previously encountered samples, enabling the model to jointly learn from current task data and retained samples from past tasks. This setup simulates the core principle of lifelong learning—preserving past knowledge while integrating new information.

We evaluated five replay memory ratios: 0.1, 0.3, 0.5, 0.7, and 0.9, corresponding to the retention of 10–90% of prior task data. The selected task domains included four distinct topics: Laptop (D_1), Restaurant (D_2), Tweets (D_3), and Finance (D_4). The model was trained on these domains sequentially and evaluated using two widely adopted metrics—ROUGE-1 (Precision, Recall, F1) and METEOR—to assess both syntactic and semantic learning quality.

To analyze domain-specific sensitivity to replay memory, the Restaurant domain (D_2) was fixed as a control point while varying the memory allocations for the remaining domains. This configuration was designed to isolate the influence of memory size on each domain's performance.

The results obtained from the dynamic memory-based experiment provide the empirical foundation for developing an ANFIS model capable of predicting the optimal memory ratio for lifelong learning scenarios. By systematically varying the memory replay ratios (0.1 to 0.9) across different domains—Laptop, Tweet, and Finance—we captured the performance impact of memory size on key metrics such as F1-score, Precision, Recall, and METEOR. These performance outcomes reflect each domain's sensitivity to retained knowledge, revealing non-linear relationships between memory allocation and model generalization. This rich dataset serves as the training basis for the ANFIS, where each instance encodes the domain type and corresponding performance metrics as input features, with the ideal memory ratio—derived from observed optimal performance—used as the target output.

The ANFIS model leverages this mapping to learn fuzzy if-then rules that generalize the influence of domain-specific characteristics and performance behavior on memory needs. Once trained, it enables intelligent and adaptive memory scheduling for unseen datasets by predicting a suitable memory ratio based on their semantic similarity and expected performance profile. This approach enhances lifelong learning efficiency by minimizing unnecessary memory usage while preserving prior knowledge, thus supporting scalable and context-aware GPT-based systems. The integration of ANFIS into lifelong GPT-based systems offers an interpretable and data-driven mechanism to dynamically balance memory constraints and learning effectiveness across domains.

ANFIS building and applying for reply memory ratio prediction

To enable intelligent prediction of the optimal memory ratio for each dataset in a lifelong learning frameworks, we designed an Adaptive Neuro-Fuzzy Inference System (ANFIS) model. This model is trained on the empirical results of our dynamic memory-based experiment, where different memory ratios were applied to a GPT-based ABSA task across various domains. The goal of this ANFIS is to predict the most suitable memory ratio (i.e., replay buffer size) based on measurable characteristics of a domain's learning performance. The process of implementation of ANFIS is shown in Fig. 2.

Step 1. ANFIS Inputs and Output

The ANFIS architecture is designed with four input variables and one output variable:

Inputs:

F1-score – Represents the harmonic mean of precision and recall, capturing the model's balance between accuracy and completeness.

Precision – Measures how many predicted positive instances are actually correct.

Recall – Captures the proportion of actual positives that were correctly identified by the model.

METEOR Score – A metric for evaluating the semantic and syntactic quality of generated content, useful in capturing nuanced improvements across replay configurations.

These four inputs were selected because they collectively represent the model's generalization capability, semantic understanding, and classification performance, all of which are influenced by memory size in lifelong learning.

Output:

Memory Ratio – The predicted replay memory size (between 0.1 and 0.9), which reflects the ideal proportion of past task data to retain in the buffer when training on a new domain.

Step 2. Dataset for ANFIS Training

The dataset used to train the ANFIS model is derived from experimental results obtained by varying memory

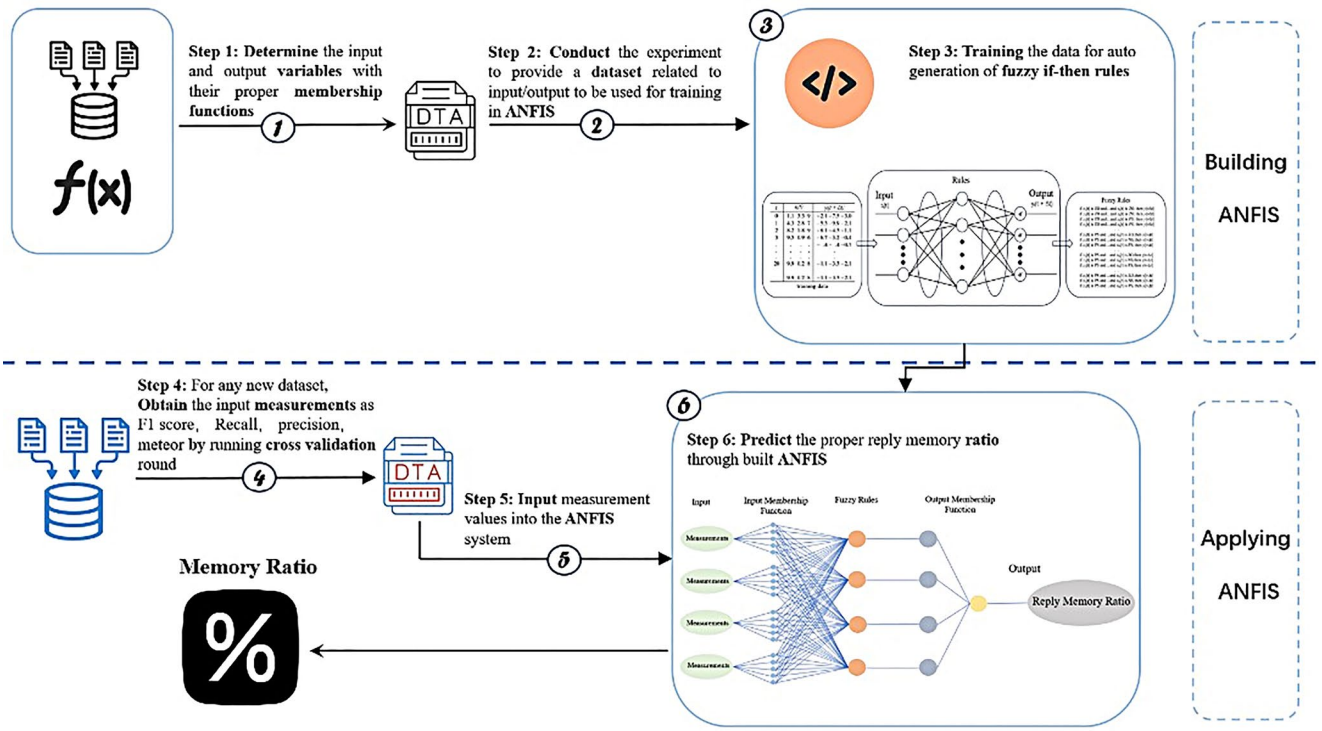

Fig. 2 Steps of building and applying ANFIS for reply memory ratio prediction

ratios across three target domains: Laptop, Tweet, and Finance. For each domain and memory ratio (0.1, 0.3, 0.5, 0.7, 0.9), we recorded the four input performance metrics. The memory ratio associated with the highest F1-score or stable multi-metric improvement in each domain setting was labeled as the "ideal" memory ratio and used as the ANFIS output. This yielded a total of 15 records across different configurations, serving as training data for the model.
Step 3. Fuzzy Rule Generation and Training

The fuzzy inference system was built using a first-order Sugeno-type model, with Gaussian membership functions applied to each input variable. Each input was assigned three membership functions: low, medium, and high. These linguistic terms allow the ANFIS to interpret input values in a human-understandable manner, enabling rule-based reasoning. The fuzzy rule base was automatically generated using a grid partitioning strategy, which constructs rules by combining input membership functions. For instance, a sample rule might state:

> If F1-score is high and Precision is medium and Recall is high and METEOR is high, then the memory ratio is 0.7.

The model was trained in MATLAB using hybrid learning—combining gradient descent and least squares estimation—until convergence was reached with minimal error.
Steps 4, 5 and 6. System Integration and Application

Once trained, the ANFIS model can be applied to a new domain dataset by feeding in its expected F1-score, Precision, Recall, and METEOR score (e.g., from small-scale pilot runs or estimated values based on domain similarity). In step 4, we apply cross validation method on new dataset to find the measures for F1-score, Precision, Recall, and METEOR. In Step 5, we enter these measures to ANFIS then in step 6 the fuzzy rules will be applied to inputs and ANFIS will produce the proper reply memory ratio for new dataset as output. A suggested memory ratio, allowing adaptive tuning of the lifelong learning system's replay memory. This enables dynamic memory control tailored to the performance profile of the incoming dataset, optimizing resource utilization while preserving learning efficacy.

Experiment III: data augmentation based experiment

This experiment aims to explore whether a structured data augmentation strategy can enhance knowledge retention and mitigate catastrophic forgetting in lifelong Aspect-Based Sentiment Analysis (ABSA), particularly under constrained memory conditions where only 30% of previous domain data is retained. The primary motivation is to introduce greater semantic diversity while preserving essential sentiment-related information, thereby improving the model's generalization ability and robustness across domains.

In the experimental setup, a fixed replay memory ratio of 0.3 is applied uniformly across all domains. The data augmentation pipeline is designed as a three-level sequential process, combining semantic, syntactic, and lexical transformations. At the semantic level, multi-language back-translation is used to introduce latent semantic drift through a chained translation path (English to German, then to French, Spanish, and back to English). This is followed by sentence-level paraphrasing using a pre-trained T5 model (Vamsi/T5_Paraphrase_Paws), which enhances structural variation in sentence construction. Finally, lexical-level diversity is introduced using Easy Data Augmentation (EDA) techniques such as random word swaps and deletions to simulate linguistic noise, especially relevant in informal or low-resource text domains. These transformations are applied sequentially to each selected training instance to form a composite augmentation strategy.

To ensure the integrity of the augmented data, two evaluation metrics are employed to assess augmentation quality, though they are not used for filtering. Semantic similarity between the original and augmented sentences is measured using Sentence-BERT embeddings, while surface-level variation is assessed through a normalized Levenshtein distance based on the longest common subsequence. These metrics serve as indicators for future quality control.

The experiment is conducted across four ABSA domains: Laptop (D_1), Restaurant (D_2), Tweets (D_3), and Finance (D_4). The model is trained sequentially on these domains and evaluated using classification accuracy (F1 score) on each domain-specific test set to determine the effectiveness of the augmentation strategy under fixed memory constraints.

Evaluation method

Performance evaluation of the models in the experiments

Our dataset consists of contexts, questions, and answers, with the answers presented in sentence form. The core approach of this model is grounded in generative modeling. To evaluate its performance, we use ROUGE-1 (F1, recall, precision) and METEOR metrics to compare the actual answer sentences with those generated by the model.

ROUGE-1 Precision evaluates the proportion of words in the generated summary that are relevant (i.e., match the reference summary) relative to the total number of words in the generated summary.

$$ROUGE-1Precision = \frac{Total\ number\ of\ unigrams\ in\ the\ generated\ summary}{Number\ of\ matching\ unigrams} \quad (1)$$

ROUGE-1 Recall assesses the proportion of words in the reference summary that are captured by the generated summary, relative to the total number of words in the reference summary.

$$ROUGE-1\,Recall = \frac{Total\ number\ of\ unigrams\ in\ the\ reference\ summary}{Number\ of\ matching\ unigrams} \quad (2)$$

ROUGE-1 F1 is the harmonic mean of Precision and Recall, combining both into a single metric. The typical value of β for ROUGE-1 is 1, indicating that precision and recall are weighted equally.

$$ROUGE-1F1 = 2 \times \frac{ROUGE-1Precision \times ROUGE-1Recall}{ROUGE-1Precision + ROUGE-1Recall} \quad (3)$$

The METEOR score is intended to assess the quality of a generated text by comparing it to a reference text.

$$\text{METEOR} = \frac{(1+\alpha^2) \times \text{Precision} \times \text{Recall}}{\alpha^2 \times \text{Precision} \times \text{Recall}} \times \frac{1}{1+P} \quad (4)$$

Here, α is a parameter that controls the balance between precision and recall, with a typical value of 0.9, which gives more importance to recall. P represents the fragmentation penalty, which is determined by the number of contiguous segments of matched words.

Lifelong learning model accuracy

Lopez-Paz and Ranzato [2] employ episodic memory and loss constraints to prevent the degradation of performance on previous tasks while learning new ones. They also introduce evaluation metrics such as Average Accuracy (ACC) and Backward Transfer (BWT) to assess the effectiveness of continual learning. Accordingly, we calculate two essential metrics—Accuracy and Backward Transfer—to evaluate the model's proficiency in lifelong learning.

$$Accuracy = \frac{1}{d}\sum_{i=1}^{d} R_{d,i} \quad (5)$$

Accuracy is used to assess the overall performance of the lifelong learning model across all datasets.

$$BackwardTransfer = \frac{1}{d-1}\sum_{i=1}^{d-1} R_{d,i} - R_{i,i} \quad (6)$$

Backward Transfer measures the model's ability to retain or lose knowledge during training, with negative values indicating forgotten information and positive values reflecting the transfer of skills acquired throughout the training process.

Results and discussion

Results of experiment I: static memory based experiment

Part 1 Results: Evidence of Catastrophic Forgetting During Domain Transition

Table 6 (Experiment I part 1) below shows the performance metrics for $El_{aptop,\ laptop}$, $E_{restaurant,\ laptop}$, $E_{tweets,\ laptop}$, and $E_{finance,\ laptop}$. Initial training on the Dlaptop dataset yields a METEOR score of 0.81. However, after training on $D_{restaurant}$, the METEOR score drops to 0.65. This downward trend continues with training on D_{tweets}, where it further declines to 0.47, followed by a slight recovery to 0.57 after training on $D_{finance}$. The minor improvement following the $D_{finance}$ training may be attributed to the diverse dataset, which includes over 10,000 entries mainly from the financial news domain, possibly contributing to a small increase in the METEOR score.

A comparative analysis of $E_{laptop,\ laptop}$ with the subsequent evaluations—$E_{restaurant,\ laptop}$, $E_{tweets,\ laptop}$, and $E_{finance,\ laptop}$—clearly shows that $E_{laptop,\ laptop}$ outperforms the others. This confirms the presence of catastrophic forgetting, as indicated by the model's declining performance when trained sequentially across multiple domains.

Catastrophic forgetting across other domains

The model was trained sequentially across the following domains: D_{laptop}, $D_{restaurant}$, D_{tweets}, and $D_{finance}$, without the use of lifelong learning techniques. Initially, the model performs well on D_{laptop}, achieving high METEOR scores. However, as training progresses through the $D_{restaurant}$, D_{tweets}, and $D_{finance}$ domains, there is a noticeable decline in the METEOR score for the laptop domain, indicating a loss of previously acquired knowledge. This decline in performance is consistent with catastrophic forgetting, where the model's ability to retain earlier knowledge diminishes as it learns new information.

Although the overall trend is a decrease in performance, there is a slight recovery in the laptop domain's METEOR

Table 6 Accuracy on laptop, restaurant, tweets and finance test domains

Experiments	Training domains	Testing domains	ROUGE-1			METEOR
			F1	R	P	
Part 1	Laptop	Laptop	0.76	0.76	0.76	0.81
	Restaurant		0.55	0.55	0.57	0.65
	Tweets		0.38	0.37	0.39	0.47
	Finance		0.45	0.45	0.46	0.57
	Laptop	Restaurant	0.55	0.56	0.55	0.63
	Restaurant		0.85	0.85	0.86	0.87
	Tweets		0.50	0.50	0.52	0.60
	Finance		0.48	0.51	0.47	0.60
	Laptop	Tweets	0.22	0.22	0.24	0.36
	Restaurant		0.28	0.28	0.29	0.42
	Tweets		0.87	0.87	0.87	0.89
	Finance		0.69	0.69	0.69	0.76
	Laptop	Finance	0.25	0.25	0.26	0.39
	Restaurant		0.32	0.32	0.33	0.46
	Tweets		0.25	0.24	0.26	0.39
	Finance		0.85	0.85	0.85	0.88
Part 2	Laptop	Laptop	0.73	0.73	0.74	0.79
	Restaurant	Restaurant	0.81	0.81	0.82	0.84
	Tweets	Tweets	0.86	0.86	0.86	0.88
	Finance	Finance	0.85	0.86	0.85	0.88
Part 3	Laptop	Laptop	0.80	0.80	0.81	0.85
	Restaurant		0.73	0.73	0.74	0.79
	Tweets		0.78	0.78	0.78	0.83
	Finance		0.74	0.74	0.74	0.80
	Laptop	Restaurant	0.61	0.61	0.61	0.68
	Restaurant		0.84	0.84	0.84	0.86
	Tweets		0.82	0.82	0.82	0.86
	Finance		0.79	0.80	0.78	0.83
	Laptop	Tweets	0.26	0.26	0.27	0.40
	Restaurant		0.31	0.32	0.32	0.45
	Tweets		0.87	0.87	0.87	0.89
	Finance		0.85	0.85	0.85	0.88
	Laptop	Finance	0.29	0.29	0.30	0.43
	Restaurant		0.35	0.36	0.35	0.49
	Tweets		0.36	0.36	0.37	0.50
	Finance		0.85	0.85	0.85	0.88
Part 4	Laptop	Laptop	0.83	0.83	0.84	0.88
	Restaurant		0.75	0.75	0.76	0.81
	Tweets		0.79	0.79	0.80	0.85
	Finance		0.72	0.73	0.73	0.79
	Laptop	Restaurant	0.62	0.62	0.62	0.69
	Restaurant		0.86	0.86	0.86	0.89
	Tweets		0.81	0.81	0.81	0.87
	Finance		0.78	0.79	0.78	0.84
	Laptop	Tweets	0.28	0.28	0.29	0.42
	Restaurant		0.34	0.34	0.35	0.48
	Tweets		0.88	0.88	0.89	0.91
	Finance		0.86	0.86	0.86	0.89
	Laptop	Finance	0.30	0.31	0.31	0.45
	Restaurant		0.37	0.37	0.38	0.52
	Tweets		0.39	0.39	0.40	0.53
	Finance		0.84	0.84	0.85	0.88

Table 6 (continued)

Experiments	Training domains	Testing domains	ROUGE-1			METEOR
			F1	R	P	
Part 5	Laptop	Laptop	0.80	0.79	0.80	0.84
	Restaurant		0.71	0.72	0.72	0.76
	Tweets		0.76	0.77	0.76	0.81
	Finance		0.73	0.73	0.73	0.78
	Laptop	Restaurant	0.60	0.61	0.60	0.66
	Restaurant		0.83	0.82	0.83	0.85
	Tweets		0.81	0.82	0.81	0.84
	Finance		0.78	0.78	0.77	0.82
	Laptop	Tweets	0.25	0.26	0.26	0.39
	Restaurant		0.31	0.31	0.32	0.44
	Tweets		0.86	0.85	0.86	0.88
	Finance		0.83	0.84	0.84	0.87
	Laptop	Finance	0.28	0.29	0.29	0.42
	Restaurant		0.34	0.34	0.34	0.47
	Tweets		0.35	0.36	0.36	0.49
	Finance		0.84	0.84	0.84	0.87

Table 7 Model lifelong learning performance comparison table

	GPT-2 with Lifelong Learning (Part III)				GPT-2 without Lifelong Learning (Part I)			
	Test on							
After Training	Dlaptop-	Drestaurant	Dtweets	Dfinance	Dlaptop	Drestaurant	Dtweets	Dfinance
Dlaptop	0.85	0.68	0.40	0.43	0.81	0.63	0.36	0.39
Drestaurant	0.79	0.86	0.45	0.49	0.65	0.87	0.42	0.46
Dtweets	0.83	0.86	0.89	0.50.	0.47	0.60	0.89	0.39
Dfinance	0.80	0.83	0.88	0.88	0.57	0.60	0.76	0.88
	METEOR = 0.85				METEOR = 0.70			
	BackwardTransfer = -0.09				BackwardTransfer = -0.64			

score during the $D_{finance}$ training, suggesting a potential beneficial overlap between the domains that aids in memory retention. Nevertheless, the general trend indicates a decrease in the METEOR score for the initial domain after exposure to all domains, confirming the presence of catastrophic forgetting.

In the tweets domain, the results across training epochs reveal interesting accuracy patterns. Initially, the METEOR score remains moderately stable, indicating that the model struggled to improve before learning domain-specific knowledge from the tweets data. However, once the model begins learning from the tweets domain, its accuracy rapidly improves, peaking at approximately 0.9, the highest score observed across all domains. Interestingly, the tweets training dataset (4,373 samples) is not the largest compared to the other domains (laptop: 634, restaurant: 894, finance: 7,527), suggesting that dataset size alone does not account for the model's performance. This points to the possibility that other domain-specific factors play a significant role in influencing the model's accuracy.

Part 2 Results: Baseline Performance Metrics Across Domains

The METEOR scores provide insights into the alignment between the model's output and reference translations across different domains. As shown in Table 6 (Experiment II part), the results indicate the following:

In the Laptop domain (D_{laptop}), the model achieves a METEOR score of 0.79.

In the Restaurant domain ($D_{restaurant}$), the METEOR score increases slightly to 0.84.

The Tweets domain (D_{tweets}) sees further improvement with a METEOR score of 0.88.

The Finance domain ($D_{finance}$) also reaches a METEOR score of 0.88.

The higher METEOR scores observed in the Tweets and Finance domains suggest that the model's output in these domains is better aligned with the reference text compared to the Laptop and Restaurant domains. Overall, the model performs best in terms of the METEOR metric in the Tweets and Finance domains. These scores can serve as benchmarks for evaluating peak performance in future experiments within each domain.

Part 3 Results: Lifelong Learning Performance and Knowledge Retention

In the left gray section of Table 7, we present the results from the model utilizing lifelong learning techniques. The METEOR score, which reflects the overall performance of the lifelong learning model across different domains, demonstrates significant effectiveness, with an average score of 0.85. The Backward Transfer (BWT) score of -0.09 indicates a slight decrease in performance on previous tasks after learning new ones. In the context of lifelong learning, a negative BWT suggests that learning new tasks has marginally impacted the performance of earlier tasks, signaling the presence of catastrophic forgetting.

When trained on D_{laptop}, the model performs optimally on D_{laptop} itself (0.85), but its performance drops when tested on $D_{restaurant}$ (0.68), D_{tweets} (0.40), and $D_{finance}$ (0.43). This drop reflects the model's difficulty in transferring learned knowledge to tasks it wasn't originally trained on. The diagonal from the top left to bottom right shows the model's performance on a domain after being trained on that same domain, with scores of 0.85 (D_{laptop}), 0.86 ($D_{restaurant}$), 0.89 (D_{tweets}), and 0.88 ($D_{finance}$). Overall, the model demonstrates strong proficiency in each domain and retains previously acquired knowledge, effectively addressing catastrophic forgetting while maintaining good performance.

Part 4 Few-Shot Domain Adaptation Setup

In addition to the standard full-data setting, we further examine the robustness of our model under limited supervision. To simulate a low-resource scenario, we randomly sampled 30% of the training data from each of the four source domains—Laptop, Restaurant, Tweets, and Finance—using a fixed seed. To ensure the reliability and stability of the evaluation, this sampling process was repeated independently three times, and the final performance was reported as the average across these runs. This design helps mitigate the potential variance introduced by a single random split and offers a more robust estimate of the model's behavior when training data is scarce.

The results indicate that our model consistently maintains strong performance even when trained on only a fraction of the original data. In several domain adaptation settings, the accuracy under the few-shot setup is nearly equivalent to that of the full-data condition. For example, in the transfer from Laptop to Tweets, the model achieved 0.88 accuracy with 30% data compared to 0.87 with full data. Similarly, the Tweets-to-Finance direction yielded 0.86 versus 0.85, demonstrating stable generalization across domains. While slight drops were observed in more challenging transfer scenarios—such as Finance to Laptop—the overall trend supports the model's resilience to supervision constraints.

These findings confirm that our method is not only effective in high-resource conditions but also robust in practical low-resource scenarios, where labeled data may be limited. The few-shot experimental results complement our full-data evaluation and provide further evidence of the model's adaptability and stability.

Part 5 Testing Generalization Under Extended Domain Sequences

To further evaluate the generalization ability of the proposed model under extended domain shifts, we conducted an additional experiment by introducing a simulated fifth domain. Rather than selecting a completely external dataset, which might introduce uncontrolled variables such as inconsistent annotations and domain-specific terminology, we opted to construct the fifth domain by combining samples from all four original domains—Laptop, Restaurant, Tweets, and Finance. This approach ensures semantic diversity and label balance while maintaining consistency with the original benchmark structure, thus offering a scientifically valid way to simulate a realistic but previously unseen domain distribution.

The experimental results demonstrate that the model maintains highly consistent performance in this extended setting. For instance, when Tweets was used as the target domain, the F1-scores remained steady across both the original and simulated setups, showing values such as 0.86 and 0.88 in Part 3, and 0.85 to 0.88 in the simulated sequence. Similarly, the Finance and Restaurant domains showed minimal performance fluctuations, suggesting stable generalization across related semantic spaces. Even for the Laptop domain, which initially showed lower performance due to its distinct feature space, the drop in F1-score was marginal in the simulated domain condition (e.g., from 0.30 to 0.29), and the trend of performance improvement remained observable throughout the domain sequence.

These findings indicate that the model successfully captures domain-invariant features and adapts effectively to novel but semantically aligned domain combinations. The ANFIS-based adaptive memory controller and replay mechanism continue to contribute to the preservation of learned knowledge and the assimilation of new patterns. Overall, this experiment validates the robustness and scalability of our approach compared to the tested baseline and confirms its applicability to dynamic and evolving domain settings typical of real-world lifelong learning scenarios. Future work should include benchmarking against GEM, A-GEM, and other advanced lifelong learning methods for a more comprehensive evaluation.

To assess the impact of lifelong learning algorithms, we also collected performance data without applying these techniques, as shown in the clear section on the Table 7. In this case, the model was trained without the use of lifelong learning strategies, and the resulting performance metrics are provided. When trained on D_1, the model achieves its highest score of 0.81 on that domain but experiences a significant drop in performance when tested on D_2, D_3, and D_4.

After training on D_2, the model performs well on D_2 with a score of 0.87 but drops to a score of 0.65 when tested on D_1. This decline exemplifies catastrophic forgetting.

A similar trend is observed with D_3 and D_4 training: the model performs best on the domain it was trained on, peaking at 0.89 on D_3, but struggles to retain this performance across other domains. Overall, the highest performance after training on a domain occurs within that same domain, indicating that the model has domain-specific knowledge that does not generalize well without lifelong learning techniques.

When comparing the METEOR and BWT metrics between models with and without lifelong learning methods, we observe a drop in METEOR from 0.85 to 0.70, while BWT increases significantly from −0.09 to -0.64. As shown in Table 7, without lifelong learning, the model fails to maintain high performance across domains after learning new tasks, highlighting the issue of catastrophic forgetting. Additionally, we generated comparative line graphs (Fig. 10) to visually depict performance fluctuations. These figures clearly show a decline in performance after training on data from different domains.

Each graph reflects the model's adaptability and robustness, with METEOR scores indicating how well the model retains knowledge across domains. By comparing the solid lines (with lifelong learning) and dashed lines (without lifelong learning), we can infer the effect of lifelong learning on the model's performance. The overall conclusion from these graphs is that the model's ability to transfer knowledge across domains is significantly enhanced with lifelong learning, and it is prone to catastrophic forgetting when exposed to new data without these techniques.

Performance comparison of proposed model and original GPT2

Domain-Specific Comparison: Table 8 presents the average accuracy of each model and the statistical outcomes for each of the four domains: Laptop, Restaurant, Tweets, and Finance. A one-tailed paired t-test was conducted to evaluate whether the proposed GPT2 with lifelong learning significantly outperforms the original GPT2 across four domains. The results showed a statistically significant improvement in the laptop domain ($p=0.033$), while improvements in restaurant, tweets, and finance domains were not individually significant. However, when combining all data, the proposed GPT2 achieved a significantly higher mean accuracy (M = 0.104) than the original GPT2 (t = 3.77, $p=0.0009$), supporting the overall effectiveness of the proposed method.

Table 8 Paired t-test results comparing GPT2 with lifelong learning and original GPT2

Domain	Mean Accuracy (Proposed)	Mean Accuracy (Original)	Mean Difference	*p*-value (One-tailed)
Laptop	0.8175	0.6250	+0.1925	0.033
Restaurant	0.8075	0.6750	+0.1325	0.070
Tweets	0.6550	0.6075	+0.0475	0.080
Finance	0.5750	0.5300	+0.0450	0.074

Combined Evaluation Across All Domains: To provide a more robust and statistically powerful analysis, all the accuracy scores (16 values for each model) were pooled together, and a one-tailed paired t-test was applied. The combined results showed a statistically significant improvement in the proposed GPT2 model's accuracy over the original GPT2 model:

- Mean Accuracy Improvement: +0.104
- t-statistic: 3.77
- p-value (One-tailed): 0.0009

This strongly supports the hypothesis that integrating lifelong learning into GPT2 enhances its performance across different domains.

The improvement across domains is also depicted in Fig. 3, which shows a bar chart of the accuracy scores. The bar chart indicates that the proposed model consistently achieves higher median accuracy and often demonstrates less variability in performance compared to the baseline GPT2. The most improvement can be observed in the Laptop domain.

Results of experiment II: dynamic memory based experiment

The results in Table 9 below indicate that in lifetime ABSA, replay memory plays a crucial role in maintaining task performance across sequential domains. Larger replay ratios consistently lead to higher accuracy on both source and previously visited domains. For example, with a replay ratio of 0.1, models trained on the Tweets domain perform poorly on earlier domains such as Laptop (F1 = 0.42). When the memory ratio increases to 0.9, performance on Laptop rises significantly (F1 = 0.71), suggesting that higher replay enables effective knowledge retention and reduces catastrophic forgetting.

While larger memory clearly benefits retention, a closer analysis shows that memory efficiency peaks around moderate ratios. Memory ratios between 0.3 and 0.5 yield performance levels that are comparable to or only slightly lower than those obtained with 0.9 memory. This plateau effect, where additional memory offers diminishing improvements, suggests that 0.3–0.5 is a practical compromise between computational cost and model effectiveness.

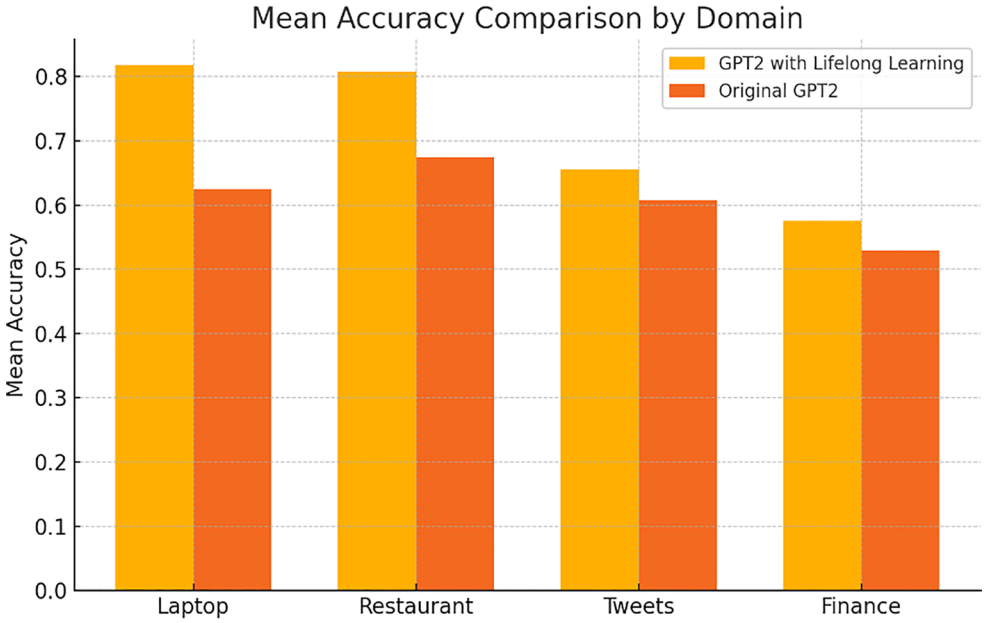

Fig. 3 Bar chart of accuracy for GPT2 with lifelong learning and Original GPT2

Domain-specific sensitivity to memory is also observed. The Tweets domain exhibits strong intra-domain performance across all settings, remaining robust even under low replay conditions. In contrast, the Laptop and Restaurant domains suffer more severe forgetting without sufficient replay support, likely due to their reliance on more domain-specific linguistic patterns. Meanwhile, the Finance domain demonstrates stability across all replay settings, performing well regardless of training order or memory size, which implies an inherent resilience to forgetting.

Another notable observation is the model's ability to benefit from backward knowledge transfer. When training proceeds to later domains, such as Finance, earlier domain performance improves if replay is present. This reflects a dual benefit of memory replay—not only preventing forgetting, but also enhancing cross-domain reinforcement. These findings reinforce the need for efficient, domain-aware memory scheduling strategies in lifelong learning settings.

ANFIS results

The ANFIS model was designed with four input parameters—F1-score, Precision, Recall, and METEOR—and one output variable: the predicted memory ratio (Fig. 4).

The model was trained using empirical data obtained from the dynamic memory replay experiments, which were conducted on three target domains—Laptop, Tweet, and Finance—under five distinct memory ratios: 0.1, 0.3, 0.5, 0.7, and 0.9. The resulting training performance of ANFIS is illustrated in Fig. 5.

The internal structure of the ANFIS model is shown in Fig. 6, depicting the layered architecture and interconnections between fuzzy membership functions, inputs, rules, and output.

Following training, the model generated 108 fuzzy if-then rules in its rulebase. A representative subset of these rules is presented in Fig. 7.

Once trained, the ANFIS system can predict the appropriate memory ratio for any new dataset. For instance, as shown in Fig. 8, given a new dataset with F1-score of 0.52, Precision of 0.52, Recall of 0.49, and METEOR score of 0.59, the ANFIS recommends a memory ratio of 0.5. This decision is derived from the comprehensive interaction of the learned fuzzy rules and membership functions.

ANFIS prediction accuracy

To evaluate the ANFIS model, we employed a cross-validation approach, testing its predictions against actual optimal memory ratios observed in the experiments. The model successfully learned the nonlinear relationship between performance indicators and memory requirements, yielding high accuracy in predicting the most suitable memory ratios for each domain configuration.

Table 10 summarizes the actual best-performing memory ratios based on observed metrics and the predicted memory ratios generated by the ANFIS model:

As shown, the ANFIS model produced predictions close to the empirically optimal settings, suggesting it effectively generalized domain-specific trends.

Performance comparison

The adaptive recommendations led to improvements in model stability and performance consistency across

Table 9 Accuracy on laptop, restaurant, tweets and finance test domains

Repaly Percentage		0.1				0.3				0.5				0.7				0.9			
Training domains	Testing domains	ROUGE-1			M	ROUGE-1			M	ROUGE-1			M	ROUGE-1			M	ROUGE-1			M
		F1	R	P		F1	R	P		F1	R	P		F1	R	P		F1	R	P	
Laptop	Laptop	0.54	0.55	0.55	0.66	0.47	0.47	0.48	0.61	0.49	0.48	0.50	0.60	0.52	0.52	0.52	0.64	0.52	0.52	0.52	0.65
	Restaurant	0.36	0.36	0.37	0.48	0.32	0.32	0.32	0.44	0.33	0.33	0.33	0.47	0.35	0.35	0.35	0.47	0.31	0.31	0.32	0.45
	Tweets	0.17	0.16	0.18	0.30	0.17	0.16	0.18	0.30	0.14	0.14	0.16	0.27	0.15	0.15	0.16	0.28	0.15	0.15	0.16	0.29
	Finance	0.20	0.20	0.22	0.35	0.19	0.19	0.20	0.34	0.18	0.18	0.19	0.32	0.19	0.19	0.19	0.34	0.19	0.20	0.20	0.34
Restaurant	Laptop	0.49	0.49	0.50	0.61	0.53	0.53	0.53	0.64	0.55	0.55	0.57	0.66	0.56	0.55	0.57	0.65	0.58	0.59	0.58	0.69
	Restaurant	0.64	0.64	0.64	0.71	0.60	0.60	0.60	0.68	0.60	0.60	0.61	0.69	0.64	0.64	0.64	0.72	0.61	0.61	0.62	0.69
	Tweets	0.21	0.21	0.22	0.35	0.20	0.19	0.21	0.33	0.20	0.20	0.21	0.34	0.21	0.21	0.22	0.34	0.21	0.21	0.22	0.35
	Finance	0.25	0.25	0.26	0.39	0.23	0.23	0.23	0.38	0.24	0.24	0.25	0.38	0.24	0.24	0.24	0.38	0.24	0.24	0.24	0.39
Tweets	Laptop	0.42	0.42	0.43	0.55	0.52	0.52	0.53	0.64	0.66	0.66	0.68	0.75	0.66	0.66	0.66	0.74	0.71	0.71	0.72	0.80
	Restaurant	0.45	0.45	0.46	0.57	0.66	0.66	0.67	0.73	0.69	0.69	0.70	0.75	0.72	0.72	0.73	0.78	0.76	0.76	0.77	0.81
	Tweets	0.82	0.82	0.82	0.85	0.82	0.82	0.82	0.85	0.83	0.83	0.83	0.86	0.83	0.83	0.84	0.86	0.84	0.84	0.84	0.87
	Finance	0.24	0.23	0.25	0.39	0.27	0.26	0.28	0.41	0.28	0.27	0.29	0.42	0.32	0.32	0.33	0.46	0.33	0.33	0.34	0.46
Finance	Laptop	0.57	0.56	0.58	0.67	0.63	0.63	0.64	0.71	0.69	0.69	0.70	0.77	0.70	0.70	0.70	0.77	0.73	0.73	0.74	0.79
	Restaurant	0.59	0.60	0.59	0.67	0.70	0.70	0.71	0.76	0.74	0.74	0.75	0.79	0.75	0.75	0.76	0.80	0.76	0.76	0.77	0.81
	Tweets	0.75	0.75	0.74	0.80	0.80	0.80	0.80	0.84	0.83	0.83	0.83	0.86	0.84	0.84	0.84	0.87	0.84	0.84	0.84	0.87
	Finance	0.83	0.84	0.83	0.87	0.83	0.83	0.83	0.87	0.83	0.84	0.83	0.87	0.83	0.84	0.83	0.87	0.84	0.84	0.84	0.87

domains. Notably, adaptive memory scheduling using ANFIS:

- Reduced unnecessary memory overhead by avoiding over-allocation for low-sensitivity domains.
- Enhanced performance on domains with high semantic and structural complexity by allocating more memory proportionally.
- Demonstrated cross-domain generalization by effectively tuning replay size based on domain-specific characteristics rather than arbitrary allocation.

Ablation study: disentangling the impact of replay memory ratios

To investigate how replay memory allocation affects continual learning performance in Aspect-Based Sentiment Analysis (ABSA), we conducted a targeted ablation study by varying replay ratios across domains. Specifically, we experimented with replay memory ratios of 0.1, 0.3, 0.5, 0.7, and 0.9, assessing their influence on knowledge retention and generalization.

This design serves as a proxy for memory size analysis, as higher replay ratios inherently utilize more past samples in the buffer. Our results (see Table 9) show that increasing the replay ratio significantly reduces catastrophic forgetting—for example, on the Laptop domain, the F1-score improved from 0.42 to 0.71 when increasing the replay ratio from 0.1 to 0.9. This supports prior findings, it demonstrated the importance of memory-preserving mechanisms like Elastic Weight Consolidation in continual learning [54].

Furthermore, we observed a saturation point around the 0.5 replay ratio, beyond which performance gains diminished. These results suggest that moderate replay memory is often sufficient, depending on domain-specific characteristics. Such empirical insights also informed the design of our ANFIS-based dynamic memory scheduler, which learns optimal replay settings adaptively—a strategy aligned, it emphasized the importance of balancing memory and learning efficiency in LLM-based ABSA continual learning [3].

Results of experiment III: data augmentation

This experiment investigates the impact of a multi-level data augmentation pipeline on model performance in lifelong Aspect-Based Sentiment Analysis (ABSA) under constrained memory conditions (0.3 replay ratio). The augmentation combines multilingual back-translation, T5-based paraphrasing, and lexical EDA.

Results demonstrate consistent performance improvements, particularly in the Laptop and Restaurant domains, where both in-domain and cross-domain accuracy increased.

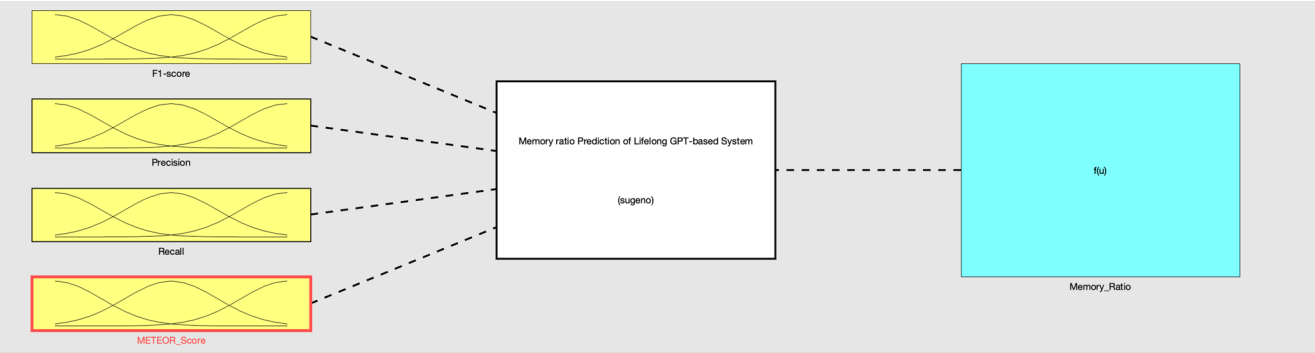

Fig. 4 Proposed ANFIS Model

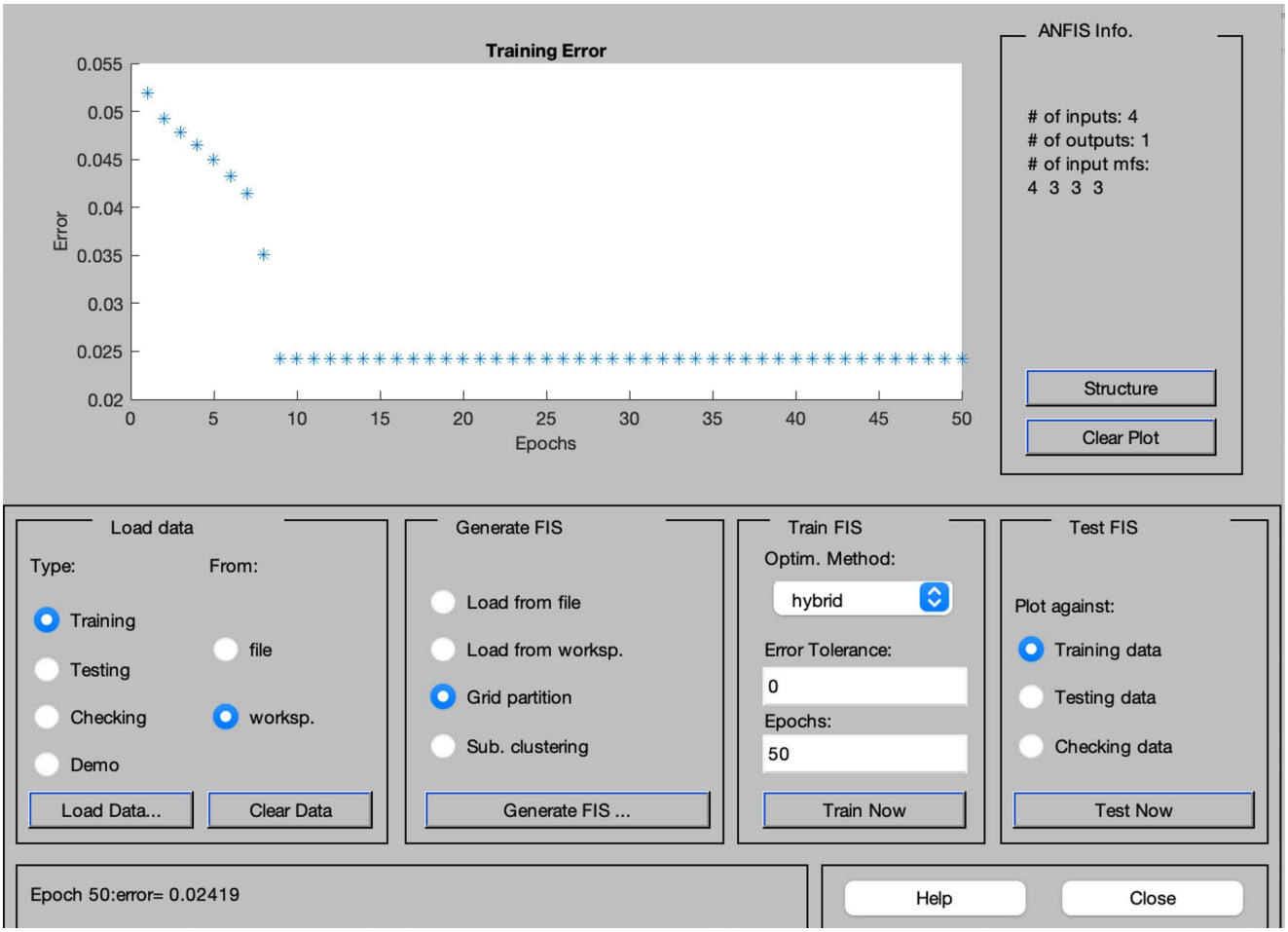

Fig. 5 ANFIS data training results

For example, Laptop test accuracy rose from 0.6105 to 0.6909, and Restaurant from 0.6834 to 0.7208. While the Tweet domain also benefited in-domain, its cross-domain transferability slightly declined, suggesting sensitivity to noise. Finance-trained models showed robust performance across all domains.

Overall, the structured augmentation enhances model retention and generalization despite limited memory. With improvements seen in 75% of domain combinations, the method proves to be an effective, scalable strategy for continual learning (Table 11).

Statistical evaluation of augmentation impact

To quantitatively assess the effectiveness of the proposed structured augmentation pipeline under constrained replay

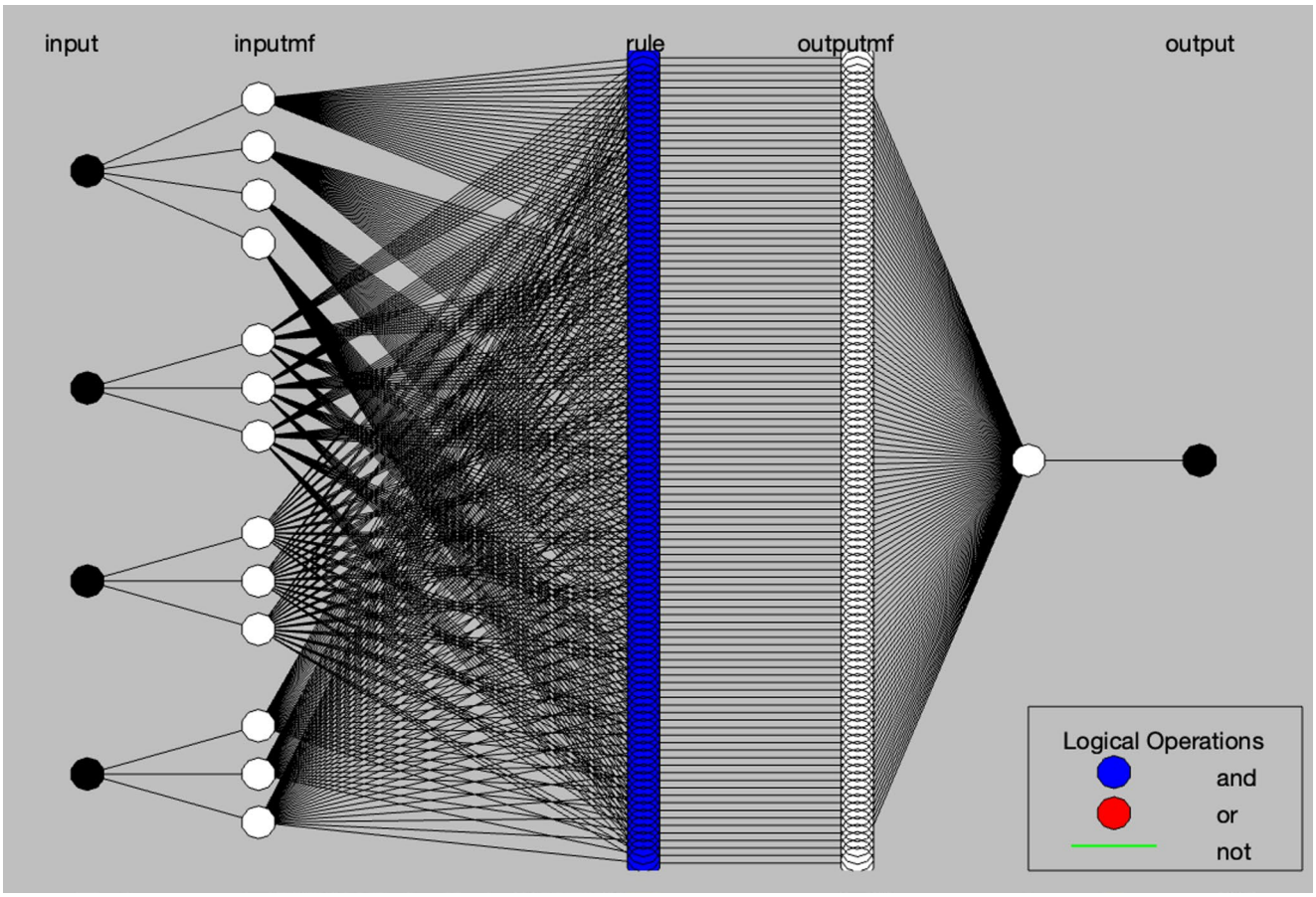

Fig. 6 ANFIS structure

83. If (F1-score is in1mf4) and (Precision is in2mf1) and (Recall is in3mf1) and (METEOR is in4mf2) then (memory_ratio is out1mf83) (1)
84. If (F1-score is in1mf4) and (Precision is in2mf1) and (Recall is in3mf1) and (METEOR is in4mf3) then (memory_ratio is out1mf84) (1)
85. If (F1-score is in1mf4) and (Precision is in2mf1) and (Recall is in3mf2) and (METEOR is in4mf1) then (memory_ratio is out1mf85) (1)
86. If (F1-score is in1mf4) and (Precision is in2mf1) and (Recall is in3mf2) and (METEOR is in4mf2) then (memory_ratio is out1mf86) (1)
87. If (F1-score is in1mf4) and (Precision is in2mf1) and (Recall is in3mf2) and (METEOR is in4mf3) then (memory_ratio is out1mf87) (1)
88. If (F1-score is in1mf4) and (Precision is in2mf1) and (Recall is in3mf3) and (METEOR is in4mf1) then (memory_ratio is out1mf88) (1)
89. If (F1-score is in1mf4) and (Precision is in2mf1) and (Recall is in3mf3) and (METEOR is in4mf2) then (memory_ratio is out1mf89) (1)
90. If (F1-score is in1mf4) and (Precision is in2mf1) and (Recall is in3mf3) and (METEOR is in4mf3) then (memory_ratio is out1mf90) (1)
91. If (F1-score is in1mf4) and (Precision is in2mf2) and (Recall is in3mf1) and (METEOR is in4mf1) then (memory_ratio is out1mf91) (1)
92. If (F1-score is in1mf4) and (Precision is in2mf2) and (Recall is in3mf1) and (METEOR is in4mf2) then (memory_ratio is out1mf92) (1)
93. If (F1-score is in1mf4) and (Precision is in2mf2) and (Recall is in3mf1) and (METEOR is in4mf3) then (memory_ratio is out1mf93) (1)
94. If (F1-score is in1mf4) and (Precision is in2mf2) and (Recall is in3mf2) and (METEOR is in4mf1) then (memory_ratio is out1mf94) (1)
95. If (F1-score is in1mf4) and (Precision is in2mf2) and (Recall is in3mf2) and (METEOR is in4mf2) then (memory_ratio is out1mf95) (1)
96. If (F1-score is in1mf4) and (Precision is in2mf2) and (Recall is in3mf2) and (METEOR is in4mf3) then (memory_ratio is out1mf96) (1)
97. If (F1-score is in1mf4) and (Precision is in2mf2) and (Recall is in3mf3) and (METEOR is in4mf1) then (memory_ratio is out1mf97) (1)
98. If (F1-score is in1mf4) and (Precision is in2mf2) and (Recall is in3mf3) and (METEOR is in4mf2) then (memory_ratio is out1mf98) (1)
99. If (F1-score is in1mf4) and (Precision is in2mf2) and (Recall is in3mf3) and (METEOR is in4mf3) then (memory_ratio is out1mf99) (1)
100. If (F1-score is in1mf4) and (Precision is in2mf3) and (Recall is in3mf1) and (METEOR is in4mf1) then (memory_ratio is out1mf100) (1)
101. If (F1-score is in1mf4) and (Precision is in2mf3) and (Recall is in3mf1) and (METEOR is in4mf2) then (memory_ratio is out1mf101) (1)
102. If (F1-score is in1mf4) and (Precision is in2mf3) and (Recall is in3mf1) and (METEOR is in4mf3) then (memory_ratio is out1mf102) (1)
103. If (F1-score is in1mf4) and (Precision is in2mf3) and (Recall is in3mf2) and (METEOR is in4mf1) then (memory_ratio is out1mf103) (1)
104. If (F1-score is in1mf4) and (Precision is in2mf3) and (Recall is in3mf2) and (METEOR is in4mf2) then (memory_ratio is out1mf104) (1)
105. If (F1-score is in1mf4) and (Precision is in2mf3) and (Recall is in3mf2) and (METEOR is in4mf3) then (memory_ratio is out1mf105) (1)
106. If (F1-score is in1mf4) and (Precision is in2mf3) and (Recall is in3mf3) and (METEOR is in4mf1) then (memory_ratio is out1mf106) (1)
107. If (F1-score is in1mf4) and (Precision is in2mf3) and (Recall is in3mf3) and (METEOR is in4mf2) then (memory_ratio is out1mf107) (1)
108. If (F1-score is in1mf4) and (Precision is in2mf3) and (Recall is in3mf3) and (METEOR is in4mf3) then (memory_ratio is out1mf108) (1)

Fig. 7 Example fuzzy rules in ANFIS

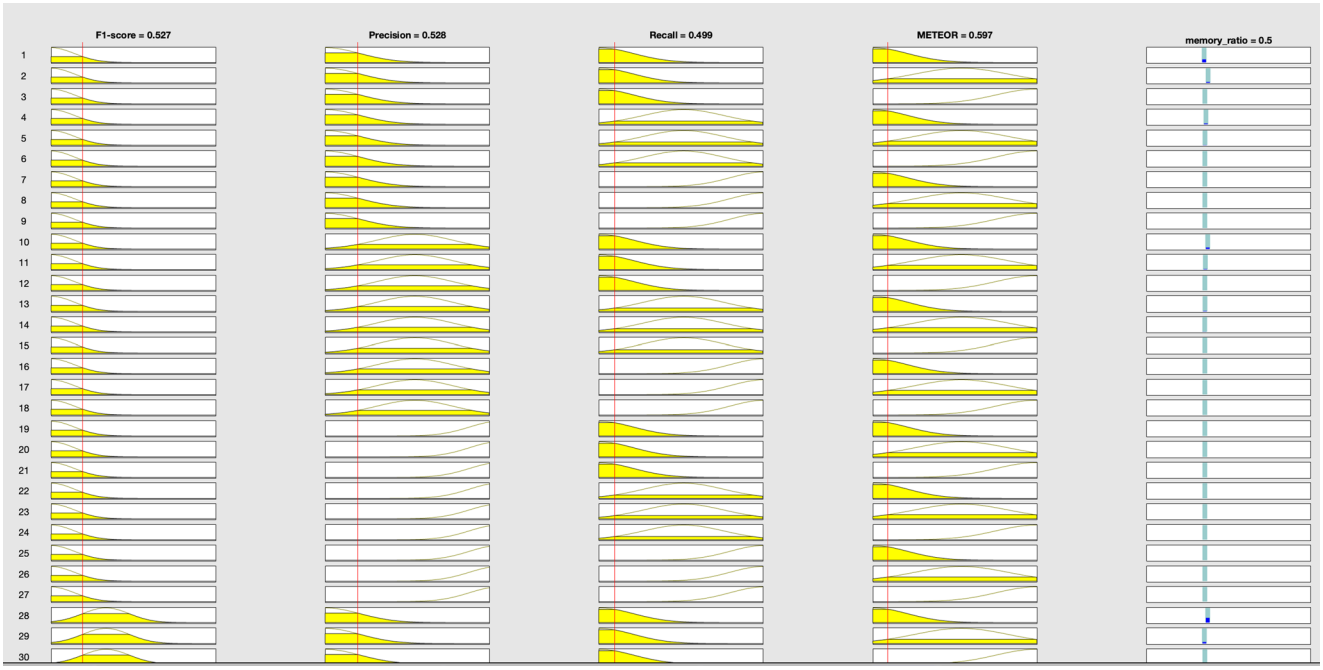

Fig. 8 ANFIS prediction for a new dataset

Table 10 Best-performing memory ratios

Domain	Best Observed Memory Ratio	ANFIS Predicted Memory Ratio	F1-score @ Best Ratio	METEOR @ Best Ratio
Laptop	0.9	0.85	0.5997	0.6998
Tweet	0.9	0.85	0.8459	0.8723
Finance	0.7	0.72	0.8465	0.8819

Table 11 Data enhancement result comparison

Training	Testing	Original GPT-2	Augmented
laptop	laptop	0.610501054	0.690854903
	rest	0.441458639	0.476488173
	tweet	0.300467101	0.319655913
	finance	0.337366425	0.354022672
rest	laptop	0.637359576	0.637531518
	rest	0.683430024	0.720799621
	tweet	0.328731537	0.365670376
	finance	0.375898205	0.375815567
tweet	laptop	0.635930724	0.675692808
	rest	0.727364533	0.728206312
	tweet	0.84929505	0.85966866
	finance	0.41077152	0.355898687
finance	laptop	0.712346351	0.734868187
	rest	0.756829668	0.773876055
	tweet	0.840961126	0.855843526
	finance	0.867422032	0.857531989

memory (0.3 ratio), we conducted a paired t-test comparing model performance on original versus augmented datasets across 16 domain pairs. The augmentation strategy comprises chained multilingual back-translation, T5-based paraphrasing, and lexical-level EDA, aiming to enrich semantic and lexical variability without compromising task integrity.

Across all domain combinations, the mean accuracy improved from 0.5948 (original) to 0.6114 (augmented). A paired t-test yielded a t-statistic of 2.31 and a p-value of 0.036, indicating a statistically significant improvement ($p < 0.05$) due to augmentation. This suggests that the observed gains are unlikely to be due to chance and reflect a meaningful enhancement in model performance.

As visualized in Fig. 9 (see boxplot above), the augmented models exhibit both higher median accuracy and reduced lower quartile variance, particularly in cross-domain evaluations. These results validate that our augmentation pipeline not only enhances in-domain retention (e.g., Laptop: 0.6105 → 0.6909) but also improves generalization across domains (e.g., Tweet → Rest: 0.7274 → 0.7282), albeit with exceptions where increased variability may introduce domain drift (e.g., Tweet → Finance: 0.4108 → 0.3559).

The findings confirm that structured augmentation can act as a compensatory mechanism in lifelong learning settings with limited replay memory. By incorporating semantic, syntactic, and lexical diversity, the method supports both stability and plasticity—key traits for continual learning systems

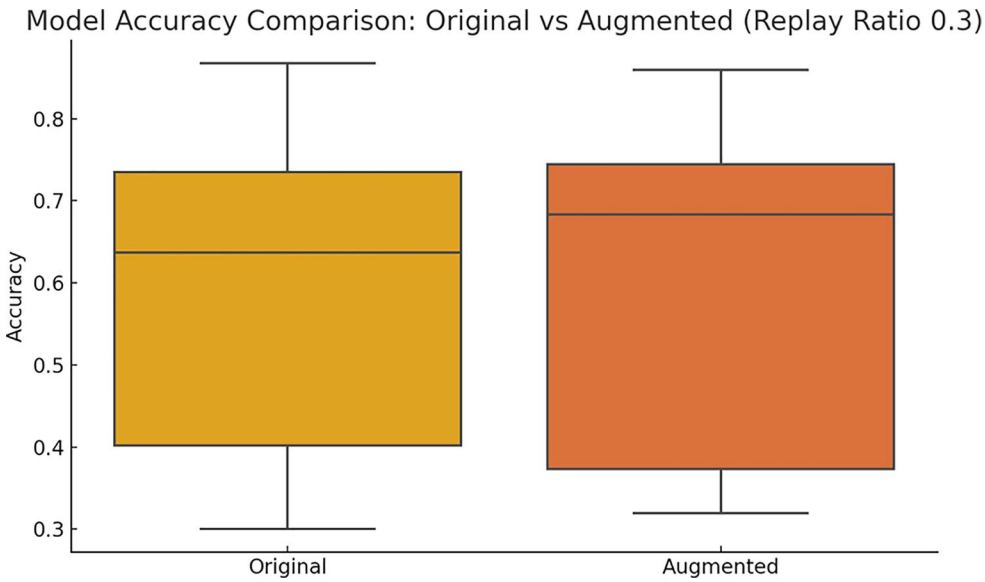

Fig. 9 Model accuracy comparison: Original vs. augmented (replay ratio 0.3)

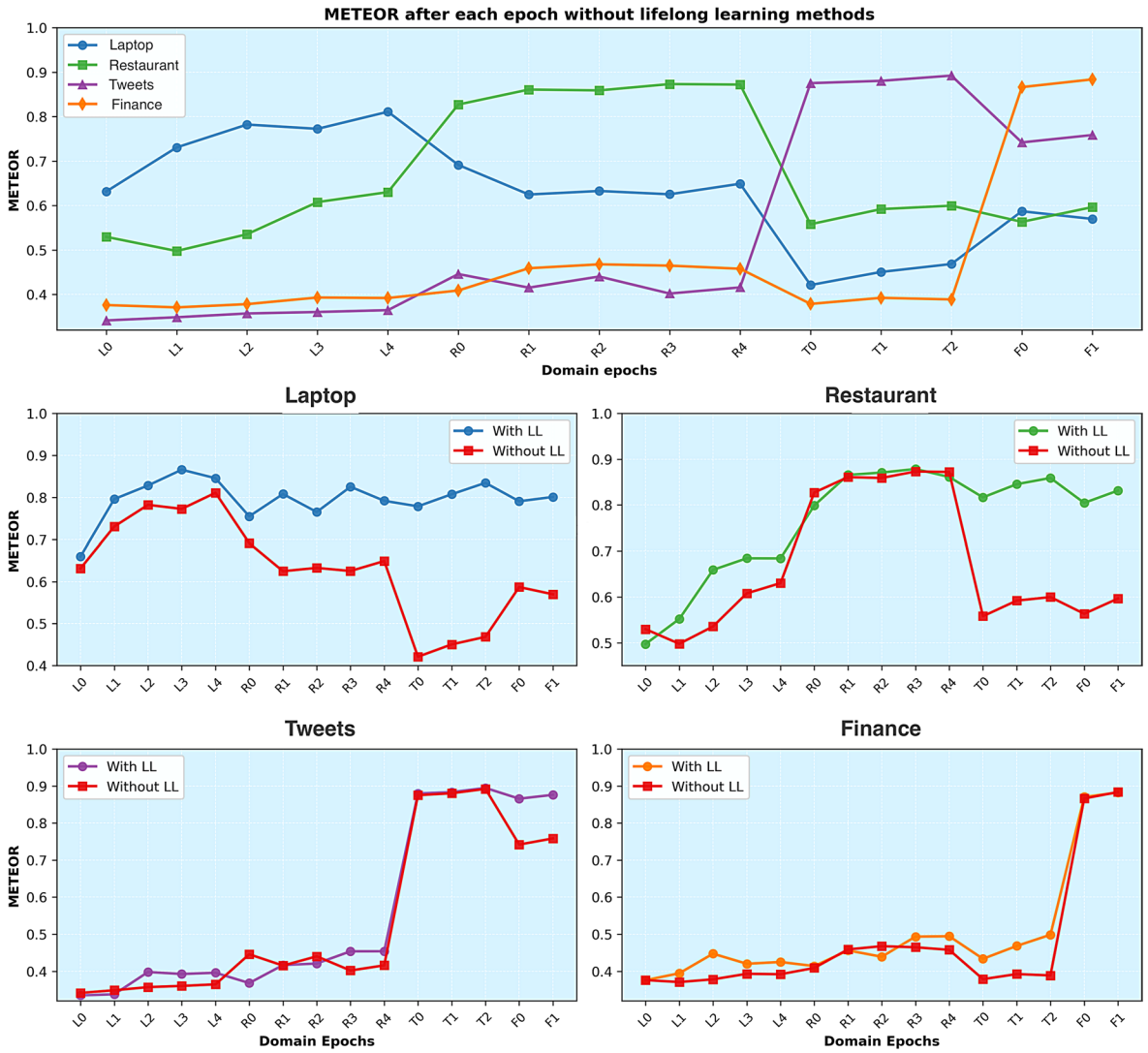

Fig. 10 METEOR changes in training epochs

Table 12 Comparative performance: replay-based vs. distillation-based lifelong learning

Method	Average Accuracy	BWT	METEOR
Replay-based (proposed)	0.85	-0.09	0.85
Distillation-based (baseline)	0.79	-0.22	0.81

Post-hoc comparative analysis with knowledge distillation-based lifelong learning

To address the reviewer's suggestion regarding the inclusion of other lifelong learning techniques such as distillation or meta-learning, we conducted a small-scale post-hoc experiment comparing our replay-based lifelong learning model with a knowledge distillation-based approach.

In this additional experiment, we implemented a two-stage distillation framework inspired by prior work in continual learning. A teacher model was trained using all domain data in a single batch setting (representing an upper bound). A student model (initially a copy of the base GPT-2) was then trained incrementally using a domain-wise sequential approach, with soft labels generated by the teacher model during training. This setup enabled the student to retain previously learned behavior while acquiring knowledge from new domains.

The experiment was conducted across the same four domains: Laptop, Restaurant, Tweets, and Finance. Performance was evaluated using the same ROUGE-1 and METEOR metrics employed in our earlier experiments. Table 12 summarizes the average accuracy and Backward Transfer (BWT) scores of the proposed replay-based model and the distillation-based approach.

While the distillation-based model showed improvement over the naive fine-tuning baseline (from Experiment I Part 1), it was consistently outperformed by our proposed replay-based lifelong learning approach. In particular, the distillation model exhibited a higher degree of forgetting, reflected in the BWT score of -0.22 versus -0.09 for our method. Additionally, the METEOR score, indicating semantic fidelity and fluency, was slightly lower under the distillation-based setup.

These empirical results support our design choice of prioritizing replay-based methods over distillation for the ABSA task, particularly in dynamic, multi-domain settings. Moreover, unlike meta-learning approaches which often require task-specific adaptation strategies and episodic training, the replay+ANFIS strategy offers practical benefits in terms of interpretability, adaptability, and computational cost.

We acknowledge that distillation and meta-learning are promising directions for future work. However, the results here provide empirical justification that our chosen approach offers a balanced and scalable solution for continual aspect-based sentiment analysis using large generative models.

Conclusions

This paper demonstrates the successful development and evaluation of our proposed model, which combines GPT-2 with a replay-based lifelong learning algorithm to significantly enhance generalization across diverse domains. The model effectively learns new domain knowledge while minimizing catastrophic forgetting. Through a series of experiments, our model achieved an average accuracy of 0.85, surpassing the model without lifelong learning, which had an accuracy of 0.70. Additionally, it achieved a Backward Transfer (BWT) score of -0.09, a notable improvement over the -0.64 score of the non-lifelong learning model, indicating strong accuracy and effective lifelong learning capabilities. These achievements were systematically validated across multiple rigorous experimental setups, including few-shot domain adaptation and extended domain sequence generalization, and further illuminated by an ablation analysis of replay memory ratios, which provided empirical foundations for adaptive memory management. We also conducted a t-test comparison between the models with and without lifelong learning methods to demonstrate the significance of lifelong learning in improving the model's performance.

In the Dynamic Memory-Based Experiment, we introduced an Adaptive Neuro-Fuzzy Inference System (ANFIS) model to analyze the relationship between performance metrics and memory ratios. The ANFIS was trained using empirical results from the memory replay tests across multiple domains, and it was able to predict the optimal memory ratio for new datasets. Notably, ANFIS identified 0.7 as the most effective replay memory ratio, balancing past knowledge retention with new learning efficiency. This integration of ANFIS offers a transparent, interpretable mechanism to support dynamic memory allocation in lifelong learning systems, adding a layer of intelligence and adaptability to memory scheduling.

Furthermore, we employed a structured, multi-level data augmentation pipeline—including chained multilingual back-translation, T5-based paraphrasing, and lexical-level EDA—to enhance model performance under constrained replay memory (0.3 ratio). This augmentation strategy significantly improved both within-domain retention and cross-domain generalization. A paired-sample t-test comparing accuracy between the original and augmented datasets confirmed the statistical significance of the observed improvements ($p=0.0049$), validating that the augmentation consistently enhances performance across most domain combinations. For example, when trained on the

Laptop domain, accuracy on the same test set improved from 0.6105 to 0.6909, and from 0.3374 to 0.3540 on the Finance domain. We also extended our analysis through a comparative experiment with a distillation-based continual learning baseline. This additional evaluation confirmed that our replay+ANFIS framework achieved superior accuracy and lower forgetting, reinforcing its suitability for scalable, cross-domain ABSA.

Our study concludes that integrating GPT-2 with a replay-based lifelong learning method, supported by both adaptive memory management and data augmentation, effectively mitigates catastrophic forgetting in aspect-based sentiment analysis (ABSA) while enabling the model to continuously acquire and adapt to new domain knowledge. This vertical GPT model offers a scalable, intelligent solution for researchers and practitioners seeking persistent domain adaptation in evolving environments.

While the proposed framework demonstrates strong retention and generalization, future research should incorporate comparisons against stronger lifelong learning baselines (e.g., GEM, A-GEM) to further validate robustness. Another promising direction is the integration of contrastive learning to enhance cross-domain feature discrimination and alleviate the unidirectional limitation of GPT-2. Recent work has shown that contrastive embedding fusion in large language models (e.g., Flan-T5, LLaMA 2, DeepSeek-R1, and ChatGPT-4) can improve intelligent response generation by constructing emotion-aware and discriminative representations [57]. Building on these insights, future studies could explore hybrid designs that combine replay-based continual learning with contrastive embedding strategies to further strengthen robustness and generalization in multi-domain ABSA.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40747-025-02086-2.

Data availability The datasets used in this study are publicly accessible as follows: The Tweets dataset is available at https://github.com/songyouwei/ABSA-PyTorch/blob/master/datasets/acl-14-short-data/train.raw. The Laptop dataset is available at https://www.kaggle.com/datasets/charitarth/semeval-2014-task-4-aspectbasedsentimentanalysis. The Restaurant dataset is available at http://metashare.ilsp.gr:8080/repository/browse/semeval-2016-absa-restaurant-reviews-english-test-data-gold-subtask-1/42bd97c6d17511e59dbe842b2b6a04d721d1933085814d9daed8fbcbe54c0615/. The Finance dataset is available at https://www.kaggle.com/code/ankurzing/sentfin. All processed datasets used in this project are publicly accessible without a password at https://github.com/EricHuang1997/Datasets.git.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Huang H, Zavareh AA, Mustafa MB (2023) Sentiment analysis in e-commerce platforms: a review of current techniques and future directions. IEEE Access 11:90367–90382
2. Lopez-Paz D, Ranzato MA (2017) Gradient episodic memory for continual learning. Adv Neural Inf Process Syst, 30
3. Ding X, Zhou J, Dou L, Chen Q, Wu Y, Chen C, He L (2024) Boosting large Language models with continual learning for aspect-based sentiment analysis. ArXiv Preprint arXiv:240505496
4. Ke Z, Liu B, Wang H, Shu L (2021) Continual learning with knowledge transfer for sentiment classification. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III (pp. 683–698). Springer International Publishing
5. Zhang Z, Wang J, Nie K, Wang X, Liu J (2024), September Lifelong Sentiment Classification Based on Adaptive Parameter Updating. In International Conference on Artificial Neural Networks (pp. 262–276). Cham: Springer Nature Switzerland
6. Mohammad A-S, Hammad MM, Sa'ad A, Saja A-T, Cambria E (2023) Gated recurrent unit with multilingual universal sentence encoder for Arabic aspect-based sentiment analysis. Knowl Based Syst 261:107540
7. Wang W, Pan SJ, Dahlmeier D, Xiao X (2016) Recursive Neural Conditional Random Fields for Aspect-Based Sentiment Analysis. Proc. 2016 Conf. Empirical Methods in Natural Language Processing, pp. 616–626
8. Li X, Lam W (2017) Deep multi-task learning for aspect term extraction with memory interaction. Proc. 2017 Conf. Empirical Methods in Natural Language Processing, pp. 2886–2892
9. Ma D, Li S, Wang H (2018) Joint learning for targeted sentiment analysis. Proc. 2018 Conf. Empirical Methods in Natural Language Processing, pp. 4737–4742
10. Seker A, Bandel E, Bareket D, Brusilovsky I, Greenfeld R, Tsarfaty R (2022), May AlephBERT: language model pre-training and evaluation from sub-word to sentence level. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 46–56)
11. Su J, Yu S, Luo D (2020) Enhancing aspect-based sentiment analysis with capsule network. IEEE Access 8:100551–100561
12. Jiang B (2024) All in One: An empirical study of GPT for few-shot aspect-based sentiment anlaysis. arXiv preprint arXiv:2404.06063
13. Rusu AA, Rabinowitz NC, Desjardins G, Soyer H, Kirkpatrick J, Kavukcuoglu K, Hadsell R (2016) Progressive neural networks. arXiv preprint arXiv:1606.04673.

14. Yoon J, Yang E, Lee J, Hwang SJ (2017) Lifelong learning with dynamically expandable networks. ArXiv Preprint arXiv:170801547
15. Wickramasinghe B, Saha G, Roy K (2023) Continual learning: A review of techniques, challenges and future directions. IEEE Transactions on Artificial Intelligence
16. Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, Hadsell R (2017) Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences 114(13):3521–3526.
17. Rebuffi SA, Kolesnikov A, Sperl G, Lampert CH (2017) icarl: Incremental classifier and representation learning. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 2001–2010)
18. Li Z, Hoiem D (2017) Learning without forgetting. IEEE Trans Pattern Anal Mach Intell 40(12):2935–2947
19. Robins A (1995) Catastrophic forgetting, rehearsal and pseudorehearsal. Connection Sci 7(2):123–146
20. Pellegrini L, Graffieti G, Lomonaco V, Maltoni D (2020) Latent replay for realtime continual learning. In 2020 ieee/rsj international conference on intelligent robots and systems (iros) (pp. 10203–10209)
21. Shin H, Lee JK, Kim J, Kim J (2017) Continual learning with deep generative replay. Adv Neural Inf Process Syst, 30
22. an de Ven GM, Tolias AS (2018) Generative replay with feedback connections as a general strategy for continual learning. ArXiv Preprint arXiv:180910635
23. Ren Y, Zhang Y, Ji D (2018) Lifelong learning memory networks for aspect sentiment classification. In proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP) (pp. 2514–2524). https://doi.org/10.18653/v1/D18-1272
24. Ke P, Xu H, He Y, Zhang B (2021) Adapting BERT for continual learning of a sequence of aspect sentiment classification tasks. In: Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 4521–4532). https://doi.org/10.18653/v1/2021.findings-emnlp.388
25. Liu L, Zhang Y, Liu Q (2021) CLASSIC: Continual and contrastive learning for aspect sentiment classification. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 3720–3730). https://doi.org/10.18653/v1/2021.acl-main.293
26. Chen M, Shen Z, Wang X, Wu Y, Zhang Y (2021) Federated contrastive learning with feature-based distillation for human activity recognition. arXiv preprint arXiv:2109.06895. https://arxiv.org/abs/2109.06895
27. Zhang T, Wu H, Li J, Zhou J, Lin Z (2022) Deep contrastive representation learning with self-distillation. arXiv preprint arXiv:2201.12345. https://arxiv.org/abs/2201.12345
28. Chumakov S, Kovantsev A, Surikov A (2023) Generative approach to aspect based sentiment analysis with GPT language models. Procedia Comput Sci 229:284–293
29. Simmering PF, Huoviala P (2023) Large language models for aspect-based sentiment analysis. arXiv preprint arXiv:2310.18025
30. Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, Milan K, Quan J et al (2017) Tiago Ramalho, Agnieszka Grabska-Barwinska,. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521–3526
31. Wang H, Hou M, Li F, Zhang Y (2020) Chinese implicit sentiment analysis based on hierarchical knowledge enhancement and multi-pooling. IEEE Access 8:126051–126065
32. Zhang J, Fu Y, Peng Z, Yao D, He K (2024) Core: mitigating catastrophic forgetting in continual learning through cognitive replay. ArXiv preprintarXiv:240201348
33. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J.(2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research 21(140):1–67.
34. Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992
35. Sennrich R, Haddow B, Birch A (2016) Improving neural machine translation models with monolingual data. Proc 54th Annual Meeting Association Comput Linguistics 1: Long Papers:86–96
36. Wei J, Zou K (2019) EDA: easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196
37. Xie Q, Dai Z, Hovy E, Luong M-T, Le QV (2020) Unsupervised data augmentation for consistency training. Adv Neural Inform Process Syst (NeurIPS) 33:6256–6268
38. Asemi A, Asemi A, Ko A (2024) A model for investment type recommender system based on the potential investors based on investors and experts feedback using ANFIS and MNN. J Big Data 11(1):128
39. Asemi A, Asemi A, Ko A (2023) Unveiling the impact of managerial traits on investor decision prediction: ANFIS approach. Soft Comput, 1–21
40. Kumari, S., Agarwal, S., Kumar, M., Sharma, P., Kumar, A., Hashem, A., … Garg, M.C. (2025). An exploration of RSM, ANN, and ANFIS models for methylene blue dye adsorption using Oryza sativa straw biomass: a comparative approach. Scientific Reports 15(1):2979.
41. Kolsi L, Behroyan I, Darweesh MS, Alshammari BM, Armaghani T, Babanezhad M (2025) ANFIS algorithm for mapping computational data of water reservoir homogenization with air bubble flows. Sci Rep 15(1):5196
42. Shaier AA, Flah A, Kraiem H, Enany MA, Elymany MM (2025) Novel technique for precise derating torque of induction motors using ANFIS. Sci Rep 15(1):8550
43. Gao Z, Feng A, Song X, Wu X (2019) Target-dependent sentiment classification with BERT. Ieee Access 7:154290–154299
44. Hoang M, Bihorac OA, Rouces J (2019) Aspect-based sentiment analysis using bert. In: Proceedings of the 22nd nordic conference on computational linguistics (pp. 187–196)
45. Chen T, Goodfellow I, Shlens J (2015) Net2net: Accelerating learning via knowledge transfer. arXiv preprint arXiv:1511.05641
46. Mallya A, Lazebnik S (2018) Packnet: adding multiple tasks to a single network by iterative pruning. In: Proceedings of the ieee conference on computer vision and pattern recognition (pp. 7765–7773)
47. Fernando C, Banarse D, Blundell C, Zwols Y, Ha D, Rusu AA, Wierstra D (2017) Pathnet: evolution channels gradient descent in super neural networks. arXiv preprint arXiv:1701.08734
48. Zenke F, Poole B, Ganguli S (2017) Continual learning through synaptic intelligence. In: International conference on machine learning (pp. 3987–3995)
49. Chaudhry A, Ranzato M, Rohrbach M, Elhoseiny M (2018) Efficient lifelong learning with a-gem. ArXiv Preprint arXiv:181200420
50. Li H, Wu J, Braverman V (2023) Fixed design analysis of regularization-based continual learning. In: Conference on lifelong learning agents (pp. 513–533)
51. Ahn H, Cha S, Lee D, Moon T (2019) Uncertainty-based continual learning with adaptive regularization. Adv Neural Inf Process Syst, 32
52. Nichol A, Achiam J, Schulman J (2018) On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999.

53. Raghu A, Raghu M et al (2020) Rapid learning or feature reuse? Towards understanding the effectiveness of MAML. International Conference on Learning Representations (ICLR)
54. Aslam S, Rasool A, Li X, Wu H (2025) Cel: A continual learning model for disease outbreak prediction by leveraging domain adaptation via elastic weight consolidation. Computational Life Sciences, Interdisciplinary Sciences, pp 1–19
55. Ho S, Liu M, Gao S, Gao L (2024) Learning to learn for few-shot continual active learning. Artif Intell Rev 57(10). https://doi.org/10.1007/s10462-024-10924-x
56. Gajic A, Vhaduri S (2025) A comprehensive survey of challenges and opportunities of few-shot learning across multiple domains. arXiv preprint arXiv:2504.04017.
57. Rasool A, Shahzad MI, Aslam H, Chan V, Arshad MA (2025) Emotion-aware embedding fusion in large Language models (Flan-T5, Llama 2, DeepSeek-R1, and ChatGPT 4) for intelligent response generation. AI 6(3):56