Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11, 448-466
2025
Publisher: Learning Gate
DOI: 10.55214/2576-8484.v9i11.10905

Learning without forgetting in NLP: Deep learning based transformer models for lifelong task incremental learning

Ambi Rachel Alex[1], Malliga Subramanian[2*], Jayanth J R[2], Keerthi Bala A T[2], Taniya Mukherjee[3]
[1]College of Engineering, Gulf university, Sanad 26489, Kingdom of Bahrain.
[2]Department of Computer Science and Engineering; Kongu Engineering College, India; mallinishanth72@gmail.com (M.S.).
[3]General Studies unit, Gulf University, Sanad 26489, Kingdom of Bahrain.

Abstract: This study investigates task incremental learning for developing transformer-based natural language processing (NLP) models capable of sequentially learning new tasks without forgetting previously acquired knowledge. It focuses on enabling models to learn from abusive comment detection to offensive comment detection while maintaining previous knowledge. Pre-trained transformer models such as ERT, ALBERT, RoBERTa, and DistilBERT were integrated with continual learning strategies, including rehearsal, pseudo-rehearsal, and regularization-based methods. Their effectiveness was evaluated on sequential abusive and offensive comment detection tasks using measures of classification accuracy and knowledge retention across multiple datasets. Among the evaluated models, BERT integrated with the Learning Without Forgetting (LwF) approach achieved the best trade-off between stability and plasticity, attaining accuracies of 91.06% for abusive and 98.8% for offensive comment detection. This demonstrates the ability of continual learning to avoid catastrophic forgetting in sequential NLP tasks. Task incremental learning enables transformer models to adapt to new linguistic challenges while retaining prior knowledge, supporting lifelong learning. The proposed framework provides valuable insights for building adaptable NLP systems applicable to content moderation, toxic speech detection, and other evolving language-based applications.

Keywords: *Abusive task, Learning without forgetting, Offensive dataset, Task incremental learning, Transformer models.*

1. Introduction

Social media platforms such as YouTube, Twitter, Facebook, etc., have become an integral part of modern communication, enabling the exchange of ideas and information [1]. The proliferation of these platforms has transformed the way individuals communicate. While these platforms provide numerous benefits, they also become channels for spreading abusive comments. Abusive language triggers cyberbullying, which targets individuals such as celebrities, politicians, etc., and groups such as specific countries, genders, religions, etc [2]. Such comments can have detrimental effects, including psychological impacts on individuals and disrupting community harmony [3]. Consequently, detecting and mitigating abusive comments has become a significant area of research within Natural Language Processing (NLP) and Artificial Intelligence (AI). Manually monitoring and detecting such social media content is impossible due to the large amount of data generated [4]. Traditional methods for detecting abusive language in text have relied on rule-based systems, classical machine learning, and deep learning algorithms [4, 5]. Common machine learning techniques for detecting offensive language include lexicon-based methods [6], n-grams [7-9] and ensemble-based learning [10, 11]. Some studies have applied and compared deep learning models, such as Long Short-Term Memory (LSTM) [4, 12], Bidirectional Long Short-Term Memory (BiLSTM) [13, 14], Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) [12] to detect abusive language. Recent advancements in deep learning, particularly the development of transformer-based models such as BERT, GPT, and their

History: Received: 2 September 2025; Revised: 10 October 2025; Accepted: 14 October 2025; Published: 7 November 2025
* Correspondence: mallinishanth72@gmail.com

variants, have significantly enhanced the capabilities of NLP systems. These models capture contextual information and have demonstrated remarkable success in various language understanding tasks, including sentiment analysis, text classification, and machine translation.

In this research, we explore the application of transformer-based models for detecting abusive comments in social media texts. Specifically, we leverage the contextual understanding and language representation capabilities of models such as BERT, RoBERTa, ALBERT, and DistilBERT to identify and classify abusive content accurately. By fine-tuning pre-trained transformer models on annotated datasets of social media comments, we aim to achieve high accuracy in abusive comment detection. However, these models typically involve training on a fixed task. The task is available at once, and the model is trained in a single batch or over several epochs using this task. After deployment, the model does not continue to learn. Any new task would require retraining the models from scratch or updating them periodically with new batches of data. In contrast, continual learning involves training models incrementally and learning from a stream of tasks that arrive over time, updating their knowledge continually. Continual learning has been extensively researched in machine learning, computer vision, and NLP. It involves incremental learning of tasks, assuming that once a task 't' is learned, its training data, or most of it, is no longer available. In NLP, tasks can range from end tasks like text classification and summarization to domain-specific corpora used for further pre-training of language models. The nature of tasks in NLP often differs from those in computer vision due to their distinct applications. Additionally, social media language is highly dynamic, with new usage patterns constantly emerging.

In this work, we incorporate continual learning techniques into our model development process to address this challenge. Continual learning enables models to learn from new data incrementally without forgetting previously acquired knowledge. Initially, we train the models using abusive comments and then apply continual learning to enhance the ability of the models in offensive text classification. We integrate several continual learning approaches into our models and evaluate their performance on both abusive and offensive tasks.

The major contributions of the proposed work include:

i. Model Development: We develop and fine-tune transformer-based models specifically for the task of abusive comment detection, demonstrating their effectiveness compared to traditional methods.
ii. We incorporate continual learning strategies to ensure that our models remain effective over time, adapting to new types of offensive language as they appear.
iii. Evaluation and validation: We rigorously evaluate the developed models using abusive and offensive tasks, ensuring their robustness and generalizability across different tasks.

The novelty of the proposed work lies in its application of continual learning specifically to NLP tasks, a focus that has not been addressed by existing studies, which predominantly apply continual learning to computer vision tasks [15]. From our literature survey, we identified that only a few studies review continual learning approaches for NLP tasks such as topic classification, intent classification, and question answering [16]. This highlights the unique contribution of our work in advancing the use of continual learning within the NLP domain. To be more specific, we could not find any work focusing on abusive and offensive texts. Understanding the relationship between these texts helps in developing better detection and mitigation strategies, ensuring that social media platforms can effectively address harmful content while fostering respectful communication.

The rest of the article is structured as follows: Section 2 reviews related work on continual learning. Section 3 introduces the task sets used in our study and describes the continual learning methods implemented. Section 4 details the experimental setup and presents the results. Section 5 discusses the findings from our experiments. Lastly, Section 6 concludes the article and suggests potential future research directions.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

2. Literature Survey

A significant challenge in incremental learning is catastrophic forgetting (CF), where a model loses previously acquired knowledge when learning new tasks. Continual learning strategies aim to mitigate this issue, ensuring that the model retains essential information from previous tasks. In this section, we review existing research on utilizing continual learning for task-incremental learning.

We begin by summarizing works that reviewed existing research on continual learning. The authors of Parisi et al. [17] and Wang et al. [18] highlighted the main challenges associated with lifelong learning in artificial systems and compared various neural network approaches that address CF to varying degrees. Gepperth and Hammer [19] formalized the concept of incremental learning, discussed the associated challenges, and provided an overview of standard approaches, their theoretical foundations, and recent applications. De Lange et al. [20] focused on task-incremental classification, where tasks are presented in a batch-like manner. Their work provided a taxonomy and comprehensive overview of state-of-the-art approaches for incremental learning, a novel framework to continuously determine the stability-plasticity trade-off of the continual learner, and an extensive experimental comparison of 10 state-of-the-art continual learning methods and 4 baselines. The Tiny ImageNet and a large-scale unbalanced iNaturalist datasets were used to evaluate the reviewed approaches. Ke and Liu [16] conducted a comprehensive review and analysis of recent advancements in continual learning within the field of NLP. Their work covered all settings of continual learning, presenting a taxonomy of existing techniques, methods to prevent CF, and strategies for knowledge transfer, which are significant for NLP tasks.

Similarly, Biesialska et al. [21] presented an extensive overview of existing research on continual learning in NLP, categorizing various machine learning paradigms and methods designed to mitigate CF. They also discussed the application of these methods to different NLP tasks and summarized the available benchmark datasets and evaluation approaches. Yang et al. [22] introduced an innovative NLP transformer model designed to reduce CF in online continual learning through attention calibration. This research proposed a general and lightweight feature calibration method to address task-incremental continual learning issues and included extensive empirical experiments on Seq2Seq language generation applications. Varshney et al. [23] proposed an incremental learning method that combines prompt-based classification with generative replay using a pre-trained language model. This approach demonstrated superior performance compared to traditional replay-based methods for lifelong intent detection across three public intent classification datasets, a text classification dataset, and two public multi-domain dialog intent detection datasets.

Wang et al. [24] presented the Meta Lifelong Learning (MeLL) framework to tackle intent classification tasks. In MeLL, a BERT-based text encoder is used to learn text representations across tasks, which is incrementally updated for lifelong learning. This framework employs global and local memory networks to capture cross-task prototype representations of different classes, functioning as a meta-learner that quickly adapts to various tasks. The results demonstrate that MeLL enhances performance compared to strong baselines while also reducing total parameters. Qian et al. [25] proposed a lifelong learning approach for hate speech classification on social media. To address CF, they utilized Variational Representation Learning along with a memory module based on Load-Balancing Self-Organizing Incremental Neural Network (LB SOINN). Their research demonstrated that the combination of variational representation learning and the LB SOINN memory module outperforms commonly used lifelong learning techniques.

A lifelong learning method based on a Bayesian optimization framework utilizing stochastic gradient descent was introduced by Chen et al. [26]. Their experimental results showed that this method significantly outperforms baseline methods, highlighting lifelong learning as a promising research direction. The authors of de Masson d'Autume et al. [27] introduced a lifelong language learning framework where the model continuously learns from a stream of text examples without relying on dataset identifiers. They proposed an episodic memory model that uses sparse experience replay and local adaptation to reduce catastrophic forgetting (CF). Experiments conducted on text

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

classification and question answering tasks demonstrated that sparse experience replay and local adaptation helped the model effectively learn from new datasets over time. Additionally, several studies have focused on continual learning in computer vision. Li et al. [28] tackled the challenge of adapting a vision system to new tasks while maintaining performance on original tasks without access to the original training data. They proposed the Learning without Forgetting (LwF) method for convolutional neural networks, which combines knowledge distillation and fine-tuning to learn parameters that are discriminative for the new task while preserving the knowledge of the original tasks. The effectiveness of their method was demonstrated on several classification tasks. A study by Huang and Asemi [29] introduced DllaGPT (Dynamic Lifelong Learning Aspect-Based Sentiment Analysis GPT), a transformer-based model capable of learning across multiple domains while reducing catastrophic forgetting. The model is sequentially fine-tuned on Laptops, Restaurants, Tweets, and Finance datasets using a data retention mechanism that preserves prior domain knowledge during new task learning. Experimental results show that DllaGPT achieves an average accuracy of 0.85 and a Backward Transfer score of -0.09, confirming its effectiveness and stability in continual, multi-domain sentiment analysis. Jain et al. [30] proposed an incremental learning approach for continuously updating domain-specific vocabularies. The proposed DocLib archive efficiently stores essential information from earlier data to support ongoing vocabulary refinement. Its performance is evaluated through a text classification task, where classification accuracy indicates the vocabulary's effectiveness in identifying new, domain-relevant content. Experimental results confirm that incremental updates sustain both the relevance and discriminative power of the vocabulary without degrading its overall performance.

To summarize, incremental learning faces the significant challenge of CF, where models lose previously learned knowledge when acquiring new tasks. Continual learning strategies aim to mitigate this issue, ensuring that models retain essential information from prior tasks. From the literature review, we find that there are a few works focusing on continual learning of sentiment classification, intent classification, hate speech classification, question-answering, etc., but there is no work focusing on continual learning of abusive and offensive texts on social media comments. Developing models to detect abusive and offensive comments on social media is essential for promoting a safer online environment and protecting the mental health of users. This necessitates the need for models that use continual learning to provide a scalable solution, ensuring the models can process and learn from new data consistently. Such models can handle a sequence of tasks within a unified architecture, leveraging shared features and improving overall performance, making them well-suited for the dynamic and evolving nature of social media content.

3. Materials and Methods

3.1. Dataset Description

The dataset used for this research has been collected from the social platforms[1] such as Twitter and Facebook activities are often based on suspicious behaviors such as racism, discrimination, abusive language, and threats, which are commonly associated with cyberbullying. The dataset has been labeled according to the presence of suspicious words used in tweets and comments. Suspicious data is tagged with 1, while non-suspicious data is tagged with 0. The dataset comprises up to 20,001 rows of sentiment data, with 12,179 rows labeled as negative sentiments, including racism, discrimination, and abuse. The remaining 7,822 rows are marked with positive or neutral sentiments, indicating that the comments are not suspicious. A sample of texts from the dataset is presented in Table 1.

[1] https://www.kaggle.com/datasets/syedabbasraza/suspicious-communication-on-social-platforms

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

Table 1.
Sample abusive texts.

Texts	Category
I loved it, I laughed my ass off	1
Ahh, you might be on to something...damn techies	1
What is your guilty pleasure Disney movie?	0
Yeah, if I tried.	0

3.2. Classifiers

In this section, we present a brief overview of the pretrained models that are used in our work.

3.2.1. Bidirectional Encoder Representations from Transformers (BERT)

BERT understands word context by considering both preceding and following words in a sentence, making it bidirectional [31]. BERT is pre-trained using two unsupervised tasks: Masked Language Modeling and Next Sentence Prediction. Pre-trained BERT can be fine-tuned on specific tasks with labeled data. The BERT architecture is based on the multilayer bidirectional transformer described by Vaswani et al. [32]. By adding an output layer, BERT may be optimized to provide state-of-the-art models for a variety of applications, including language inference and question answering, without requiring significant changes to the task-specific architecture.

3.2.2. DistilBERT

DistilBERT, a variant of BERT, offers a faster and more lightweight model while preserving much of BERT's performance through knowledge distillation [33]. DistilBERT undergoes pre-training on extensive text datasets using Masked Language Modeling to develop broad language understanding. It can then be fine-tuned on specific text classification tasks with labeled data, allowing it to adapt to the needs of these tasks. Its compact design makes it well-suited for various NLP applications, including text classification, where it provides both high accuracy and efficiency.

3.2.3. A Lite BERT (ALBERT)

ALBERT is a smaller variant of the BERT model, intended to reduce size and computing requirements while preserving performance [34]. ALBERT accomplishes this by employing factorized embedding parameterization and cross-layer parameter sharing. ALBERT divides the size of the embedding matrix into two smaller ones. This minimizes the number of parameters while still providing a high-level representation of the input task. Instead of having separate settings for each layer, ALBERT spreads them across several layers. This reduces the overall number of parameters, making the model lighter. Such improvements increase efficiency, allowing it to handle NLP tasks with fewer resources and faster processing times while still producing good results.

3.2.4. Robustly Optimized BERT Approach (RoBERTa)

By improving its pre-training strategy, RoBERTa expands on BERT and achieves better results in a range of NLP tasks. It uses self-attention techniques and a stack of encoders in the same transformer-based encoder design as BERT [35]. Like BERT, RoBERTa analyzes text in a bidirectional manner, considering context from the left and right sides. RoBERTa employs dynamic masking, which modifies the masked tokens in each epoch to expose the model to a wider range of text variations. In addition, RoBERTa only focuses on Masked Language Modeling and can be further refined on labeled datasets for specialized text classification tasks, modifying the pre-trained model to fit the requirements of the tasks. RoBERTa has been extensively pre-trained and fine-tuned, which makes it highly effective for text classification and other NLP applications.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

3.3. Continual Learning

It refers to the ability of a model to learn and adapt to new tasks over time without forgetting previously acquired knowledge [18, 36]. This is necessary in dynamic environments where data characteristics or user needs change over time. Continual learning techniques aim to mitigate CF, where learning new tasks causes the model to lose knowledge of previously learned tasks [37]. There are three types of continual learning methods, namely task incremental learning, domain incremental learning, and class incremental learning [16, 36, 38], which are briefly introduced below.

Task Incremental Learning: In this setting, tasks are learned sequentially, and the models are aware of the task they are currently working on [20, 38-40]. Task labels are provided during both training and testing. For instance, a model first learns to classify sentiments in movie reviews, then learns to classify news articles by topic, and finally learns to detect spam in emails.

Domain Incremental Learning: In this learning, the same task is presented in different domains, but the models do not know which domain they are currently classifying [38, 41]. The model must generalize across domains while learning to differentiate between them. For example, a sentiment analysis model learns to classify text as positive or negative across different domains, such as product reviews, movie reviews, and social media posts.

Class Incremental Learning: In class incremental learning, new classes are introduced sequentially, and the models must learn to classify them alongside previously learned classes [36, 38, 40]. A model learns to classify animals in a sequence, first learning to identify cats and dogs, then learning to identify tigers and so on.

In our work, we employ task incremental learning to first learn hope speech, followed by hate speech sequentially. The main features of task incremental learning are outlined as follows:

- Sequential Learning: The model is trained on tasks in a sequence, where each task comprises a distinct set of classes.
- Task Identifiers (Task-IDs): Unique identifiers, known as Task-IDs, are assigned to each task. These identifiers are accessible during both the training and testing phases, and the classes across tasks may or may not overlap.
- Training and Testing: During training, the model sequentially learns from the data associated with each task. During testing, the model uses the provided Task-ID to classify the relevant task.
- Class Overlap: The classes across different tasks can either be disjoint or overlapping. Disjoint tasks have entirely separate classes, while non-disjoint tasks may share some classes.

3.3.1. Task Incremental Learning

In task incremental learning, CF refers to the phenomenon where a model, when trained on new tasks, forgets the knowledge it acquired from previous tasks. This occurs because the model updates its parameters to learn for the new task, at the cost of performance on earlier tasks. CF prevention involves decreasing the amount of parameter sharing between tasks or avoiding significant changes to previously learned parameters. Addressing CF is important in task incremental learning because it ensures that the model retains knowledge from all previously learned tasks. Regularization, replay approaches, and parameter isolation are generally used to prevent CF, allowing the model to learn new tasks while maintaining its ability to perform well on previously learned tasks [16, 37, 42, 43]. In this work, we propose using two task incremental learning approaches, namely, replay-based and regularization-based approaches, to address CF in the developed models. To test the incremental learning capability of the models, we sequentially introduce two tasks: hope speech, followed by hate speech. Below, we summarize the steps in task incremental learning approaches.

3.3.2. Replay-based Task Incremental Learning

This is an approach to continual learning where a model is trained on new tasks while simultaneously using or replaying data from previously learned tasks. There are two types of replay-based methods, namely Rehearsal and Pseudo-Rehearsal.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

3.3.2.1. Rehearsal approach for Task Incremental learning

The rehearsal approach for task incremental learning is designed to reduce catastrophic forgetting (CF) by revisiting a subset of data from previously learned tasks. The objective is to combine the loss from the current task and the loss from the replayed data. The significant components of this approach include:

Storing of data: A small buffer stores a sample of data from each task the model has already learned. This sample is carefully chosen so that it represents the most essential aspects of each task.

Periodic Rehearsal: While training on new tasks, the model periodically rehearses by retraining on stored data from previous tasks. By integrating this rehearsal data with the new task data, the model strengthens its understanding of earlier tasks while simultaneously learning the new ones.

Preserving knowledge on previous tasks: The rehearsal approach helps the models in preserving knowledge from previous tasks by regularly revisiting stored data. This minimizes the forgetting of the parameters of the models when learning new tasks, thereby helping to prevent CF.

The steps in rehearsal-based approach are listed below:

Step 1: Train the models on Hope Speech dataset.

Step 2: Initialize the buffer(M) with a subset of hope speech comments

Step 3: Train the models on offensive comments (D_{Off}) combined with replayed data from the buffer

Step 4: Compute the loss on task 2, L_{Off} and the loss on replayed data, L_{Abu}

Step 5: Combine both losses as in Equation (1):

$$L(\theta) = \lambda L_{Off} + (1 - \lambda) L_{Abu}(\theta) \quad (1)$$

where λ is a hyperparameter balancing the importance of new and replayed data and θ are the parameters being updated. And L_{Off} and L_{Abu} are calculated as Equation (2) and Equation (3)

$$L_{Off}(\theta) = \left(\frac{1}{|D_{Off}|}\right) \sum_{(x,y)\epsilon D_{Off}} L(f(x;\theta), y) \quad (2)$$

$$L_{Abu}(\theta) = \frac{1}{|M|} \sum_{(x,y)\epsilon M} L(f(x;\theta), y) \quad (3)$$

Step 6: Update θ by minimizing the combined loss $L(\theta)$.

Step 7: Update the memory buffer with a subset of the previous task data by replacing older samples.

3.3.2.2. Pseudo-Rehearsal Method

Since the rehearsal-based approach involves storing a small part of real data from each task, which can become memory-intensive as the number of tasks grows, we have used the pseudo-rehearsal approach instead. This method generates synthetic data on the fly, reducing the need for extensive memory storage. In this case, the loss for both tasks is computed similarly to the rehearsal approach, with a small change in how the loss on pseudo-replayed data is calculated, as shown in Equation (4).

$$L_{Abu}(\theta) = \frac{1}{|D_G|} \sum_{(x,y)\epsilon D_G} L(f(x;\theta), y) \quad (4)$$

Where G is the model that generates synthetic samples from the previous tasks, and D_G is the data generated by the generative model. Both methods aim to balance learning new tasks while retaining knowledge from previous tasks, thus addressing the problem of CF in continual learning scenarios.

3.3.3. Regularization-Based Task Incremental Learning

Regularization methods do not require storing data from previous tasks, unlike rehearsal and pseudo-rehearsal methods. This makes them more memory-efficient, especially as the number of tasks increases. Regularization methods work by adding regularization terms to the loss function that penalize changes to important parameters. In this work, we use a regularization-based approach called Learning Without Forgetting (LwF). The details of LwF are given below.

LwF: It is a regularization-based approach for task-incremental learning that mitigates catastrophic forgetting (CF) by preserving the knowledge of previously learned tasks while learning new ones. It

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

works on the principle of using the predictions of the models on previous tasks as a form of soft target regularization while training on new tasks. The steps involved in the LwF approach are as follows:
The following notations are used

- x is the input data
- $y1$ and y2 are the true labels of the abusive and offensive tasks, respectively.

Step 1: Train the abusive texts

i. Train the models on the abusive texts and calculate the loss as in Equation (5)

$$L_{Abu}(\theta) = CrossEntropyLoss(f_{\theta}(x), y_1) \quad (5)$$

Where $f_{\theta}(x)$ is the model output with parameters θ.

ii. Save the parameters such as weights and biases of the developed models and their outputs of the initial task (abusive task). These outputs will serve as soft targets when training on the next (offensive texts) task. Then, compute the soft targets, Z_{Abu} for an abusive task using the model trained on the abusive task as given in Equation (6)

$$Z_{Abu} = f_{\theta^*_{Abu}}(x) \quad (6)$$

Where θ^*_{Abu} are the optimal parameters after training on the abusive task.

Step 2: Train on offensive texts using regularization with LwF

i. Initialize the training with the previously learned model parameters θ^*_{Abu} and train the models using offensive texts

ii. Compute the loss for the offensive comments as shown in Equation (7)

$$L_{Off}(\theta) = CrossEntropyLoss(f_{\theta}(x), y_2) \quad (7)$$

iii. Use Equation (8) to compute the loss on the soft targets from the previous tasks using the saved outputs.

$$L_{KD}(\theta) = KLDivergence(Softmax(f_{\theta}(x)/T), Softmax(Z_{Abu}/T)) \quad (8)$$

Where T is the temperature parameter used to soften the probability distributions. Higher temperatures produce softer probability distributions, making it easier to transfer knowledge.

Step 3: Calculate the total loss

i. Combine the loss on the offensive task and the knowledge distillation loss from the abusive task to find the total loss. (The knowledge distillation loss helps to retain the information from previous tasks). This is done using Equation (9).

$$L_{total}(\theta) = L_{Off}(\theta) + \lambda L_{KD}(\theta) \quad (9)$$

Where λ is the balancing parameter used to control the trade-off between retaining old knowledge on abusive texts and learning new information from offensive texts.

Step 4: Update the parameters of the models

i. Update the model parameters as in Equation (10) by minimizing the combined loss to ensure that while the model learns the offensive task, it does not forget the knowledge acquired from the abusive task.

$$\theta^* = argmin_{\theta} L_{total}(\theta) \quad (10)$$

The workflow of the proposed research is shown in Figure 1.

4. Experimental Setup and Results

4.1. Experimental Setup

To implement and assess the performance of continual learning approaches in comparison to pre-trained models, we adapted the continual hyperparameter selection framework introduced in Ritter et al.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

[3] to ensure a fair comparison with existing methods. The experimental setup utilized in this study is detailed in Table 1. All models were developed using the PyTorch library.

Table 1.
Experimental Setup.

Parameters	Values set
Stochastic Gradient Descent with momentum as an optimizer	0.8
Training epochs	100
Early stopping	5
Batch size	128
Softmax Temperature (T)	2
Knowledge Distillation Strength (λ)	0.5

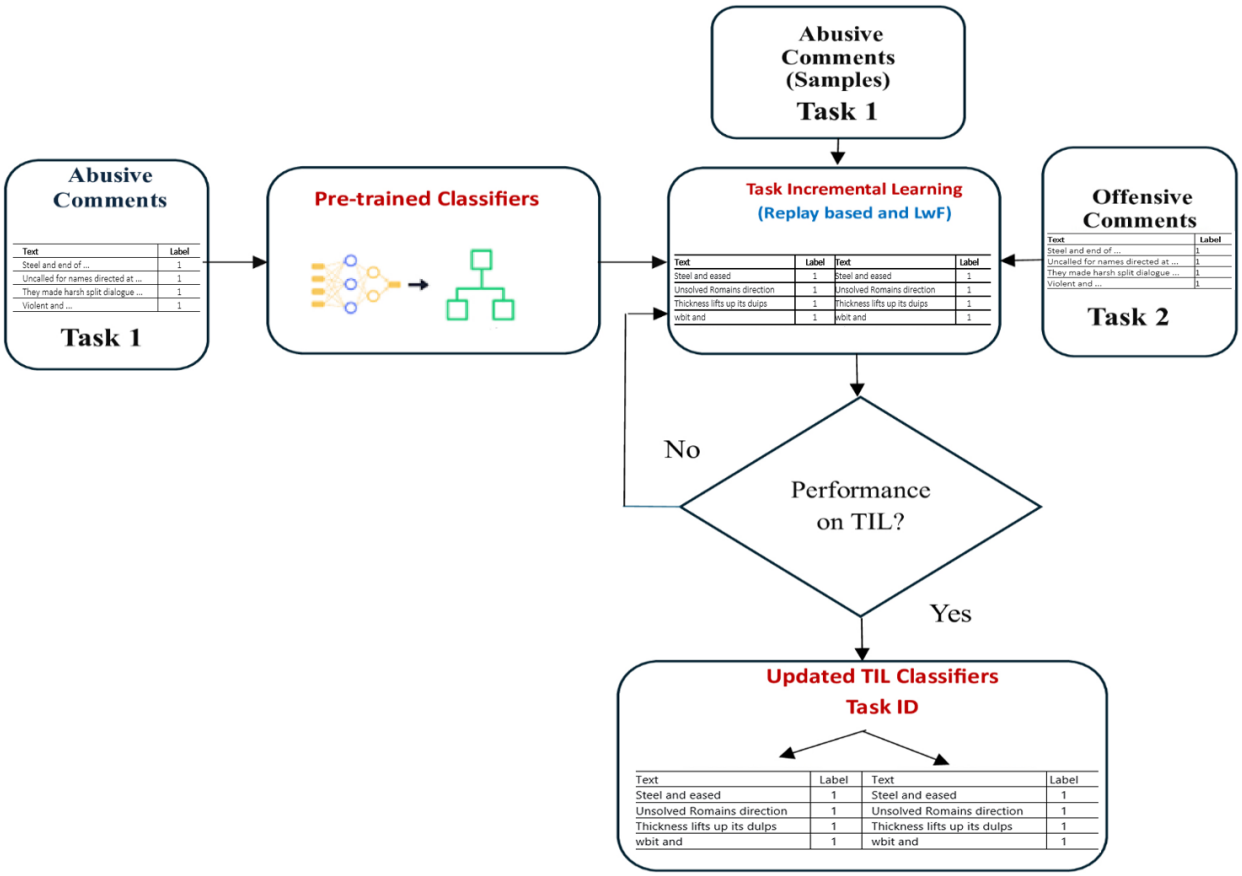

Figure 1.
Proposed workflow.

Our work utilizes Gradient Episodic Memory (GEM) for replay-based task incremental learning and implements Learning without Forgetting (LwF) from scratch using Python. A multi-head configuration is employed, where the model shares hidden layers but has separate output heads for different tasks. Specifically, a two-head configuration is implemented, with one head classifying abusive comments and the other classifying offensive comments.

4.2. Experimental Results

We then present the results from a series of experiments conducted on the developed models. Initially, the pre-trained models were tested on abusive texts, followed by offensive texts, without any retraining. Subsequently, we implemented task incremental learning using both replay-based and regularization-based approaches. Table 2 shows the performance of the models on abusive comments.

Table 2.
Performance of the classifiers on abusive comments.

Classifiers	Class	Accuracy	Recall	Precision	F1 Score
BERT	Abusive	91.33	89.83	95.56	92.61
	Non-Abusive		93.62	85.76	89.51
	Macro F1	91.06			
	Weighted F1	91.38			
ALBERT	Abusive	77.76	74.00	87.27	80.09
	Non-Abusive		83.50	67.74	74.80
	Macro F1	77.45			
	Weighted F1	78.00			
RoBERTa	Abusive	90.98	95.51	83.90	89.33
	Non-Abusive		88.01	96.77	92.18
	Macro F1	90.76			
	Weighted F1	91.06			
DistilBERT	Abusive	92.03	92.90	93.88	93.39
	Non-Abusive		90.70	89.27	89.98
	Macro F1	91.68			
	Weighted F1	92.05			

Next, we evaluated the models on an offensive dataset to assess their continual learning capability. The offensive comments dataset, sourced from Huggingface[2] contains 18,858 rows labeled as offensive and 10,364 rows labeled as non-offensive. Initially, the models were tested without any retraining or fine-tuning, and their performance is detailed in Table 3.

Table 3.
Performance of the classifiers on offensive comments without retraining.

Classifiers	Class	Accuracy	Recall	Precision	F1 Score
BERT	Offensive	62.24	45.90	92.14	61.27
	Not Offensive		92.70	47.90	63.16
	Macro F1	62.22			
	Weighted F1	61.93			
ALBERT	Offensive	73.19	66.48	89.65	76.35
	Not Offensive		85.69	57.84	69.06
	Macro F1	72.70			
	Weighted F1	73.80			
RoBERTa	Offensive	69.24	69.87	80.30	74.73
	Not Offensive		68.05	54.79	60.71
	Macro F1	67.72			
	Weighted F1	69.83			
DistilBERT	Offensive	60.17	43.69	89.94	58.81
	Not Offensive		90.89	46.41	61.44
	Macro F1	60.13			
	Weighted F1	59.73			

As seen in Table 3, the performance of the models on the offensive dataset is relatively lower compared to the abusive dataset. This difference arises because the models were trained on abusive

[2]https://huggingface.co/datasets/christinacdl/offensive_language_dataset

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

comments and therefore learned parameters specific to that dataset. When applied to the offensive dataset, the knowledge gained from the previous task did not significantly enhance the classification accuracy. Therefore, we incorporated continual learning approaches into the developed models to preserve existing knowledge while acquiring new knowledge for different tasks. First, we implemented replay-based approaches, namely rehearsal and pseudo-rehearsal methods, as discussed in Section 3.3. The performance of the models using replay-based approaches is presented in Tables 4 and 5. These tables depict the performance of the models on both the old task (abusive comments) and the new task (offensive comments) and clearly show that after incorporating replay-based approaches, the performance on the abusive dataset declined slightly compared to before their implementation, while performance on the offensive dataset improved significantly. This indicates that the models are attempting to balance both datasets using the replay-based approach. Notably, pseudo-rehearsal approaches demonstrated better performance compared to rehearsal approaches.

Table 4.
Performance of the classifiers on abusive comments using Replay-based approaches.

a. Rehearsal-based approach.

Classifiers	Class	Accuracy	Recall	Precision	F1 Score
BERT	Abusive	73.20	93.91	66.05	77.55
	Non-Abusive		53.06	89.97	66.75
	Macro F1	72.15			
	Weighted F1	72.08			
ALBERT	Abusive	73.50	87.63	67.92	76.53
	Non-Abusive		59.76	83.24	69.58
	Macro F1	73.05			
	Weighted F1	73.00			
RoBERTa	Abusive	75.80	71.20	77.83	74.36
	Non-Abusive		80.28	74.13	77.08
	Macro F1	75.72			
	Weighted F1	75.74			
DistilBERT	Abusive	76.10	75.74	75.40	75.93
	Non-Abusive		76.47	76.80	76.27
	Macro F1	76.10			
	Weighted F1	76.10			

b. Pseudo Rehearsal approach

Classifiers	Class	Accuracy	Recall	Precision	F1 Score
BERT	Abusive	89.34	81.56	96.59	88.44
	Non-Abusive		97.12	84.04	90.11
	Macro F1	89.28			
	Weighted F1	89.28			
ALBERT	Abusive	78.00	76.00	80.00	78.00
	Non-Abusive		81.00	77.00	79.00
	Macro F1	78.00			
	Weighted F1	78.00			
RoBERTa	Abusive	86.62	85.48	87.47	86.47
	Non-Abusive		87.76	85.80	86.77
	Macro F1	86.62			
	Weighted F1	86.62			
DistilBERT	Abusive	90.72	98.36	85.32	91.38
	Non-Abusive		83.08	98.06	89.95
	Macro F1	90.67			
	Weighted F1	90.67			

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

Table 5.
Performance of the classifiers on Offensive comments using Replay-based approaches.

a. Rehearsal-based approach

Classifiers	Class	Accuracy	Recall	Precision	F1 Score
BERT	Offensive	87.60	92.05	83.65	87.65
	Not Offensive		83.52	91.98	87.55
	Macro F1	87.60			
	Weighted F1	87.60			
ALBERT	Offensive	73.40	94.56	65.32	77.26
	Not Offensive		54.02	91.56	67.95
	Macro F1	72.61			
	Weighted F1	72.40			
RoBERTa	Offensive	81.60	81.59	80.25	80.91
	Not Offensive		81.61	82.88	82.24
	Macro F1	81.58			
	Weighted F1	81.61			
DistilBERT	Offensive	90.80	90.79	90.04	90.42
	Not Offensive		90.80	91.51	91.15
	Macro F1	90.79			
	Weighted F1	90.80			

b. Pseudo Rehearsal approach

Classifiers	Class	Accuracy	Recall	Precision	F1 Score
BERT	Offensive	98.02	97.72	97.73	98.01
	Not Offensive		98.32	98.31	98.03
	Macro F1	98.02			
	Weighted F1	98.02			
ALBERT	Offensive	93.00	88.00	97.00	92.00
	Not Offensive		97.00	88.00	93.00
	Macro F1	93.00			
	Weighted F1	93.00			
RoBERTa	Offensive	95.50	94.40	96.52	95.45
	Not Offensive		96.60	94.52	95.55
	Macro F1	95.50			
	Weighted F1	95.50			
DistilBERT	Offensive	98.10	96.80	99.38	98.07
	Not Offensive		99.40	96.88	98.12
	Macro F1	98.10			
	Weighted F1	98.10			

In addition to replay-based approaches, we have also integrated LwF into the developed models, and the results are shown in Tables 6 and 7. From these tables, we infer that the performance of the models significantly increased after integrating LwF.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

Table 6.
Performance of the classifiers on Abusive comments using LwF.

Classifiers	Class	Accuracy	Recall	Precision	F1 Score
BERT	Abusive	91.06	97.52	86.36	91.60
	Non-Abusive		84.60	97.15	90.44
	Macro F1	91.02			
	Weighted F1	91.02			
ALBERT	Abusive	76.92	93.68	70.16	80.23
	Non-Abusive		60.16	90.49	72.27
	Macro F1	76.25			
	Weighted F1	76.25			
RoBERTa	Abusive	80.90	93.08	74.85	82.97
	Non-Abusive		68.72	90.85	78.25
	Macro F1	80.61			
	Weighted F1	80.61			
DistilBERT	Abusive	87.74	98.24	81.19	88.90
	Non-Abusive		77.24	97.77	86.30
	Macro F1	87.60			
	Weighted F1	87.60			

Table 7.
Performance of the classifiers on Offensive comments using LwF.

Classifiers	Class	Accuracy	Recall	Precision	F1 Score
BERT	Offensive	98.80	99.12	98.49	98.80
	Not Offensive		98.48	99.11	98.80
	Macro F1	98.80			
	Weighted F1	98.80			
ALBERT	Offensive	95.82	94.92	96.66	95.78
	Not Offensive		96.72	95.01	95.86
	Macro F1	95.82			
	Weighted F1	95.82			
RoBERTa	Offensive	95.52	92.76	98.18	95.39
	Not Offensive		98.28	93.14	95.64
	Macro F1	95.52			
	Weighted F1	95.52			
DistilBERT	Offensive	97.86	98.00	97.73	97.86
	Not Offensive		97.72	97.99	97.86
	Macro F1	97.86			
	Weighted F1	97.86			

Further to evaluate the performance of the developed task incremental models, we adapt the metrics recommended by Akhter et al. [2], such as average accuracy, average incremental accuracy, forgetting rate, forward transfer, and backward transfer. Measuring these metrics in incremental learning is crucial for a comprehensive evaluation of the models' performance. These metrics assess the ability of the models to learn new tasks while retaining knowledge from previous ones, balancing performance across tasks, and mitigating the effects of catastrophic forgetting (CF). They also help evaluate how well knowledge is transferred between tasks, both forward and backward, providing insights into the overall effectiveness in sequential and continual learning settings. A brief note on these metrics is provided below:

Average accuracy: It refers to the mean performance of a model across all tasks in a task-incremental learning setting. After the model has learned all tasks, average accuracy is calculated by evaluating the model on each task and computing the average of these performance scores. It is calculated as given in Equation (11).

$$A_T = \frac{1}{T}\sum_{t=1}^{T} a_{T.t} \quad (11)$$

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

Where T is the number of tasks and $a_{T.t}$ refers to the performance of the model on the testing set of tasks after the model has continually trained all T tasks, that is, the accuracy on task 't' after learning the t^{th} task.

Average incremental accuracy: It measures how well a model retains knowledge of previous tasks while learning new ones. It is calculated as the average accuracy of the model on each task at the time that task was learned. This metric highlights the ability of the models to maintain performance on earlier tasks as it progresses through new tasks and is represented as given in Equation (12).

$$A_{IA} = \frac{1}{T}\sum_{t=1}^{T} a_{t.t} \quad (12)$$

where $a_{t.t}$ is the accuracy on task 't' immediately after learning task 't'.

4.3. Forgetting Rate

It measures the extent to which the performance of a model on previously learned tasks degrades as it learns new tasks. It is calculated by comparing the accuracy of the model on a task immediately after it was learned to its accuracy on the same task after learning subsequent tasks. A higher forgetting rate indicates more significant forgetting, which is a key challenge in task incremental learning, and is calculated as follows in Equation (13):

$$F_T = \frac{1}{T-1}\sum_{t=1}^{T-1}(a_{t,t} - a_{T,t}) \quad (13)$$

where $a_{t,t}$ represents the test accuracy for task t immediately after it is learned, and $a_{T,t}$ represents the accuracy for task t after training on the final task T. We average these values across all trained tasks except the last one, since the final task does not exhibit forgetting. A lower forgetting rate indicates higher knowledge preservation.

Forward transfer: This refers to the ability of a model to use knowledge from previously learned tasks to improve learning and performance on a new task. Positive forward transfer occurs when knowledge gained from previous tasks helps improve performance on subsequent tasks. It is measured by comparing the performance on a new task with and without prior task knowledge. It is given as mentioned in Equation (14).

$$FWT_T = \frac{1}{T}\sum_{t=1}^{T}(a_{t,t} - a_{0,t}) \quad (14)$$

where $a_{0,t}$ refers to the performance of training task t when it is trained from scratch individually and $a_{t,t}$ represents the test accuracy for task t immediately after it is learned.

Backward transfer: It evaluates the impact that learning new tasks has on the performance of the models on previously learned tasks. Positive backward transfer occurs when learning a new task enhances the performance on earlier tasks, while negative backward transfer indicates that learning new tasks has led to a decline in performance on previous tasks. The formula to find the backward transfer rate is given in Equation (15).

$$BWT_T = \frac{1}{T-1}\sum_{t=1}^{T-1}(a_{T,t} - a_{t,t}) \quad (15)$$

All the above metrics have been measured, and these values are presented in Tables 8 to 10.

Table 8.
Performance of the classifiers using the reply-based rehearsal approach.

Metrics/ Models	Average Accuracy	Average Incremental Accuracy	Forgetting Rate	Forward Transfer	Backward Transfer
BERT	80.4	87.6	0.2	4.01	0.01
ALBERT	73.45	79.4	9.79	5.21	0.01
RoBERTa	78.7	81.6	0.01	8.59	7.43
DistilBERT	83.45	90.8	0.01	8.67	9.79

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

Table 9.
Performance of the classifiers using the reply-based pseudo rehearsal approach.

Metrics/ Models	Average Accuracy	Average Incremental Accuracy	Forgetting Rate	Forward Transfer	Backward Transfer
BERT	93.68	98.8	1.98	98.02	10.66
ALBERT	85.53	96.32	7.36	92.64	21.58
RoBERTa	91.06	97.75	4.5	95.5	13.38
DistilBERT	94.41	99.05	1.9	98.1	9.28

Table 10.
Performance of the classifiers using LwF.

Metrics/ Models	Average Accuracy	Average Incremental Accuracy	Forgetting Rate	Forward Transfer	Backward Transfer
BERT	94.93	92.4	1.06	98.8	7.74
ALBERT	86.37	84.98	2.78	95.82	18.9
RoBERTa	88.21	84.65	7.12	95.52	14.62
DistilBERT	92.8	90.87	3.86	97.86	10.12

5. Findings and Discussion

This section presents and discusses the impact of using task incremental learning for the sequential classification of abusive and offensive comments. The focus of this study is to evaluate the effectiveness of task incremental learning in enabling models to learn new tasks, such as offensive comment detection, without forgetting previously learned tasks, like abusive comment detection.

Prior to retraining, the models demonstrated good performance on abusive comments but exhibited a decline in accuracy when applied to offensive comments. This decrease is attributed to the models having optimized their parameters specifically for abusive text, which are not effective when evaluating offensive comments. Among the models, DistilBERT achieved an accuracy of 92.03% on the abusive dataset, while only 60.17% on the offensive dataset. Following the implementation of replay-based approaches, we observed a decline in the performance of the models on abusive comments. In particular, the rehearsal approach led to a high reduction in accuracy for abusive text classification, whereas the pseudo-rehearsal approach demonstrated comparatively better results. Notably, models like BERT, ALBERT, and DistilBERT maintained performance levels on abusive comments that were nearly consistent with those observed before applying incremental learning. Furthermore, these models exhibited significantly improved performance on offensive texts compared to their performance prior to retraining. Unlike rehearsal-based methods that rely on storing and replaying a limited subset of actual data, pseudo-rehearsal generates synthetic data that broadly represents the distribution of abusive tasks, leading to improved results. This synthetic data helps prevent overfitting and better preserves knowledge, which enhances the models' ability to generalize across tasks. Additionally, pseudo-rehearsal effectively mitigates catastrophic forgetting by providing diverse and representative examples, allowing the models to adapt to new tasks without compromising performance on earlier ones.

Following the implementation of LwF, a notable improvement in performance on offensive comments was observed across the models, with BERT achieving 98.8%, ALBERT 95.82%, and RoBERTa 95.52%. The increase in performance on offensive comments with LwF is attributed to the ability of the models to adapt effectively to new tasks while preserving knowledge of previous tasks through knowledge distillation. LwF regularizes the learning process, ensuring that the models retain the required knowledge from earlier tasks, namely abusive comment classification, while learning to classify offensive comments. This regularization helps prevent catastrophic forgetting (CF) and allows the models to perform well on the offensive dataset. Except for BERT, all models experienced a decline in performance on abusive comments when using LwF compared to the pseudo-rehearsal approach, with accuracy of ALBERT dropping from 78% to 76.92%, RoBERTa's from 86.6% to 80.9%, and DistilBERT's from 90.72% to 87.74%. These results indicate that BERT, when used with LwF, surpasses all other models in accuracy for both abusive and offensive tasks. The improved performance

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

of BERT with LwF for both abusive and offensive datasets is due to its robust bidirectional architecture. Figure 2 comprehensively compares the accuracy achieved by all the evaluated models. It demonstrates that the LwF consistently delivers superior performance across both task types.

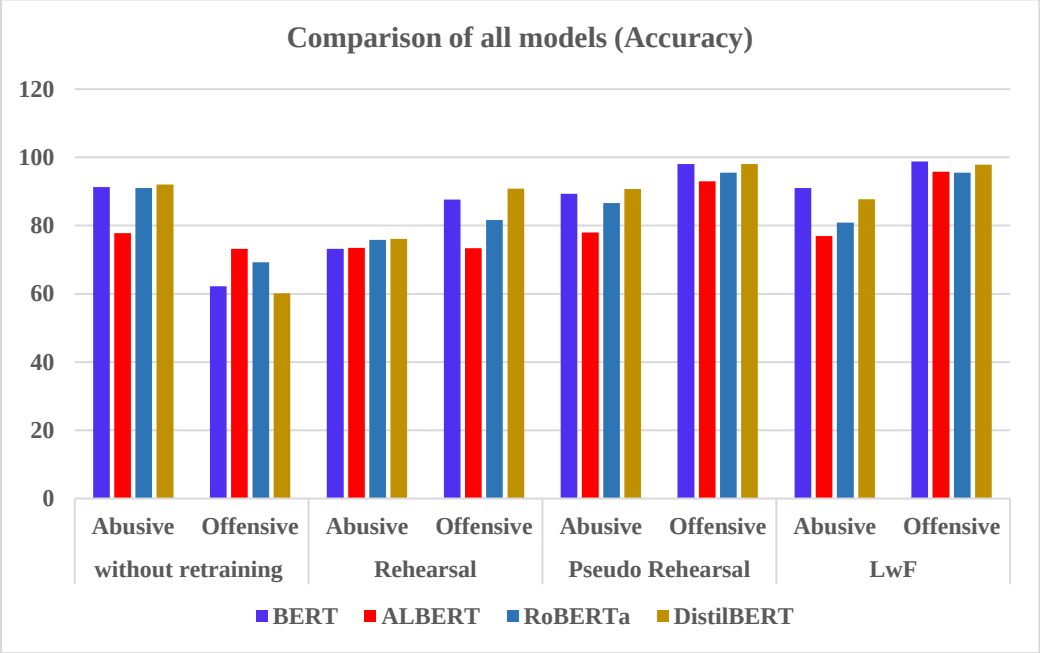

Figure 2.
Comparison of all models on both tasks.

Unlike metrics such as accuracy, recall, precision, and F1 score, which focus on single-task performance, metrics such as average accuracy, average incremental accuracy, forgetting rate, forward transfer, and backward transfer provide a comprehensive view of how well models perform in task incremental learning by assessing their ability to retain knowledge, adapt to new tasks, and manage performance across multiple tasks. From Tables 8, 9, and 10, we can infer that the average accuracy and average incremental accuracy are high for BERT with LwF compared to other models. High values of these metrics indicate improved performance. Likewise, this model demonstrates a lower forgetting rate compared to others, indicating that BERT with LwF effectively maintains performance on previously learned tasks as new tasks are introduced, thereby reducing CF. Additionally, when evaluating forward and backward transfer, BERT with LwF shows high forward transfer and low backward transfer compared to the other models. All these findings support our results, showing that BERT with LwF outperforms other models using replay-based approaches. To facilitate a comprehensive understanding of the performance of the models in incremental learning, Figure 3 presents a comparative analysis of all models within a single diagram.

Overall, these findings suggest that integrating advanced models and techniques like LwF significantly enhances performance in applications requiring continuous learning and knowledge preservation.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

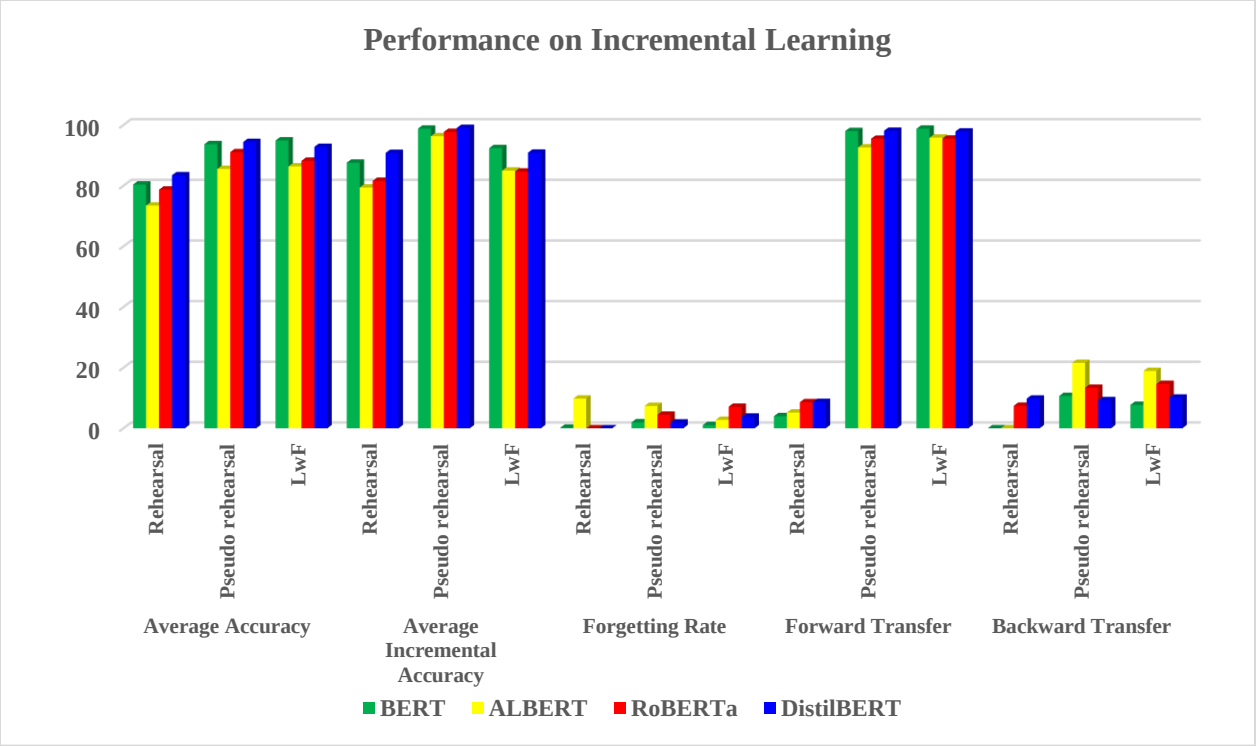

Figure 3.
Performance of all models on incremental learning.

6. Conclusion and Future Work

In this study, we investigated task incremental learning for sequential classification of abusive and offensive texts using BERT, ALBERT, DistilBERT, and RoBERTa models. To add incremental learning capability, we integrated approaches such as rehearsal-based, pseudo-rehearsal, and LwF approaches. The results demonstrate that BERT with LwF provides better performance compared to the other models, including ALBERT, DistilBERT, and RoBERTa. BERT with LwF excels in maintaining high average accuracy and average incremental accuracy, while also demonstrating a low forgetting rate and high forward transfer. These results highlight the effectiveness of BERT with LwF in preserving knowledge of previously learned tasks and adapting efficiently to new tasks. Furthermore, it is understood that the performance of sequential learning applications can be significantly improved by integrating incremental learning approaches into transformer models, where a series of related tasks need to be learned and adapted over time.

In the future, we intend to explore various approaches to further advance task incremental learning in sequential applications. Examining the integration of advanced techniques, such as meta-learning or continual learning frameworks, could yield further improvements in managing CF and enhancing forward transfer. Moreover, experimenting with a wider range of tasks beyond abusive and offensive text classification may offer insights into the general applicability of our findings. Lastly, investigating the effects of different hyperparameters and fine-tuning strategies on the performance of task incremental learning models could provide valuable guidance for optimizing their effectiveness in real-world scenarios.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 11: 448-466, 2025
DOI: 10.55214/2576-8484.v9i11.10905

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

References

[1] S. A. Castaño-Pulgarín, N. Suárez-Betancur, L. M. T. Vega, and H. M. H. López, "Internet, social media and online hate speech. Systematic review," *Aggression and Violent Behavior*, vol. 58, p. 101608, 2021. https://doi.org/10.1016/j.avb.2021.101608

[2] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. AbdelMajeed, and T. Zia, "Abusive language detection from social media comments using conventional machine learning and deep learning approaches," *Multimedia Systems*, vol. 28, pp. 1925-1940, 2022. https://doi.org/10.1007/s00530-021-00784-8

[3] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 583–593). Stroudsburg, PA: Association for Computational Linguistics*, 2011.

[4] B. R. Chakravarthi *et al.*, "Detecting abusive comments at a fine-grained level in a low-resource language," *Natural Language Processing Journal*, vol. 3, p. 100006, 2023. https://doi.org/10.1016/j.nlp.2023.100006

[5] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: A systematic review," *Language Resources and Evaluation*, vol. 55, pp. 477-523, 2021. https://doi.org/10.1007/s10579-020-09502-8

[6] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 4, pp. 215-230, 2015.

[7] G. I. Sigurbergsson and L. Derczynski, "Offensive language and hate speech detection for Danish," *arXiv preprint arXiv:1908.04531*, 2019. https://doi.org/10.48550/arXiv.1908.04531

[8] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on twitter," *arXiv preprint arXiv:1706.01206*, 2017. https://doi.org/10.48550/arXiv.1706.01206

[9] A. Khan, A. Ahmed, S. Jan, M. Bilal, and M. F. Zuhairi, "Abusive language detection in urdu text: Leveraging deep learning and attention mechanism," *IEEE Access*, vol. 12, pp. 37418-37431, 2024. https://doi.org/10.1109/ACCESS.2024.3370232

[10] R. Pelle, C. Alcântara, and V. P. Moreira, "A classifier ensemble for offensive text detection," in *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, 2018.

[11] R. Ahuja, A. Banga, and S. Sharma, *Detecting abusive comments using ensemble deep learning algorithms. In Malware Analysis Using Artificial Intelligence and Deep Learning*. Cham: Springer International Publishing, 2020.

[12] A. Baruah, L. Wahlang, F. Jyrwa, F. Shadap, F. Barbhuiya, and K. Dey, "Abusive language detection in Khasi social media comments," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2024.

[13] K. B. Nelatoori and H. B. Kommanti, "Attention-based bi-lstm network for abusive language detection," *IETE Journal of Research*, vol. 69, no. 11, pp. 7884-7892, 2023. https://doi.org/10.1080/03772063.2022.2034534

[14] S. A. Agnes, A. A. Solomon, and D. J. C. Tamilmaran, "Abusive comment detection in social media with bidirectional LSTM model," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1368-1373). IEEE*, 2023.

[15] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner, "Online continual learning in image classification: An empirical survey," *Neurocomputing*, vol. 469, pp. 28-51, 2022. https://doi.org/10.1016/j.neucom.2021.10.021

[16] Z. Ke and B. Liu, "Continual learning of natural language processing tasks: A survey," *arXiv preprint arXiv:2211.12701*, 2022. https://doi.org/10.48550/arXiv.2211.12701

[17] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54-71, 2019. https://doi.org/10.1016/j.neunet.2019.01.012

[18] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5362-5383, 2024. https://doi.org/10.1109/TPAMI.2024.3367329

[19] A. Gepperth and B. Hammer, "Incremental learning algorithms and applications," in *European Symposium on Artificial Neural Networks (ESANN)*, 2016.

[20] M. De Lange *et al.*, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3366-3385, 2021. https://doi.org/10.1109/TPAMI.2021.3057446

[21] M. Biesialska, K. Biesialska, and M. R. Costa-Jussa, "Continual lifelong learning in natural language processing: A survey," *arXiv preprint arXiv:2012.09823*, 2020. https://doi.org/10.18653/v1/2020.coling-main.574

[22] P. Yang, D. Li, and P. Li, "Continual learning for natural language generations with transformer calibration," in *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, 2022.

[23] V. Varshney, M. Patidar, R. Kumar, G. Shroff, and L. Vig, "Prompt augmented generative replay via supervised contrastive training for lifelong intent detection," presented at the Findings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2022) (pp. 1113–1127). Seattle, WA: Association for Computational Linguistics, 2022.

[24] C. Wang *et al.*, "Mell: Large-scale extensible user intent classification for dialogue systems with meta lifelong learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.

[25] J. Qian, H. Wang, M. ElSherief, and X. Yan, "Lifelong learning of hate speech classification on social media," *arXiv preprint arXiv:2106.02821*, 2021. https://doi.org/10.48550/arXiv.2106.02821

[26] Z. Chen, N. Ma, and B. Liu, "Lifelong learning for sentiment classification," *arXiv preprint arXiv:1801.02808*, 2018. https://doi.org/10.48550/arXiv.1801.02808

[27] C. de Masson d'Autume, S. Ruder, L. Kong, and D. Yogatama, "Episodic memory in lifelong language learning," presented at the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS 2019) (pp. 13122–13131). Vancouver, BC, Canada, 2019.

[28] Z. Li, D. Hoiem, and D. Forsyth, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[29] H. Huang and A. Asemi, "An experimental study on dynamic lifelong learning with GPT for mitigating catastrophic forgetting in aspect-based sentiment analysis," *IEEE Access*, vol. 13, pp. 90 316–90 332, 2025.

[30] M. Jain, H. Kaur, B. Gupta, J. Gera, and V. Kalra, "Incremental learning algorithm for dynamic evolution of domain specific vocabulary with its stability and plasticity analysis," *Scientific Reports*, vol. 15, p. 272, 2025. https://doi.org/10.1038/s41598-024-78785-6

[31] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019: Minneapolis, Minnesota.

[32] A. Vaswani *et al.*, "Attention is all you need," presented at the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NeurIPS 2017) (pp. 5998–6008). Long Beach, CA, USA, 2017.

[33] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019. https://doi.org/10.48550/arXiv.1910.01108

[34] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019. https://doi.org/10.48550/arXiv.1909.11942

[35] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019. https://doi.org/10.48550/arXiv.1907.11692

[36] S. Liu *et al.*, "A comprehensive review of continual learning with machine learning models," in *International Conference on Image, Vision and Intelligent Systems, pp. 504-512. Singapore: Springer Nature Singapore*, 2023.

[37] H.-G. Doan, H.-Q. Luong, T.-O. Ha, and T. T. T. Pham, "An efficient strategy for catastrophic forgetting reduction in incremental learning," *Electronics*, vol. 12, no. 10, p. 2265, 2023. https://doi.org/10.3390/electronics12102265

[38] G. M. Van de Ven, T. Tuytelaars, and A. S. Tolias, "Three types of incremental learning," *Nature Machine Intelligence*, vol. 4, pp. 1185-1197, 2022. https://doi.org/10.1038/s42256-022-00568-3

[39] Q. Gao, X. Shan, Y. Zhang, and F. Zhou, "Enhancing knowledge transfer for task incremental learning with data-free subnetwork," *Advances in Neural Information Processing Systems*, vol. 36, pp. 68471-68484, 2023.

[40] S. Tian, L. Li, W. Li, H. Ran, X. Ning, and P. Tiwari, "A survey on few-shot class-incremental learning," *Neural Networks*, vol. 169, pp. 307-324, 2024. https://doi.org/10.1016/j.neunet.2023.10.039

[41] H. Shi and H. Wang, "A unified approach to domain incremental learning with memory: Theory and algorithm," *Advances in Neural Information Processing Systems*, vol. 36, pp. 15027-15059, 2023.

[42] E. L. Aleixo, J. G. Colonna, M. Cristo, and E. Fernandes, "Catastrophic forgetting in deep learning: A comprehensive taxonomy," *arXiv preprint arXiv:2312.10549*, 2023. https://doi.org/10.48550/arXiv.2312.10549

[43] J. Huang *et al.*, "Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal," *arXiv preprint arXiv:2403.01244*, 2024. https://doi.org/10.48550/arXiv.2403.01244