IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 June 2024, accepted 23 July 2024, date of publication 29 July 2024, date of current version 7 August 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3435537*

RESEARCH ARTICLE

# Fine-Tuning of Distil-BERT for Continual Learning in Text Classification: An Experimental Analysis

**SAHAR SHAH[1], SARA LUCIA MANZONI[1], FAROOQ ZAMAN[2], FATIMA ES SABERY[3], FRANCESCO EPIFANIA[4], AND ITALO FRANCESCO ZOPPIS[1]**

[1]Department of Informatics, Systems and Communication, University of Milano-Bicocca, 20126 Milan, Italy
[2]Department of Computer Science, University of Information Technology, Lahore 54890, Pakistan
[3]Department of Economics and Management Sciences, Faculty of Law, Economics and Social Sciences, Hassan II University of Casablanca, Mohammedia 20000, Morocco
[4]Social Things srl, 20135 Milan, Italy

Corresponding author: Sahar Shah (s.shah19@campus.unimib.it)

This work was supported by the Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy.

**ABSTRACT** Continual learning (CL) with bidirectional encoder representation from transformer (BERT) and its variant Distil-BERT, have shown remarkable performance in various natural language processing (NLP) tasks, such as text classification (TC). However, the model degrading factors like catastrophic forgetting (CF), accuracy, task dependent architecture ruined its popularity for complex and intelligent tasks. This research article proposes an innovative approach to address the challenges of CL in TC tasks. The objectives are to enable the model to learn continuously without forgetting previously acquired knowledge and perfectly avoid CF. To achieve this, a task-independent model architecture is introduced, allowing training of multiple tasks on the same model, thereby improving overall performance in CL scenarios. The framework incorporates two auxiliary tasks, namely next sentence prediction and task identifier prediction, to capture both the task-generic and task-specific contextual information. The Distil-BERT model, enhanced with two linear layers, categorizes the output representation into a task-generic space and a task-specific space. The proposed methodology is evaluated on diverse sets of TC tasks, including Yahoo, Yelp, Amazon, DB-Pedia, and AG-News. The experimental results demonstrate impressive performance across multiple tasks in terms of F1 score, model accuracy, model evaluation loss, learning rate, and training loss of the model. For the Yahoo task, the proposed model achieved an F1 score of 96.84 %, accuracy of 95.85 %, evaluation loss of 0.06, learning rate of 0.00003144. In the Yelp task, our model achieved an F1 score of 96.66 %, accuracy of 97.66 %, evaluation loss of 0.06, and similarly minimized training losses by achieving the learning rate of 0.00003189. For the Amazon task, the F1 score was 95.82 %, the observed accuracy is 97.83 %, evaluation loss was 0.06, and training losses were effectively minimized by securing the learning rate of 0.00003144. In the DB-Pedia task, we achieved an F1 score of 96.20 %, accuracy of 95.21 %, evaluation loss of 0.08, with learning rate 0.0001972 and rapidly minimized training losses due to the limited number of epochs and instances. In the AG-News task, our model obtained an F1 score of 94.78 %, accuracy of 92.76 %, evaluation loss of 0.06, and fixed the learning rate to 0.0001511. These results highlight the exceptional performance of our model in various TC tasks, with gradual reduction in training losses over time, indicating effective learning and retention of knowledge.

**INDEX TERMS** Continual learning, natural language processing, text classification, fine-tuning, Distil-BERT.

## I. INTRODUCTION

Sentiment Analysis (SA), the automated process of detecting sentiment or emotion in text data, has gained significant attention and finds many applications in various domains

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

like social media analysis, customer feedback analysis, and market research [1]. One of the fundamental tasks in SA is to categorize text documents into predefined sentiment categories such as positive, negative, or neutral [2]. However, ensuring accuracy and adaptability of SA models over the time is challenging due to the dynamic nature of textual data [3]. While TC is a critical area within NLP that holds immense significance across various domains, including SA, document categorization, and information retrieval. In TC techniques, SA are employed to determine the sentiment or emotion expressed in textual data [4], [5]. Document categorization involves organizing text documents into predefined categories based on their content, facilitating efficient information organization and retrieval. These applications demonstrate the broad impact and relevance of TC methods in enabling effective analysis and management of textual data across different fields [6]. With the increasing size and complexity of textual data, the need for scalable and adaptable TC models becomes crucial. However, several challenges hinder the development of effective CL approaches in this domain [7]. One significant challenge is the scalability of the proposed methods. TC tasks typically involve large datasets, requiring efficient handling of incremental updates to accommodate the continuous influx of new data. Managing the scalability of models is essential to ensure their effectiveness and efficiency in handling the ever-growing volume of textual data [8], [9].

CL, which enables models to learn from new data while retaining previous knowledge, is of utmost importance in TC and SA tasks. In these domains, sentiment and text patterns use constantly evolve, necessitating models to consistently adapt and refine their understanding of sentiments and text. This dynamic nature of SA underscores the need for CL approaches that can effectively capture evolving sentiment trends and maintain model performance over the time [10]. Indeed, the scalability of TC models is a crucial consideration for real-world applications dealing with large datasets, specially when one want to capture CL domain. In practical applications, where massive amounts of data need to be processed accurately and efficiently, the scalability of the model plays a significant role in ensuring its effectiveness and practicality. Future research should aim to address these scalability challenges to develop TC models that can handle large datasets and numerous classes while maintaining high performance and efficiency [11]. CF is a significant challenge in CL, as it can lead to the loss of previously acquired knowledge when learning new information. The objective of CL is to strike a balance between acquiring new knowledge and retaining the knowledge of earlier tasks. Mitigating CF is crucial to ensure that the model maintains performance on earlier tasks while adapting to new ones [12]. Developing effective strategies, such as regularization techniques, rehearsal methods, or knowledge distillation, can help alleviate this issue and enable CL models to preserve and transfer knowledge across tasks effectively. Addressing CF is a vital aspect of CL to achieve robust and adaptive TC models. The explicitly addressing how a TC model handles CF is a crucial role to ensure its ability to retain and apply knowledge from past tasks in a CL scenario [13].

The Distil-BERT is a compressed version of the BERT model, and is developed by researchers at Hugging Face to offer similar performance with a smaller size and faster speed. The left side component serves as an educational tool to explain the complex internal workings of the transformer layer in the BERT model, which is foundational for understanding how the entire model processes and transforms the input data to generate meaningful representations. This component includes several key elements. First, the multi-head attention mechanism starts with the input being projected into three different vectors: Query (Q), Key (K), and Value (V). This allows the model to focus on different parts of the input sequence simultaneously, capturing various relationships and dependencies within the data. The attention scores for each head are computed and then concatenated to form the final result. After the attention mechanism, the result is added back to the original input through a residual connection, which helps preserve the original information and stabilize the training process. Layer normalization is then applied to ensure the outputs have a stable mean and variance, which speeds up training and improves overall model performance. Next, the feed-forward neural network (FFNN) consists of two fully connected layers with a ReLU activation in between them. This part of the model applies non-linear transformations to the input data, enabling the capture of more complex patterns. The output of the FFNN is also added back to the input of this sub-layer through another residual connection and then normalized. The purpose of this component is to illustrate the structure of a single transformer layer within the BERT model, showing the detailed architecture and highlighting the key processes that occur within each layer. It provides a clear understanding of how the multi-head attention mechanism works together with the FFNN and residual connections. Additionally, it emphasizes the importance of layer normalization in stabilizing and improving the training process. It employs a transformer encoder architecture, featuring self-attention mechanisms and FFNN. Distil-BERT uses Word-Piece embedding to handle out-of-vocabulary words by breaking them into sub words. The model undergoes pretraining on unlabeled text and fine-tuning on labeled data for specific NLP tasks like TC [14]. It captures contextual information from both left and right contexts of words and can be applied to tasks such as TC and many more. Moreover, computational efficiency of Distil-BERT model is a vital consideration when deploying with TC in the real-world applications. The programmer should consider the model parameters that impact the practicality and scalability of the model's like training time, memory requirements, and inference speed [15]. These considerations are crucial for practical implementations of the TC, where efficient

training and inference processes are essential for real-time or resource-constrained environments. In the context of Distil-BERT, TC refers to training the model to classify text documents into different classes/categories using the fine-tuning process.

The first main goal of this paper, is to integrate the CL memory-based technique with the Distil-BERT model for the TC. This integration performs continuous learning in such a way that the model doesn't forget the previously learned task and avoid CF. The second main goal of the paper includes the designing of a task independent model through which we are able to train different several tasks on the same model. To achieve the aforementioned goals several modifications and steps are involved that have impressive performance in diverse NLP tasks specifically in TC. Our research centers on the application of pretrained Distil-BERT to the realm of CL in TC. In our proposed methodology, we deal with Distil-BERT model by adding the two linear layers on the head of Distil-BERT. We categorize the output representation of Distil-BERT into two stages, a task generic space and a task specific space. This is accomplished through the utilization of two auxiliary tasks: next sentence prediction, which facilitates the acquisition of task generic information, and task identifier prediction, which aids in acquiring task specific representations. By following this approach, Distil-BERT captures contextual information from both preceding and subsequent words. We train the model on five different tasks i.e, Yahoo, Yelp, Amazon, DB-Pedia and AG-News. The Distil-BERT model trains on each task individually and systematically save the best model in his memory, ensuring the preservation of learned information by avoiding CF. For each task during this mechanism the two added layers individually initialized from scratch with random weights this is what we called task specific space while the task generic space remains the same for each task. These two linear layers took vectors in 768 dimensions from each input task and then classify the provided input into number of classes respectively.

The rest of the paper is organized as, next to Introduction is section II, namely Literature review. We classify the literature review section into different topics of TC, involving different transformer based models with involving CL techniques. In this section, we captured the overall theme of the state of the art (SOTA) in a summary table. In section III, we have explained the proposed methodology that shows how we have designed this proposed novel approach, what methodologies and strategies we have adopted to achieve our goals. In section IV, we discussed datasets statistics, i.e., used epochs, classes, batch size and other specific details. The section V, is experimental discussion in which achieved results of the proposed model are discussed in detail. The second last section of the paper is VI- results comparison with SOTA, this shows the experimental comparison based on accuracy parameter between the presented approach and the approaches already existed in SOTA. The last section, VII, namely conclusion and future directions, deals with the conclusion of the overall proposed work and then suggest some future directions for more better advancements.

**Contributions:**

In this paper, we have made several significant and notable contributions in the field of TC compatible with CL approach using Distil-BERT model by tackling out the important and key challenges. Through presenting this novel approaches, these contributions enhance the existing body of the knowledge by addressing critical challenges in the field of TC and CL using Distil-BERT model. The key contributions of our proposed work can be summarized as below.

- We integrated of the memory-based CL technique with the Distil-BERT model for TC. This integration ensures that the model retains previously learned tasks and prevents CF by storing the best model for each task in memory, our approach preserves learned information.
- Utilized a pre-trained Distil-BERT model to extract meaningful representations (features) from text data. Distil-BERT's transformer layers capture contextual information from input text, enabling our model to classify text documents into different categories based on their contents.
- Proposed a novel task-independent architecture that effectively handles any sequences of the input tasks, ensuring better overall performance in CL scenarios. Our model can adapt to new tasks while retaining previously acquired knowledge.
- Optimized the framework by significantly reducing the computation running time, training losses, and increasing the learning rate of the proposed model. Through careful analysis and efficient algorithm, we achieved substantial improvements in the speed of our approach, and learning rate, these factors make the model more practical and scalable for real-world applications.
- We draw a comparison between the proposed designed and previously existing models in the SOTA based on model performance parameter, i.e., accuracy.
- Our proposed model presents significant advancements than the previously designed models in terms of accuracy. However, the other important and key model performance parameters i.e., F1 score, evaluation loss, and learning rate of the model with optimal results are only considered by the proposed model.

## II. LITERATURE REVIEW

Numerous studies in the field of CL for TC and SA using transformer based models have been accomplished. These studies have provided valuable insights into lifelong learning, incremental learning, and transfer learning approaches to TC based on transformer models. Techniques such as regularization, rehearsal, and distillation have been proposed to mitigate CF and retain knowledge from past aspects. The research findings contribute to the advancements of TC techniques that can handle evolving sentiment patterns and dynamic textual data in real-world applications. This section

is divided into different topics to enhance the readability, clarity and define the approaches used for each specific task. In addition, this article incorporates a brief table 1, that summarize the proposed techniques, achievements, applications, limitations and key parameters for each topic. By incorporating this well-organized table, the readability of the section is enhanced, and researchers can conveniently access the valuable information for their own research with respect to the models and techniques used, and gain a better understanding for more advancements.

### A. CLASSIC MODEL FOR TC IN DIL SETTINGS

The authors, in [16], proposed a novel model called continual and contrastive learning of aspect sentiment classification tasks (CLASSIC), for TC in domain incremental learning (DIL). CLASSIC operates in a DIL setting, eliminating the need for task information during testing. It leverages Adapter-BERT to incorporate pre-trained BERT without fine-tuning and addresses CF. The model introduces contrastive CL for knowledge transfer (KT) between tasks and distillation from old to new tasks, enhancing classification accuracy. CLASSIC consists of three sub-systems: contrastive ensemble distillation (CED), contrastive knowledge sharing (CKS), and contrastive supervised learning (CSL). The architecture is designed for aspect sentiment classification (ASC) in DIL, utilizing Adapter-BERT. During training, CLASSIC takes hidden states and a task ID, but during testing, no task ID is required. The model's outputs are task-specific features used for constructing a classifier. The framework follows contrastive learning principles and is termed contrastive CL.

### B. TC IN IOT

The Internet of Things (IoT) connects smart devices through the internet, generating vast amounts of textual data [17]. TC is a challenge in this context, and language models like BERT and Distil-BERT have shown promise in handling it. This study compares BERT and Distil-BERT for TC in English and Brazilian Portuguese using different datasets. This highlights the importance of dataset balance and its impact on model performance, with unbalanced datasets showing lower accuracy. Additionally, the use of lightweight models like Distil-BERT allows for efficient execution on low computational resources while maintaining performance comparable to larger models. The experimental results indicate that Distil-BERT is 40 % smaller, 45 % faster, and retains 96 % language comprehension skills compared to BERT. The study highlights the effectiveness of Distil-BERT in different languages and emphasizes the importance of dataset quality.

### C. CL IN ASC TASKS VIA BERT BASED MODEL

The paper [18] focuses on the topic of CL in the context of ASC tasks. While previous CL techniques have been proposed for document sentiment classification (SC), research specifically targeting CL in ASC is limited. The authors introduce BERT based CL (B-CL), a new CL system that addresses two key challenges in ASC with CL. These challenges include transferring knowledge from previous tasks to facilitate better model learning, and maintaining performance on previous tasks to prevent forgetting or degradation. B-CL is a capsule network-based model that incorporates forward knowledge transfer (FKT) and backward knowledge transfer (BKT) mechanisms. It effectively utilizes knowledge gained from previous tasks to improve performance on both new and old tasks. The model utilizes a novel building block called continual learning adopter (CLA), inspired by Adapter-BERT. CLA employs capsules and dynamic routing to identify similarities between previous and new tasks, facilitating the transfer of shared knowledge. Task masks are employed to protect task-specific knowledge and prevent CF. Extensive experiments are conducted to validate the effectiveness of B-CL, comparing it with various baselines. The results demonstrate the superior performance of B-CL in ASC with CL scenarios. This paper contributes by highlighting the need for CL approaches in ASC and proposing the B-CL model, which incorporates the CLA into a pre-trained BERT model, enabling effective ASC in CL.

### D. SUPERVISED CONTRASTIVE LEARNING FRAMEWORK FOR ABSA BASED ON BERT

In [19], the authors introduce a novel approach to aspect-based sentiment analysis (ABSA) by focusing on improving sentiment prediction for unknown testing aspects. They address this challenge by leveraging sentiment features and propose a BERT-based supervised contrastive learning framework. The main contributions of this research are twofold. Firstly, they approach ABSA from a new perspective, emphasizing the enhancement of SA for unknown testing aspects by leveraging sentiment features. Secondly, they introduce the BERT-SCon framework, which utilizes supervised contrastive learning to distinguish sentiment features based on sentiment polarity and pattern. The proposed BERT-SCon framework achieves SOTA performance on five benchmark datasets, demonstrating the effectiveness of the approach. The architecture of the framework consists of four components: data augmentation, feature extractor using a pre-trained BERT model, SC with a soft max function, and contrastive learning to bring together representations of the same sentiment polarity/pattern and differentiate representations from different classes.

### E. NLP AND ML FOR CLASSIFICATION OF FURNITURE TIP-OVER INCIDENTS

This paper proposes an improved method for classifying furniture tip-over incidents using a combination of NLP techniques and machine learning (ML) algorithms. The proposed model architecture is based on a pretrained RoBERTa model, enhanced with layer normalization, dropout layers, and a linear classifier [20]. The study compares the proposed model with other transformer-based models like BERT,

RoBERTa, DeBERTa, ALBERT, DistilBERT, and MPNet. The models were trained on injury narratives from the united states-consumer product safety commission (U.S-CPSC) dataset, addressing challenges such as imbalanced classes and domain-specific jargon. The text data was preprocessed, tokenized, and encoded for input into the models. The computational complexity of the proposed model is estimated based on the number of attention layers and linear operations. The study found that the proposed model achieved improved classification results compared to the default transformer-based models, showcasing its potential for streamlining classification tasks in various datasets. The experimental analysis demonstrated that the use of machine learning techniques can reduce human effort and enhance efficiency in reviewing and classifying incident reports.

### F. TRANSFER LEARNING IN TC FOR COVID-19

TC is a widely studied problem in information retrieval and data mining. It has applications in various domains such as healthcare, marketing, entertainment, and content filtering. Researchers have recently focused on developing automated systems for TC using NLP and data mining techniques. NLP enables the categorization of documents with different types of texts, and social media platforms generate vast amounts of data for experimentation. Transfer learning, a technique that leverages knowledge from unlabeled data for tasks with limited labeled data, has gained attention in TC. In this study [21], transfer learning classification models are applied to coronavirus disease (COVID-19) fake news and extremist-non-extremist datasets. The researchers emphasize the potential of transfer learning to improve accuracy with less human supervision compared to active and supervised learning. NLP transformers, particularly the attention-based transformer models, have shown promising accuracy in various applications. The study demonstrates the effectiveness of transformers in predicting real and fake news related to COVID-19. Fake news dissemination through social media platforms is a significant concern, and distinguishing between real and fake news is challenging. Existing approaches have limitations, motivating the development of hybrid methods. The study applies nine transfer learning models to COVID-19 datasets and evaluates their performance using metrics. Reliable repositories are used for data collection, and the results highlight the effectiveness of transfer learning models in binary TC.

### G. MTL MODELS FOR PEER REVIEW COMMENTS

This paper introduces two model transformation learning (MTL) models, leveraging BERT and Distil-BERT, for evaluating peer-review comments [22]. The models detect multiple features simultaneously, improving performance and reducing model size compared to previous methods. MTL enhances data efficiency and can be seen as a form of inductive transfer learning. BERT and Distil-BERT, pre-trained language models, are effective tools for NLP tasks. The study demonstrates the superiority of BERT-based models over GloVe-based ones and suggests deploying the MTL-BERT model for high accuracy or the MTL-Distil-BERT model for resource efficiency on peer-review platforms. The MTL models proposed in this study consider three features of high-quality peer reviews, containing suggestions, mentioning problems, and using a positive tone. The comparison between BERT-based single task learning (STL) models and the previous GloVe-based method demonstrates significant improvements in detecting a single feature. The study acknowledges limitations, such as the need to explore additional tasks, consider alternative parameter sharing approaches, and evaluate the model in real-world settings. These findings lay the groundwork for ongoing work to comprehensively evaluate peer review comments and enhance peer assessment.

### H. MULTIMODEL-DEEP LEARNING FOR SHORT-TEXT MULTI-CLASS CLASSIFICATION

This paper presents a multimodel-based deep learning framework for short-text multi-class classification with an imbalanced and minimal dataset [23]. The framework consists of an encoder layer using Distil-BERT for dynamic word embeddings, followed by word-level and sentence-level long short term memory (LSTM) networks to extract deep semantic information. A max-pooling layer reduces dimensionality, and a Soft Max layer performs multi-class classification. The proposed approach achieves SOTA performance while being faster and lighter for deployment on mobile devices. Distil-BERT reduces complexity, and the bidirectional LSTM (BI-LSTM) enhances the model's ability to handle polysemous words. The framework addresses data imbalance, small text, and multi-classification tasks and is applicable to real-world scenarios, particularly mobile devices. The smaller model size makes it suitable for Artificial Intelligence (AI) technology in smart devices, while BERT-based models are more suited for large-scale cloud computing and research institutes.

### I. CSIC MODEL FOR SC TASKS

The research article introduces a novel approach called contrastive supervised learning with iterative combination (CSIC) for SC tasks [24]. The objective of CSIC is to overcome the limitations of static models that cannot adapt to new domains due to storage constraints or privacy concerns. The proposed CSIC method combines the original network, trained on old tasks, with a fine-tuned network trained on new tasks using knowledge distillation. This iterative combination allows for CL without increasing the network's size. BERT, a highly performant model in SC, is chosen as the backbone model for CSIC. To address CF, where the network's performance on old tasks deteriorates when learning new tasks, CSIC linearly combines the original network and fine-tuned network using knowledge distillation. Importantly, after training, the combined network can be

converted back to the standard BERT structure, eliminating the need for additional parameters or structures for old and new tasks. CSIC consists of four network components: the original network for old tasks, the fine-tuned network for new tasks, the middle network that integrates knowledge from both networks, and the final combined network converted from the middle network. All networks utilize BERT as the backbone. During the learning of a new task, CSIC involves three phases: linearly combining the original and fine-tuned networks to create the middle network, performing an additional retraining phase for the middle network to prevent CF, and converting the middle network into the final combined network with the same size as the original network. The effectiveness and efficiency of the CSIC approach are evaluated through extensive experiments on 16 popular review datasets. The results demonstrate that CSIC outperforms strong baseline methods for continual SC.

### J. CL FOR TC IN BERT MODEL

The article titled ''Addressing CL Challenges in TC through Information Disentanglement based Regularization'' addresses the notable issue of CL in the context of TC tasks [25]. CL entails a model's ability to learn from an ongoing stream of data while retaining previously acquired knowledge, without experiencing CF. To tackle this challenge, the researchers propose an innovative approach that utilizes regularization based on information disentanglement. The primary goal is to enhance the model's CL capabilities, enabling effective learning from new data classes while preserving the knowledge of previously learned classes. To accomplish this, the proposed method introduces a regularization term into the loss function of the TC model. This regularization term encourages the model to disentangle shared information, which is relevant across different tasks, from task-specific information, which pertains to each specific classification task. By disentangling this information, the model can better isolate and retain crucial knowledge for specific tasks, preventing interference or forgetting when learning new tasks. To evaluate the effectiveness of the proposed approach, the researchers conducted experiments on various well-established benchmark datasets. The results demonstrate that the information disentanglement based regularization significantly enhances the model's performance in CL scenarios. By mitigating CF, the approach enables the model to maintain its performance on previously learned tasks while adapting well to new tasks. Furthermore, the proposed method achieves competitive accuracy compared to SOTA for CL methods, highlighting its effectiveness in the domain of TC.

### K. NEURAL LANGUAGE MODELS (NLMS)

In [26], authors explores how NLMs, like Transformers, convolution neural networks (CNNs) and LSTMs, understand and process verb argument structures in German, a language where word order is flexible in subordinate clauses. The researchers developed a new testing method using minimal variation sets. These sets include sentences with correct and incorrect verb structures to see if the models can tell the difference between grammatically right and wrong sentences. Transformers generally scored higher than LSTMs and even humans but tended to overgeneralize, sometimes accepting sentences that were not very plausible. LSTMs had trouble, especially with frequent argument structures like double nominatives. Human evaluations were also part of the study. Annotators rated the naturalness of sentences on a scale, providing a benchmark for the models. These ratings confirmed that grammatical rules, such as nominative and dative alignment, play a role in how sentences are judged. The study showed that while NLMs can understand some syntactic rules, their performance is affected by biases and overgeneralization. In their experiments, the team created sentences by changing the positions and case assignments of arguments in template sentences, resulting in both acceptable and unacceptable variations. They found that both LSTMs and Transformers performed better than random guessing in identifying correct verb structures. This suggests a need for better models or training methods that more closely mimic human language understanding.

### L. CORRECTNESS OF SENTENCE VIA LANGUAGE MODELS

The authors introduced the ItaCoLA corpus in [27], a collection of almost 10,000 Italian sentences, each labeled as acceptable or unacceptable. The aim is to help researchers study how well language models can judge the grammatical correctness of sentences in languages other than English. The authors, explain how they created the ItaCoLA corpus by manually transcribing sentences from various linguistic sources and labeling them based on acceptability. About 30% of these sentences are also tagged with additional linguistic features to capture specific grammatical phenomena. The study also explores using a multilingual model, XLM-RoBERTa, which is trained on multiple languages. This model benefits from the multilingual training, but is still not as effective as models trained specifically on one language. They described several experiments to test the performance of neural language models like BERT on this new corpus. They compare how well these models perform on Italian sentences versus English sentences from a similar corpus (CoLA). The experiments include tasks within the same domain (in-domain) and tasks with different data (out-of-domain). The findings show that BERT-based models work well for Italian sentences, almost as well as for English, though there are some differences in handling certain grammatical structures.

### M. DATASET FOR LANGUAGE MODELS

A new dataset called EsCoLA, which contains 11,174 Spanish sentences labeled as either acceptable or unacceptable, is presented in [28]. This dataset is designed to help researchers test how well language models can understand and generate grammatically correct Spanish sentences. The sentences were taken from well-known Spanish grammar

books and include annotations for various grammatical features, such as sentence structure, verb use, and specific Spanish language phenomena like the use of 'ser' and 'estar.' Human experts also evaluated the sentences to provide a benchmark for the models' performance. Their judgments matched the dataset labels most of the time, but there were some disagreements, which the researchers noted. This human evaluation helps show the upper limit of how well we can expect the models to perform. The researchers ran several experiments to see how well different language models (IXABERTesv2, RoBERTa-large-bne, XLM-RoBERTa-large, and mDeBERTa-v3) performed when fine-tuned on the EsCoLA dataset. They tested these models on both sentences from the same sources used to create the dataset (in-domain) and sentences from different sources (out-of-domain). The mDeBERTa-v3 model performed the best, followed by RoBERTa-large-bne. The other models, XLM-RoBERTa-large and IXABERTesv2, did not perform as well.

## III. PROPOSED METHODOLOGY

We utilized a pre-trained Distil-BERT model to fine-tune for TC tasks, with our primary focus on CL. This section is divided into subsections to describe the proposed methodology with a clear vision. The architecture of Distil-BERT and its components is depicted in figure 1. While the overall proposed model architecture and flow-chart are shown in figures 2 and 3 respectively.

### 1) PRE-PROCESSING ON DATASETS

The data sets or tasks are not ready to use, directly. Rather, it contains extra characters, punctuations, padding, truncating, and require tokenizer, handling special characters etc., that we do not need for the next steps. So, the pre-processing is the very first step that perform on tasks data to convert the task's data to our desired format. In our case, we have five different tasks i.e., yahoo (Tweets), Yelp (Tweets), Amazon (Tweets), DB-Pedia (Tweets) and AG-News (Tweets). The first step typically involves, cleaning, tokenization, lemmatization or stemming, handling abbreviations and acronyms, handling rare words and at last encoding and vectorization. In text cleaning which we have removed unnecessary characters or symbols, such as special characters, punctuation marks and trailing white spaces. In the tokenization step, we have split the text into individual words or subwords called tokens. In lemmatization or stemming, we reduce words to their base or root form. For abbreviations and acronyms, we expanded the words to their full forms to ensure consistent representation and improve understanding. Next, we performed, removing or replacing rare words with a special token that help us in reducing noise and improving model performance. At last, in encoding and vectorization stage, the textual data converted to numerical representation through word embedding technique (Word2Vec). Shortly, in pre-processing step we have removed all the unnecessary characters, punctuations, replacing etc., and now the task's data are ready to use for next steps.

### 2) DATA SPLITTING

The pre-processed data is then fed into the data splitting block. This block divide the pre-processed data into two types of data sets, i.e., training set, and test set. The training set is the portion of the considered datasets that contains labeled samples that are used to train the Distil BERT model on tweets with their associated labels to estimate the model's parameters. The Distil BERT model learns from the training sets by adjusting the weights/coefficients of the neurons. The goal of training the Distil BERT model is to minimize the difference between the predicted outputs of the text and the true labels of the text for the training.

The second category, of data splitting, is testing set. In this category, the datasets assess the performance and generalization capability of the Distil BERT model trained on tweets present in datasets. The testing set contains labeled samples that the Distil BERT has not seen during the training phase. The testing set is used to evaluate the model's performance of the Distil BERT on unseen data and estimate how well it is performs in TC. The goal is to assess the model's ability to generalize its learned patterns and make accurate predictions on new tweets.

In output, for training the model we obtained training matrices block represents the output data generated during the training phase, including matrices for F1 score, accuracy, and learning rate to monitor the model's progress. Similarly, the test matrices block contains evaluation matrices obtained when the trained model is applied to the test set, providing insights into the model performance.

### 3) FEATURES EXTRACTION AND TEXT CLASSIFICATION

Features extraction involves utilizing a pre-trained Distil-BERT model to extract meaningful representations (features) from text data without further training the model on a specific task. The TC is a specific task in NLP that involves categorizing or assigning predefined labels or categories to a given piece of text. In our case, we have five different tasks i.e., yahoo, Yelp, Amazon, DB-Pedia and AG-News we did feature extraction, and TC on these tasks. The general block diagram of TC with feature engineering is shown in figure, 4.

The goal is to automatically classify text documents into different classes or categories based on their content. The Distil-BERT model, specifically, consists of six transformer layers that perform computations to extract features and capture contextual information from the input data. To begin, the integer IDs are transformed into embeddings, dense vector representations that encode the meaning of words in the input text. These embeddings carry semantic information, reflecting the relationships between words within the sentence context. These embeddings, along with the position embeddings that indicate word positions within the input sequence, are passed through the transformer layers. Each transformer layer consists of two sub-layers: the multi-head self-attention mechanism and a FFNN. The self-attention mechanism enables the model to assign varying

**TABLE 1.** Comparative analysis.

| Keywords | Proposed Techniques | Achievements | Applications | Limitations |
|---|---|---|---|---|
| CL, ASC, DIL, KT, CED, 2021, [16] | CLASSIC model is used to address SC in the context of DIL for ASC. | Can learn and adopt from new data without forgetting the previously learned data. | Can be used whenever the objective is to enhance the classification's accuracy. | Dependence on contrastive learning to reduce its effectiveness in handling new tasks. |
| Distil-BERT, Transformer, Big Data, 2022, [17] | Proposed two language models (English, Brazilian) for TC using various datasets. | Offers comparable performance to BERT while being significantly smaller and faster. | Distil-BERT for TC tasks in the context of the IoT. | It faces challenges with memory limitations on devices with very low resources. |
| Adapter BERT, CL, B-CL, ASC Transfer learning, 2021, [18] | Continually learn in ASC for transferring knowledge from previous tasks to avoid degrading and forgetting. | Transfer the knowledge in an efficient and effective way without forgetting the learned data. | Can be fit where enhancing model learning in ASC with CL systems is required. | The model performance drops for very different tasks. |
| Sentiment prediction, SC, BERT model, 2021, [19] | ABSA leverages supervised contrastive learning based on BERT is proposed to extract sentiment characteristic from sentence related to specific aspect. | Improved the sentiment prediction for unknown testing aspect, integration of BERT model, enhanced ABSA performance. | Needed where enhancing SA performance by leveraging sentiment features and the BERT-SCon framework is required. | Using supervised contrastive learning needs a lot of labeled data, which can be hard to get. |
| Text Analysis, Transformers, Pattern Classification, NLP, 2022, [20] | An enhanced method for furniture tip-over incident classification, utilizing NLP techniques and ML algorithms is proposed. | Outperforms for transformer-based models in classifying furniture tip-over incidents. | Broader applicability in automating classification tasks for various domains, improving productivity and accuracy. | The model does not work well for tasks outside of furniture tip-over incidents because it is trained specifically for that domain. |
| TC, Transfer Learning, Accuracy, 2022, [21] | Explores the application of transfer learning classification models to COVID-19 fake news on extremist and non-extremist datasets. | Improves TC accuracy with less human supervision compared to active and supervised learning approaches. | Applications in domains such as healthcare, marketing, entertainment, and content filtering. | This method does not work well for very specific tasks or when there is a lot of noisy data. |
| Peer Assessment, Text Analysis, Data Mining, 2021, [22] | Two multitask learning (MTL) models, utilizing BERT and Distil-BERT, for evaluating peer-review comments are proposed. | Demonstrates the superiority of BERT-based models over GloVe-based. | The MTL-BERT model can be deployed on peer-review platforms for high accuracy in evaluating peer-review comments. | The multitask learning method can be less effective when tasks are not closely related, leading to negative transfer. |
| Distil-BERT, Multi-Classification Tasks, LSTM, 2022, [23] | Multimodel-based deep learning framework is presented for short-text multi class classification with an imbalanced and minimal dataset. | Effective in addressing data imbalance, small text, and multi-classification tasks. | Best for real-world scenarios, particularly on mobile devices, where efficient short-TC is required. | The model struggles with very imbalanced datasets and still needs adjustments to fit specific hardware limitations. |

**TABLE 1.** *(Continued.)* **Comparative analysis.**

| Keywords | Proposed Techniques | Achievements | Applications | Limitations |
|---|---|---|---|---|
| Continual SC, Fine-tuned Network, BERT, CF, 2021, [24] | CSIC addresses limitations of static models in SC, enabling adaptability to new domains. It combines an original network with a fine-tuned network using knowledge distillation for CL. | Adaptability to new domains, CL without increasing the size of the network, Mitigation of CF, Utilization of BERT model. | CSIC can be applied in scenarios where storage constraints or privacy concerns prevent the use of large static models. | Combining networks repeatedly can make training complicated and possibly unstable. The success of this method depends on the quality of the knowledge distillation process. |
| CL, TC, CF, Knowledge Retention, Accuracy, 2021, [25] | Addressing the challenge of CL in TC tasks by leveraging information disentanglement. | The approach enhances the model's ability to learn from new classes while retaining knowledge of previously learned ones. | Has various potential applications in domains where models need to continuously learn from new data. | This model makes the training process more complex and slow. This method also needs careful adjustment to keep old knowledge while learning new information. |
| NLMs, Transformers, LSTMs, 2019, [26] | Developed a testing method using minimal variation sets, creating sentences with both correct and incorrect verb structures by altering positions and case assignments. | Transformers generally performed better than LSTMs and humans in recognizing correct verb structures. | Valuable for analyzing how different neural models understand and process syntactic rules in languages with flexible word order, such as German. | Transformers tended to overgeneralize, sometimes accepting sentences that were not grammatically correct. LSTMs, on the other hand, had difficulties with frequent argument structures like double nominatives. |
| ItaCoLA corpus, Italian sentences, acceptability judgment, NLMs, 2021, [27] | Developed the ItaCoLA corpus, a collection of nearly 10,000 Italian sentences labeled for grammatical acceptability. | Demonstrates that BERT-based models can perform well on Italian sentences, almost matching their performance on English sentences. | This resource is beneficial for linguistic research, particularly in the area of syntactic and grammatical analysis. | Multilingual models like XLM-RoBERTa, although beneficial, do not yet match the effectiveness of models trained on a single language. |
| EsCoLA, Spanish sentences, grammatical acceptability, language models, 2024, [28] | The EsCoLA dataset is created by sourcing 11,174 Spanish sentences from well-known grammar books, annotated for grammatical features and labeled as acceptable or unacceptable. | The mDeBERTa-v3 achieved the best performance, followed by RoBERTa-large-bne, demonstrating the effectiveness of these models. | It is useful for linguistic research, especially in syntactic and grammatical analysis of the Spanish language. | Improvements needed in training and fine-tuning language models to better capture the nuances of Spanish grammar. |

levels of importance to words in the input sequence by considering their relevance to one another. It produces weighted representations of the words, emphasizing the most contextually significant words for each position in the sequence. Subsequently, the FFNN processes the self-attention outputs. It applies non-linear transformations to the representations, capturing higher-level features and interactions among the different words in the sequence. This process of feeding the input through the transformer layers is repeated iteratively, typically six times for Distil-BERT. With each iteration, the model refines and extracts increasingly complex features from the input data. The two added linear layers on the head of Distil-BERT are the classification layers, responsible for mapping the representations generated by the transformer layers to the specific number of classes relevant to the TC task. By utilizing these extracted features, the

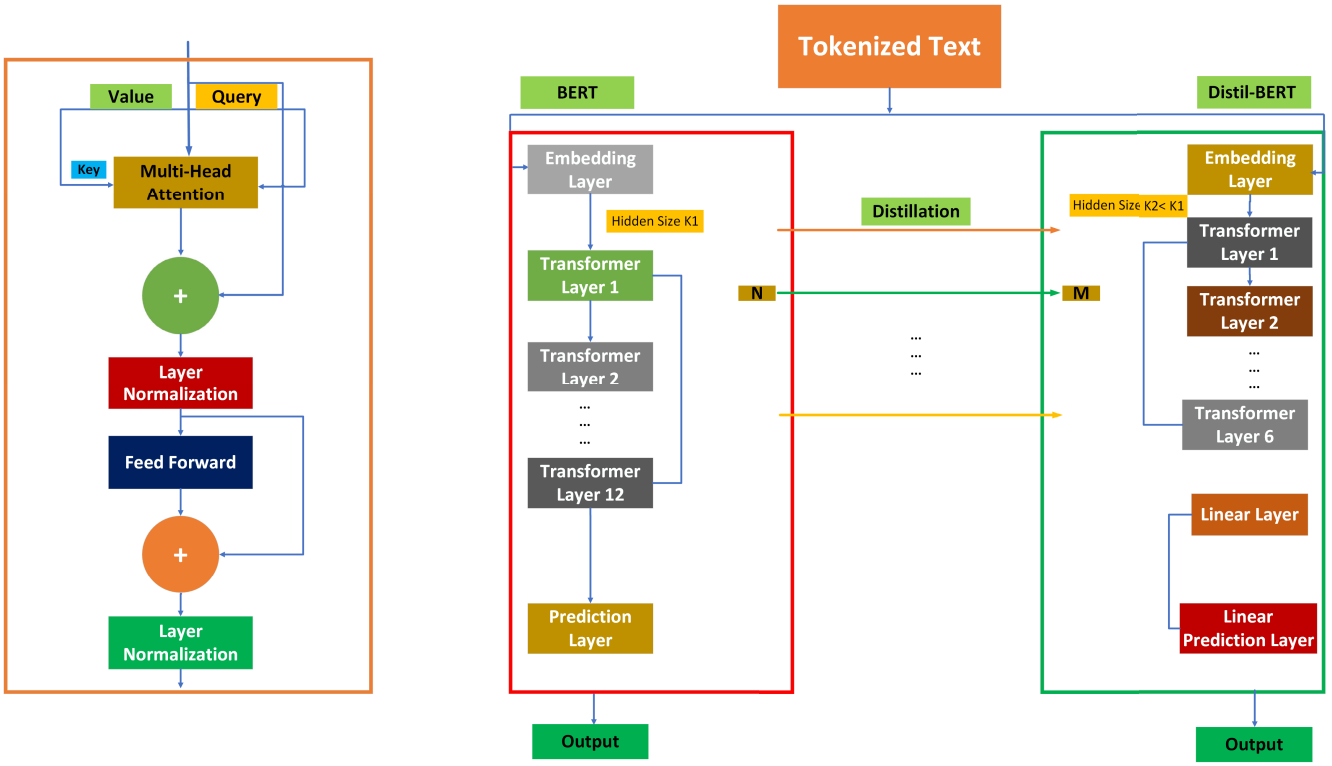


**FIGURE 1.** Distil BERT model architecture and components.

two layers make predictions about the class labels associated with the input text. For the input datasets i.e., Yahoo, Yelp, Amazon, DB-Pedia and AG-News our Distil-BERT model classifies 10, 5, 5, 14, and 4 classes respectively.

$$\mathbf{y} = W_1\mathbf{z} + W_2\mathbf{x} + \mathbf{b} \tag{1}$$

In equation 1, the variable term y represents the task-specific space representation, the term z represents the task-generic space representation and x represents the input features. While, $W_1$ and $W_2$ are the weight matrices that transform the task-generic space representation and input features, respectively. Moreover, **b** represents the bias vector and $f(z, y; \theta)$ represents the output of the model which is a function that takes the task-generic space representation z, the task-specific space representation y, and the model parameter $\theta$ as inputs, and produces an output.

#### 4) MEMORY-BASED CONTINUAL LEARNING TECHNIQUE

This proposed idea presents a novel memory-based CL approach to address the issue of CF in TC tasks. The proposed methodology utilizes the power of pre-trained Distil-BERT and enables the incorporation of new tasks while retaining knowledge from previously learned and saved tasks. To implement memory-based CL, the Distil-BERT model is trained on five tasks. This approach involves training the Distil-BERT model sequentially on multiple text classification tasks. Throughout the training process, the best-performing model for each task is systematically saved in memory to ensure that previously learned knowledge is retained and can be reused. The architecture of Distil-BERT is enhanced with two additional linear layers. These layers are crucial in dividing the output representation into a task-generic space and a task-specific space. The task-generic space is designed to capture information that is common across all tasks, providing a stable foundation that supports general understanding. On the other hand, the task-specific space captures unique characteristics relevant to each individual task. For each new task, the task-specific layers are initialized with random weights, allowing the model to adapt and learn new information without disrupting the task-generic knowledge. Auxiliary tasks play a vital role in this approach. Two auxiliary tasks are incorporated during training: next sentence prediction and task identifier prediction. The next sentence prediction task helps the model capture task-generic information by training it to predict the logical sequence of sentences. The task identifier prediction task aids in capturing task-specific information by requiring the model to identify which task the input data belongs to. By training the model on these auxiliary tasks alongside the main text classification tasks, the model can simultaneously learn task-generic and task-specific contextual information. During the training process, the task-generic layers remain consistent, ensuring that the model retains the knowledge acquired from previous tasks. The task-specific layers, however, are updated for each new task, allowing the model to adapt to new challenges while preserving the integrity of

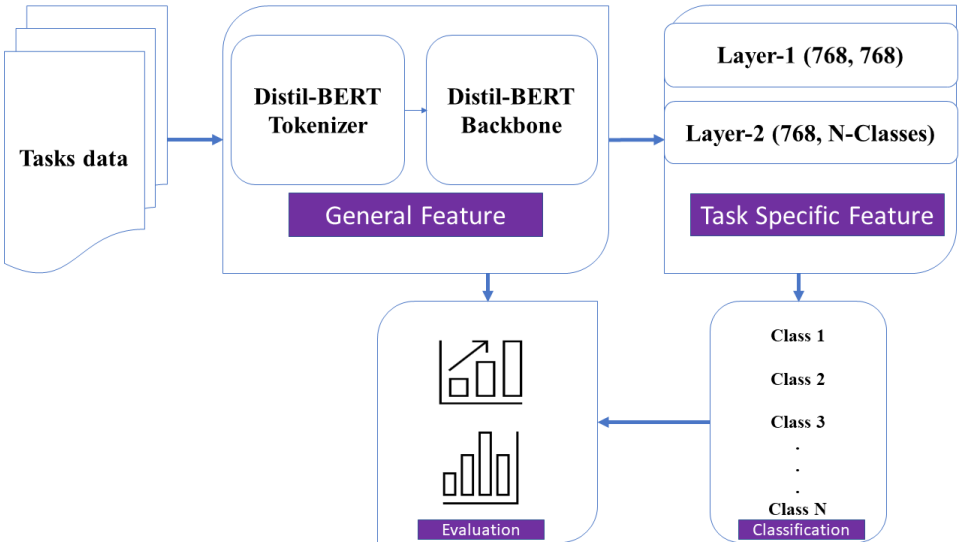


**FIGURE 2.** Proposed model architecture.

previously learned information. This strategy ensures that the model can accurately perform on previously learned tasks even when new tasks are introduced, effectively preventing catastrophic forgetting. The systematic saving of the best models and the incorporation of auxiliary tasks enable the model to maintain high performance and continually adapt to new tasks without losing previously acquired knowledge.

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta\nabla_{\theta_{\text{old}}}\mathcal{L}(\theta_{\text{old}}, \mathcal{D}_{\text{old}}) - \eta\nabla_{\theta}\mathcal{L}(\theta, \mathcal{D}_{\text{new}}) \quad (2)$$

Equation 2, explain the concept of CL memory-based approach, in a formulation for updating the model's parameters with new tasks while preserving knowledge from previously learned tasks. $\theta$, represents the all parameters of the model, and $\theta_{\text{old}}$ represent the parameters of the model before updating with new tasks. Additionally, $\mathcal{D}_{\text{new}}$ represent the dataset for the new task, and $\mathcal{D}_{\text{old}}$ represent the datasets for previously learned tasks. "$L(\theta_{\text{old}}, \mathcal{D}_{\text{old}})$" represents the loss function computed on the datasets for previously learned tasks using the parameters $\theta_{\text{old}}$. $L(\theta, \mathcal{D}_{\text{new}})$ represents the loss function computed on the dataset for the new task using the updated parameters $\theta$.

"$\eta$" represents the learning rate, which controls the size of the parameter updates.

#### 5) FINE-TUNING

Fine-tuning refers to the process of taking a pre-trained language model, in our case Distil BERT, and further training it on a specific task or domain. For each new task, we employ the same Distil-BERT backbone, complemented by two additional linear layers serving as the classification head. The second-to-last layer incorporates 768 input features and 768 output features, while the final layer comprises 768 input features, with the output features configured according to the number of classes specific to the task. The addition of two extra linear layers at the top of the Distil-BERT model divides the output representation into a task-generic space and a task-specific space. To achieve this division, auxiliary tasks are incorporated during training. One auxiliary task is next sentence prediction, which captures task-generic information, while the other is task identifier prediction, which captures task-specific representations. By training the model on these auxiliary tasks alongside the main TC tasks, the model can capture both task-generic and task-specific contextual information.

## IV. DATASETS STATISTICS

In this paper, we have considered tweets of five different tasks i.e., Yahoo, Yelp, Amazon, Dp-Pedia, AG-News for experimental analysis. The considered different statistics i.e., classes, epochs, batch size, types of tweets, trains and tests set of each dataset for experiments are explained below and shown in Table 2. The flexibility of our approach allows tasks to be added in any order, demonstrating robustness to the sequence of introduction.

### A. YAHOO

The dataset Yahoo is a collection of user-generated reviews and ratings for various products and services. It covers a wide range of domains, including electronics, movies, books, and more. The dataset consists of textual reviews along with associated ratings or sentiment labels, indicating the sentiment expressed in the review (positive, negative, or neutral). It is a popular dataset used for SA and opinion mining tasks. In our case, the Yahoo dataset consists of 10 classes, representing different categories or topics for the tweets format. The considered number of epochs for Yahoo are 14 with batch size 16. It includes a training set with 20,000 instances, which are used to train the model, and a test set with 7,600 instances, which are used to evaluate the model's performance.

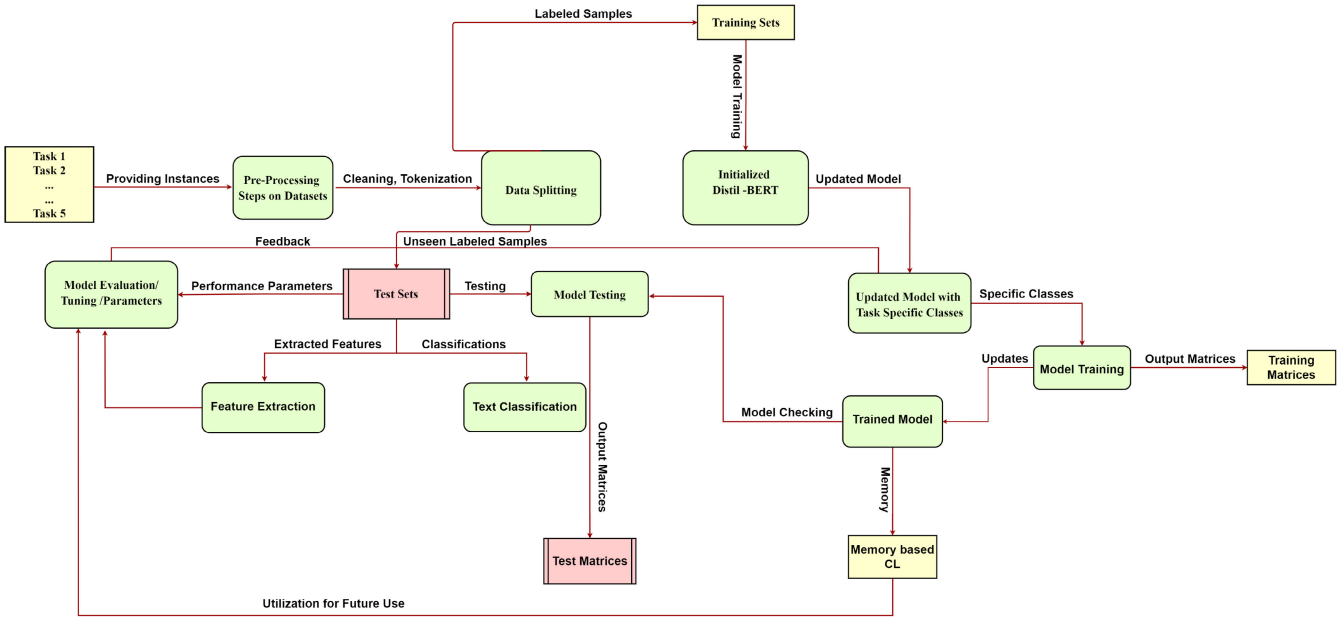


**FIGURE 3.** Flowchart of the proposed work.

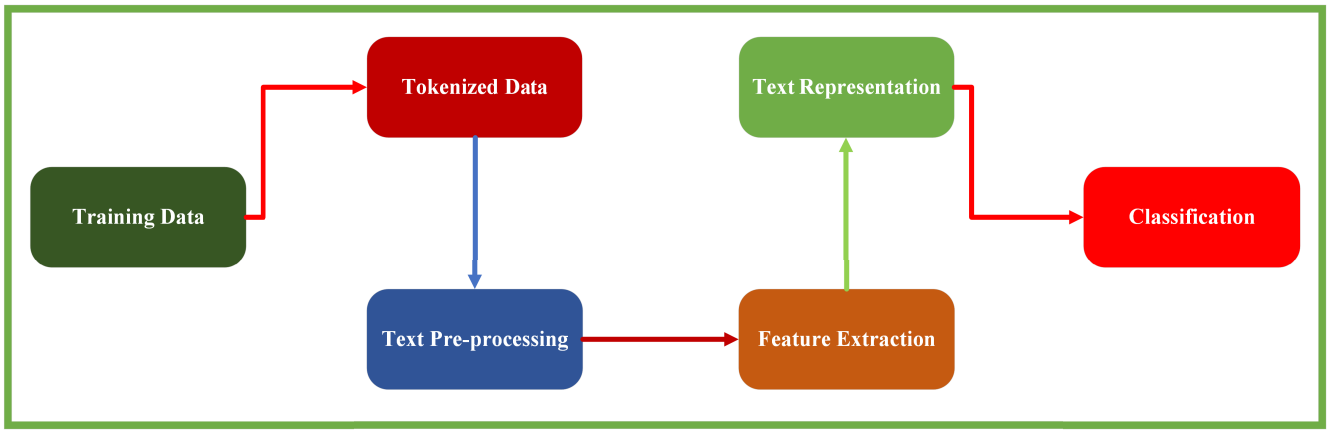


**FIGURE 4.** Text classification with feature engineering.

### B. YELP

The dataset Yelp comprises user-generated reviews for local businesses, such as restaurants, bars, and hotels.

The dataset include classes or labels that represent different aspects or sentiment categories, such as food quality, service, ambiance, etc. It consists of separate training and test sets, where the training set is used for model training, and the test set is used for evaluation. This dataset contains 14 epochs with 5 classes, representing different sentiment tweets, such as positive, negative, or neutral tweets. It consists of a training set with 10,000 instances for model training and a test set with 7,600 instances for model evaluation.

### C. AMAZON

The Amazon dataset contains customer reviews and associated ratings for products sold on the Amazon platform. It includes separate training and test sets, where the training set is used for model training, and the test set is used for evaluation. This dataset consists of 14 classes, representing various topics or concepts of tweets derived from Wikipedia, with 14 epochs. It includes a training set with 28,000 instances and a test set with 7,600 instances used for model training and evaluation.

### D. DB-PEDIA

The dataset DB-Pedia is derived from Wikipedia and consists of textual descriptions and associated labels or categories. The dataset typically includes a set of predefined classes representing different topics or concepts in tweets, such as people, places, organizations, etc. In our experiments, we have considered the DB-Pedia dataset that consists of 14 classes and 2 epochs, representing various topics or

**TABLE 2. Dataset statistics.**

| Datasets | Classes | Epochs | Batch Size | Types | Trains | Tests |
|---|---|---|---|---|---|---|
| Yahoo | 10 | 14 | 16 | Text | 20000 | 7600 |
| Yelp | 5 | 14 | 16 | Text | 10000 | 7600 |
| Amazon | 5 | 14 | 16 | Text | 10000 | 7600 |
| Dp-Pedia | 14 | 2 | 16 | Text | 28000 | 7600 |
| AG-News | 4 | 3 | 16 | Text | 8000 | 76000 |

concepts derived from Wikipedia. It includes a training set with 28,000 instances used for model training and a test set with 7,600 instances for evaluation.

### E. AG-NEWS

The dataset AG-News comprises news articles categorized into different topics or classes, such as sports, business, technology, and world news. This dataset includes predefined classes representing the different news categories with text. This dataset contains 3 epochs and 4 classes, representing different news categories. It comprises a training set with 8,000 instances for model training and a larger test set with 76,000 instances for evaluation of tweets.

The benefits achieved using diverse datasets are significant. The selected datasets cover a wide range of domains, including reviews, news articles, and product ratings, which helps create a robust model capable of handling various types of textual data. Datasets like Yelp and Amazon, which focus on user reviews and sentiments, enable the model to excel in SA tasks, improving its ability to detect and classify sentiments accurately. Additionally, using datasets with different classes and topics, such as DB-Pedia and AG-News, ensures that the model generalizes well across various types of content, making it more versatile and effective in real-world applications.

The division of each dataset into training and test sets allows for rigorous evaluation of the model's performance, aiding in fine-tuning and enhancing the accuracy and generalization capabilities of the model. By leveraging these diverse datasets, the research ensures that the proposed model is effective across multiple domains and robust enough to handle different types of textual data, thus enhancing its overall applicability and performance in various NLP tasks.

## V. RESULTS DISCUSSION

In our proposed methodology, we have enhanced the Distil-BERT model by improving performance parameters of the model such as F1 score, accuracy, and learning rate, while reducing evaluation loss and training loss. These improvements are achieved by integrating a memory-based CL technique and designing a task-independent architecture with two additional linear layers placed at the head of the Distil-BERT model. We divide the output representation of Distil-BERT into two spaces: a task-generic space and a task-specific space. The comprehensive results of the proposed model are detailed in Table 3.

**TABLE 3. Results of proposed pipeline.**

| Tasks | F1-Score | Accuracy | Evaluation Loss | Initial Learning Rate |
|---|---|---|---|---|
| Yahoo | 96.84 | 97.84 | .06 | 0.00003144 |
| Yelp | 96.66 | 97.98 | .06 | 0.00003189 |
| Amazon | 95.82 | 98.02 | .06 | 0.00003144 |
| DP-Pedia | 96.20 | 98.32 | .08 | 0.0001972 |
| AG-News | 94.78 | 96.22 | .06 | 0.0001511 |

We conducted a series of experiments with various hyperparameters and settings, including the number of layers, learning rate, batch size and epochs. Our analysis results, indicate that despite the shared language features among tasks, there is no significant performance degradation in individual tasks. This suggests that our tasks remain largely independent regarding task-specific features. Notably, our novel methodology demonstrates significant advancements over the current SOTA models in the field of CL, especially in TC. These improvements highlight the effectiveness of our approach in enhancing performance across the diverse text-based tasks.

It's important to note that the x-axis values represent the global steps toward the completion of instances for each task in each performance parameter, and this value is consistent across all tasks, being 4381. Meanwhile, the y-axis values vary for each task and performance parameter.

### A. MODEL F1 SCORE

The F1 score is a metric commonly used in classification tasks to evaluate model performance. It combines precision and recall into a single measure to provide a balanced assessment of the model's accuracy. The x-axis coordinates represent the steps required to fully train the model, while the y-axis coordinates show the model's performance in terms of the F1 score, with 1 indicating optimal performance and 0 indicating poor performance. Figure 5 illustrates the F1 scores obtained by the proposed model at different evaluation points or epochs during training across various tasks.

The starting evaluation point for the model is step 313, with initial F1 scores for Yahoo, Yelp, Amazon, DB-Pedia, and AG-News tasks being 0.7358, 0.5626, 0.5168, 0.9874, and 0.9199, respectively. The graph begins after the completion of one epoch. DB-Pedia, being the quickest task, completes its F1 evaluation by step 626 with a score of 0.9620, as it has only 2 epochs. However, during this interval (steps 313 to 626), the F1 score for the Yelp task decreases from

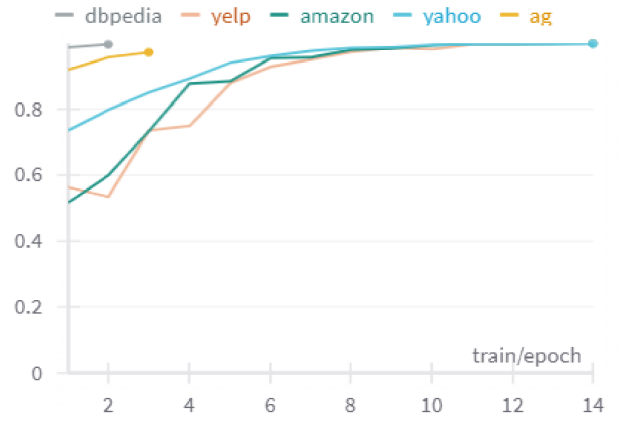


**FIGURE 5.** Model F1 score.

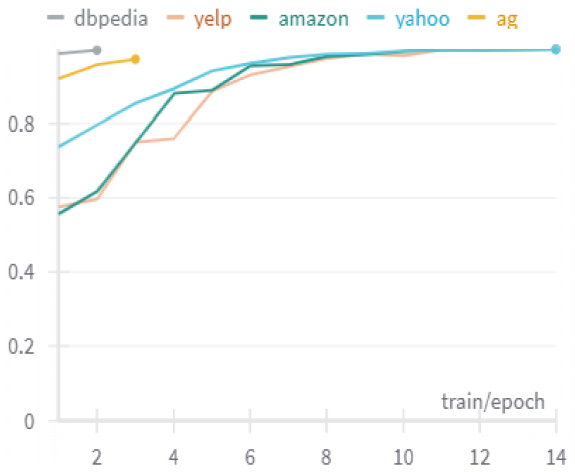


**FIGURE 6.** Model accuracy.

0.5626 to 0.5346. This degradation occurs because the model encounters complex instances that initially a challenge for its performance. Over time, as the model learns from these instances, its performance improves.

The AG-News task completes its F1 evaluation by step 939 with a score of 0.9478, as it has 3 epochs. By step 1252, the F1 scores for Yahoo, Amazon, and Yelp tasks have improved to 0.8928, 0.8776, and 0.749, respectively. From steps 2504 to 4382, the F1 scores for Yahoo, Amazon, and Yelp tasks stabilize at approximately 0.9684, 0.9582, and 0.9666, respectively. This stabilization indicates that the model has fully learned from the instances and epochs, resulting in consistent performance.

The F1 score helps visualize how the model's performance evolves over time, providing insights into its ability to balance precision and recall across different classes, instances, or categories.

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

In equation 3, F1 score shows a measure of the test's accuracy.

### B. MODEL ACCURACY

In this subsection, we analyze the performance of the proposed model by examining its prediction accuracy. The accuracy graph in 6 shows the model's accuracy percentage on the evaluation set at different instances during various training epochs. The figure tracks the model improvements in accuracy over time, indicating the convergence or stability of its performance. The x-axis represents the training progress, showing the number of steps necessary to fully train the model, while the y-axis represents the model's accuracy, ranging from 0 to 1, with 1 indicating optimal performance in terms of accuracy and 0 indicating poor accuracy.

The tasks DB-Pedia and AG-News concluded their evaluations earlier due to having fewer instances and epochs. These tasks begin providing results after the first epoch. The model's accuracy for DB-Pedia ranges from 0.9374 to 0.9832 between steps 313 and 626. Similarly, for AG-News, the accuracy ranges from 0.9198 to 0.9622 between steps 313 and 939.

For the remaining tasks, the model's accuracy gradually increases over time due to the implemented novel strategies. We observed that the accuracy for Yelp and Amazon tasks shows a steady response. For the Yelp task, from step 939 to 1252, the accuracy ranges from 0.7502 to 0.7592, indicating a consistent performance. Similarly, for the Amazon task, from step 1252 to 1565, the accuracy ranges from 0.88 to 0.8882, showing a stable performance.

Finally, the model's accuracy reaches its final magnitudes from step 2500 to 4382, achieving 0.9784 for Yahoo, 0.9798 for Yelp, and 0.9802 for Amazon. These values indicate that the model has fully learned from the instances, resulting in consistent and high performance. The accuracy graph helps in visualizing the model's performance changes over time, providing insights into its ability to correctly classify different instances or categories.

$$\text{Accuracy} = \frac{\sum \text{CorrectPredictions}}{\sum \text{AllPredictions}} \tag{4}$$

Equation 4, represents a used metric in classification tasks, measuring the ratio of correctly predicted instances to the total number of instances. It serves as a crucial measure of model performance, indicating the effectiveness of classification algorithms in accurately assigning labels to input data.

### C. MODEL EVALUATION LOSS

Model evaluation loss, also known as validation loss, represents the error or discrepancy between the model's predictions and the ground truth labels on the evaluation set. It measures how well the model performs on unseen data. In the evaluation loss graph as shown in 7, the horizontal axis indicates the progression of model training, measured in steps, while the vertical axis represents the model's evaluation

loss, with values ranging from 0 to 1. A value of 1 indicates maximum loss, while 0 signifies no loss.

We monitored the evaluation loss of our model for each task individually and then conducted a combined analysis for all tasks. Our findings indicate that the evaluation loss for the model gradually decreases over time, ultimately approaching zero.

The analysis starts at step 313, following the first epoch. At this initial point, the evaluation loss values for Yahoo, Yelp, Amazon, DB-Pedia, and AG-News are 0.9121, 0.9845, 1.004, 0.09761, and 0.3077, respectively. DB-Pedia and AG-News concluded their evaluations earlier, at steps 626 and 939, respectively, due to fewer epochs. During their analysis periods, the evaluation loss for DB-Pedia decreased from 0.09761 to 0.08177, and for AG-News, it decreased from 0.3077 to 0.0663.

For the Yelp task, between steps 939 and 1252, the evaluation loss remained steady, indicating the model had stabilized, and no significant changes were observed. For the remaining tasks, i.e., Yahoo, Yelp, and Amazon, the evaluation loss gradually decreased over time. However, between steps 1878 and 2191, the evaluation loss for the Amazon task temporarily increased from 0.1559 to 0.1736, likely due to the introduction of new instances.

Finally, from step 3443 to 4382, the evaluation loss for all remaining tasks i.e., Yahoo, Yelp, and Amazon—decreased, approaching to zero i.e., 0.0600. This trend indicates that our proposed methodology effectively reduces evaluation loss over time, enhancing the model's performance on unseen data.

$$\text{Evaluation Loss} = \frac{1}{N}\sum_{i=1}^{N}(\text{GroundTruth}_i - \text{Predicted}_i)^2 \tag{5}$$

Equation 5, represents the evaluation loss, it is calculated as the mean squared difference between the predicted and actual values across all samples in the dataset. In Equation 5, $\frac{1}{N}$ represents the number of instances or samples evaluated. $\text{GroundTruth}_i$ represents the ground truth label for the $i$-th instance, and $\text{Predicted}_i^2$ represents the predicted label for the $i$-th instance. This equation calculates the mean squared error (MSE) between the ground truth and predicted labels for over all instances.

### D. LEARNING RATE OF THE MODEL

The learning rate of any model refers to the step size or the rate at which the model's parameters are updated during training. It determines how quickly or slowly the model learns from the training data, significantly impacting the model's convergence, stability, and overall performance. Adjusting the learning rate is crucial for finding the optimal value that leads to faster convergence and better generalization.

In Figure 8, the horizontal axis represents the progression of model training, measured in steps, while the vertical axis represents the learning rate on a logarithmic scale. The values

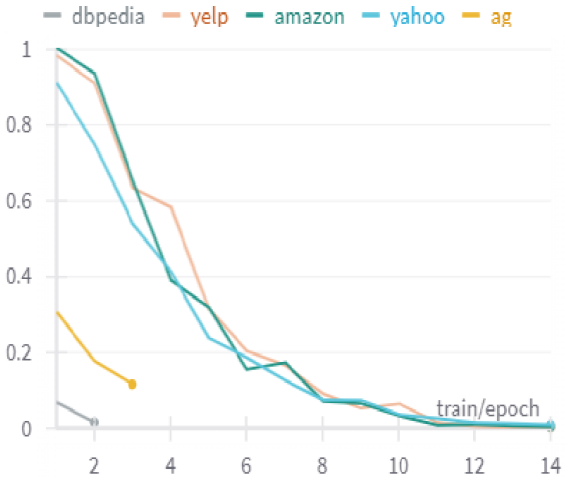


**FIGURE 7.** Model evaluation loss.

on the y-axis start from 0, followed by 5e-5, 0.0001, and so on, up to 0.0002, indicating the decreasing magnitude of the learning rate over time.

Initially, the learning rate is high to allow the model to quickly learn from the provided instances. As training progresses, the learning rate decreases to enable more precise adjustments. The tasks DB-Pedia, AG-News, and Yahoo start from step 1, while Yelp starts from step 3, and Amazon from step 4.

At the beginning, the learning rate gradually decays, reaching a peak value of 0.0002 for all tasks within the first 1000 steps. After reaching this peak, the learning rate for each task exponentially decreases, eventually hitting zero at different points. DB-Pedia and Amazon reach a learning rate of zero (0.0001972, and 0.00003144) at steps 626 and 644, respectively, indicating they have fully learned from the provided instances. AG-News concludes its learning at step 939 by dropping its learning loss to 0.0001511, while Yelp and Yahoo end their learning at step 4382, with the learning rate dropping to zero (0.00003189, and 0.00003144).

This analysis, shows that the model's learning rate decreases over time, ensuring a balance between rapid learning and precise adjustments for optimal performance.

$$\text{Learning Rate}(t) = \frac{a}{(1+bt)} \tag{6}$$

The above equation 6, represents a learning rate schedule where the learning rate decreases over the time according to a specified schedule. In Equation 6, $t$ represents the global step or iteration during training, while $a$ and $b$ are parameters that control the decay rate and initial learning rate, respectively.

### E. TRAINING LOSSES OF THE MODEL

Training loss is the reciprocal of learning rate that represents the error or discrepancy between the model's predictions and the actual labels on the training set. It measures how well the model is fitting the training data. The goal during training is to minimize the training loss, indicating that the

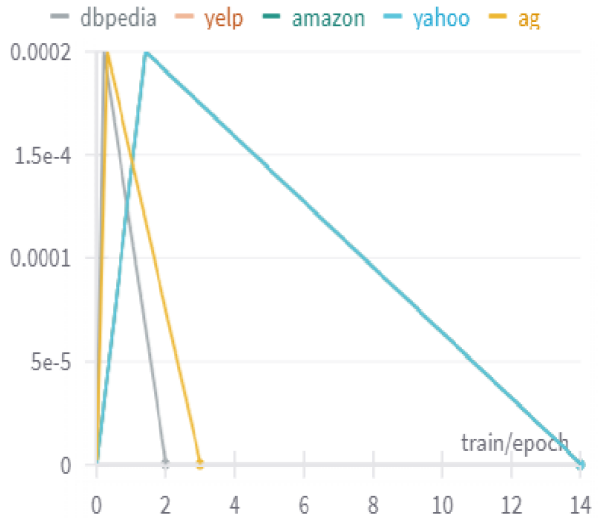


**FIGURE 8.** Model learning rate.

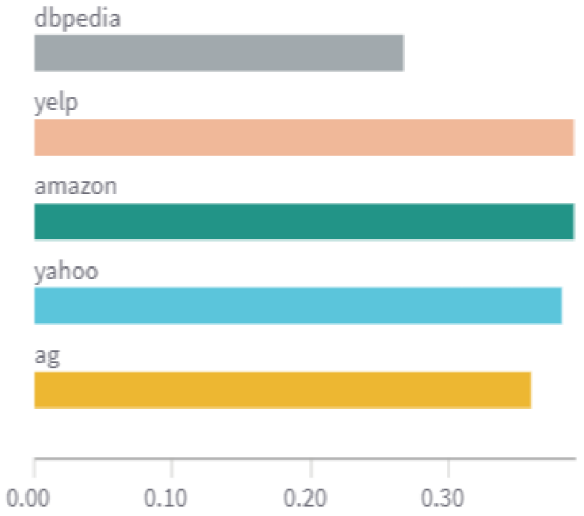


**FIGURE 9.** Training losses of the model.

model is learning to make accurate predictions on the training data. Our model successfully minimized the training losses throughout the training process.

Figure 9 illustrates the training losses of our proposed model. This figure shows that different magnitudes of the losses over time. The tasks DB-Pedia and AG-News calculated the model losses quickly and ended their contributions early due to having only 2 and 3 epochs, respectively.

Initially, the training losses were higher because the model was still in learning phase and capturing the provided instances. However, once the model analyzed and learned from these instances, it retained this knowledge for future predictions.

## VI. RESULTS COMPARISON WITH SOTA

In this section, we present a comprehensive comparison of our proposed model's performance with existing SOTA models [29], [30], focusing specifically on the accuracy metric to evaluate the effectiveness of our approach in TC. For more clarity, we compared the performance metrics for each task across three different models, as shown in Tables 4 to 8. In contrast to our proposed approach, the existing models only report accuracy in their experimental discussions, neglecting other critical performance metrics. Notably, these models do not consider or discuss the F1 score, which is a vital parameter. Additionally, they fail to address key factors such as model evaluation loss, learning rate, and training losses, all of which are essential for a comprehensive assessment of any model's performance. To facilitate a visual representation of this comparison, we have included a bar graph, as depicted in Figure 10. The x-axis lists the compared models along with the datasets used, while the y-axis shows the corresponding accuracy of each model for these datasets. This graph provides a clear visualization of the distinctions between our proposed model and the compared SOTA models.

To ensure a fair and unbiased comparison, we carefully selected two relevant articles, from the SOTA. These articles were chosen because they conducted experiments on the same datasets used in our evaluation, but with different models aiming for the same objectives.

In [29], the authors introduced a method to enhance TC performance by utilizing LSTM networks combined with word embedding techniques. Leveraging LSTM's ability to capture long-range dependencies and representing words as dense vectors through word embeddings, the authors aimed to improve the accuracy of TC tasks. Their experiments on Yahoo, Yelp, Amazon, DB-Pedia, and AG-News datasets reported accuracy scores of 92.5, 94.4, 94.3, 91.67, and 91.43, respectively.

Similarly, in [30], the authors addressed the task of binarizing word embeddings with minimal information loss using a technique called near-lossless binarization (NLB), which combines quantization and clustering methods. NLB aims to preserve semantic information by maintaining similarities between word embeddings in the binary space. The paper evaluated NLB on various downstream tasks such as SA and named entity recognition (NER), demonstrating its effectiveness in compressing word representations while preserving their quality. However, the specific model used for

**TABLE 4.** Performance metrics comparison for yahoo task.

| Model | Task 1-Yahoo | | | |
|---|---|---|---|---|
| | F1 Score | Accuracy | Evaluation Losses | Learning Rates |
| Proposed Model | 96.84 | 97.84 | .06 | 0.00003144 |
| LSTM Model | — | 92.5 | — | — |
| Character Level CNNN | — | 67.1 | — | — |


**FIGURE 10.** Results comparison.

**TABLE 5.** Performance metrics comparison for yelp task.

| Model | Task 2-Yelp | | | |
|---|---|---|---|---|
| | F1 Score | Accuracy | Evaluation Losses | Learning Rates |
| Proposed Model | 96.66 | 97.98 | .06 | 0.00003189 |
| LSTM Model | — | 94.4 | — | — |
| Character Level CNNN | — | 88 | — | — |

**TABLE 6.** Performance metrics comparison for amazon task.

| Model | Task 3- Amazon | | | |
|---|---|---|---|---|
| | F1 Score | Accuracy | Evaluation Losses | Learning Rates |
| Proposed Model | 95.82 | 98.02 | .06 | 0.00003144 |
| LSTM Model | — | 94.3 | — | — |
| Character Level CNNN | — | 84.5 | — | — |

**TABLE 7.** Performance metrics comparison for DB-Pedia task.

| Model | Task 4- DB-Pedia | | | |
|---|---|---|---|---|
| | F1 Score | Accuracy | Evaluation Losses | Learning Rates |
| Proposed Model | 96.20 | 98.32 | .08 | 0.0001972 |
| LSTM Model | — | 91.67 | — | — |
| Character Level CNNN | — | 97.4 | — | — |

generating the word embeddings was a Character-level CNN. Their experiments on Yahoo, Yelp, Amazon, DB-Pedia, and AG-News datasets reported accuracy scores of 67.1, 88, 84.5, 97.4, and 88.1, respectively.

In our proposed approach, we conducted experiments on the same datasets i.e., Yahoo, Yelp, Amazon, DB-Pedia, and AG-News using Distil-BERT model. We observed several performance parameters, but here we focus on comparing the accuracy scores. Our model achieved accuracy scores of 97.84, 97.98, 98.02, 98.32, and 96.22, respectively.

By comparing our proposed model with these SOTA results, we aim to provide a comprehensive understanding of the model's performance in terms of accuracy. This analysis offers valuable insights into the advancements and improvements achieved by our proposed model, reinforcing its potential contribution to the field of TC.

## VII. CONCLUSION AND FUTURE DIRECTIONS

Our research highlights the effectiveness of integrating pre-trained Distil-BERT with a memory-based CL technique

**TABLE 8. Performance metrics comparison for AG-News task.**

| Model | Task 5-AG-News | | | |
|---|---|---|---|---|
| | F1 Score | Accuracy | Evaluation Losses | Learning Rates |
| Proposed Model | 94.78 | 96.22 | .06 | 0.0001511 |
| LSTM Model | — | 91.43 | — | — |
| Character Level CNNN | — | 88.1 | — | — |

for TC tasks. By incorporating task-specific linear layers and employing a dynamic setup, our approach allows for the seamless addition of new tasks, ensuring scalability and adaptability. We systematically save the best model after each training epoch to address the challenge of CF, preserving the performance of previously learned tasks. Our experiments, which involved various hyperparameters and settings, confirmed the reliability of our methodology. Notably, our approach maintains task independence, with no significant decline in performance across individual tasks despite shared language features. This indicates that the tasks remain largely unaffected by each other's specific features. These results underscore the potential of our approach to effectively support CL across a wide range of text-based tasks.

As a future direction, we aim to explore the transfer ability of our approach to other NLP tasks beyond TC, such as SA and machine translation. Additionally, we plan to investigate the impact of alternative pre-training techniques and model architectures on the performance of our approach. We also intend to optimize the computational efficiency of our methodology, particularly for large-scale datasets. Through these avenues, we anticipate further improving the performance and applicability of our approach within the field of NLP.

## REFERENCES


[1] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Netw. Anal. Mining*, vol. 11, no. 1, p. 81, Dec. 2021.

[2] M. Bordoloi and S. K. Biswas, "Sentiment analysis: A survey on design framework, applications and future scopes," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 12505–12560, Nov. 2023.

[3] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022.

[4] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023.

[5] F. Es-Sabery, K. Es-Sabery, J. Qadir, B. Sainz-De-Abajo, A. Hair, B. García-Zapirain, and I. De La Torre-Díez, "A MapReduce opinion mining for COVID-19-related tweets classification using enhanced ID3 decision tree classifier," *IEEE Access*, vol. 9, pp. 58706–58739, 2021.

[6] K. P. Gunasekaran, "Exploring sentiment analysis techniques in natural language processing: A comprehensive review," 2023, *arXiv:2305.14842*.

[7] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *Int. J. Res. Marketing*, vol. 36, no. 1, pp. 20–38, Mar. 2019.

[8] F. Es-sabery, K. Es-sabery, H. Garmani, J. Qadir, and A. Hair, "Evaluation of different extractors of features at the level of sentiment analysis," *Infocommunications J.*, vol. 14, no. 2, pp. 85–96, 2022.

[9] J. M. Duarte and L. Berton, "A review of semi-supervised learning for text classification," *Artif. Intell. Rev.*, vol. 56, no. 9, pp. 9401–9469, Sep. 2023.

[10] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," *Decis. Anal. J.*, vol. 3, Jun. 2022, Art. no. 100073.

[11] J. Hurtado, D. Salvati, R. Semola, M. Bosio, and V. Lomonaco, "Continual learning for predictive maintenance: Overview and challenges," *Intell. Syst. Appl.*, vol. 19, Sep. 2023, Art. no. 200251.

[12] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Netw.*, vol. 113, pp. 54–71, May 2019.

[13] 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, May 2023.

[14] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, N. Alzakari, A. Abou Elwafa, and H. Kurdi, "Impact of dataset size on classification performance: An empirical evaluation in the medical domain," *Appl. Sci.*, vol. 11, no. 2, p. 796, Jan. 2021.

[15] X. Yao, T. Huang, C. Wu, R.-X. Zhang, and L. Sun, "Adversarial feature alignment: Avoid catastrophic forgetting in incremental task lifelong learning," *Neural Comput.*, vol. 31, no. 11, pp. 2266–2291, Nov. 2019.

[16] Z. Ke, B. Liu, H. Xu, and L. Shu, "CLASSIC: Continual and contrastive learning of aspect sentiment classification tasks," 2021, *arXiv:2112.02714*.

[17] R. Silva Barbon and A. T. Akabane, "Towards transfer learning techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for automatic text classification from different languages: A case study," *Sensors*, vol. 22, no. 21, p. 8184, Oct. 2022.

[18] Z. Ke, H. Xu, and B. Liu, "Adapting BERT for continual learning of a sequence of aspect sentiment classification tasks," 2021, *arXiv:2112.03271*.

[19] B. Liang, W. Luo, X. Li, L. Gui, M. Yang, X. Yu, and R. Xu, "Enhancing aspect-based sentiment analysis with supervised contrastive learning," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 3242–3247.

[20] B. Rodrawangpai and W. Daungjaiboon, "Improving text classification with transformers and layer normalization," *Mach. Learn. Appl.*, vol. 10, Dec. 2022, Art. no. 100403.

[21] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi, "A fine-tuned BERT-based transfer learning approach for text classification," *J. Healthcare Eng.*, vol. 2022, pp. 1–17, Jan. 2022.

[22] Q. Jia, J. Cui, Y. Xiao, C. Liu, P. Rashid, and E. F. Gehringer, "ALL-IN-ONE: Multi-task learning BERT models for evaluating peer assessments," 2021, *arXiv:2110.03895*.

[23] J. Tong, Z. Wang, and X. Rui, "A multimodel-based deep learning framework for short text multiclass classification with the imbalanced and extremely small data set," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, Oct. 2022.

[24] S. Wang and J. Liu, "Continual learning for sentiment classification by iterative networks combination," in *Proc. 5th Int. Conf. Comput. Sci. Artif. Intell.*, Dec. 2021, pp. 150–155.

[25] Y. Huang, Y. Zhang, J. Chen, X. Wang, and D. Yang, "Continual learning for text classification with information disentanglement based regularization," 2021, *arXiv:2104.05489*.

[26] C. Rochereau, B. Sagot, and E. Dupoux, "Neural language modeling of free word order argument structure," 2019, *arXiv:1912.00239*.

[27] D. Trotta, R. Guarasci, E. Leonardelli, and S. Tonelli, "Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus," 2021, *arXiv:2109.12053*.

[28] N. Bel, M. Punsola, and V. Ruíz-Fernández, "EsCoLA: Spanish corpus of linguistic acceptability," in *Proc. Joint Int. Conf. Comput. Linguistics, Language Resour. Eval.*, May 2024, pp. 6268–6277.

[29] A. C. Adamuthe, "Improved text classification using long short-term memory and word embedding technique," *Int. J. Hybrid Inf. Technol.*, vol. 13, no. 1, pp. 19–32, Mar. 2020.

[30] J. Tissier, C. Gravier, and A. Habrard, "Near-lossless binarization of word embeddings," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 7104–7111.

**SAHAR SHAH** received the M.Sc. degree in electronics from the Department of Electronics, University of Peshawar, Pakistan, in 2016, and the M.Phil. degree in electronics from the Department of Electronics, Quaid-i-Azam University, Islamabad, Pakistan, in 2019. He is currently pursuing the Ph.D. degree in computer science with the Department of Informatics, Systems and Communication (DISCo), University of Milano-Bicocca, Italy. He is a Visiting Ph.D. Student with Social Things Srl Company, Milan, and a Developer in innovative ICT solutions and algorithms in the fields of artificial intelligence (AI), the Internet of Things (IoT), and human–computer interaction. He is also a Research Collaborator with the School of Computing, National University of Computer and Emerging Sciences (NUCES), Lahore, Pakistan; and the Department of Computer Science, University of Hassan II, Casablanca, Morocco. He was a Lecturer in electronics with the Higher Education Department of Pakistan, and taught several subjects at master's degree course, such as digital logic design, computer networking, and data communication. He has supervised and designed many projects, i.e., digital voting machine, smart home appliances, robotic arm to improve the solar panels efficiency, and many more. He has published several international journal research articles in prestigious journals, such as *Symmetry* (MDPI), Micromachines, *Big Data* (Hindawi), *Scientific Programming*, *Industrial Internet of Things*, and *Wireless Communication and Mobile Computing*; and many conferences. His current research interests include the artificial intelligence, text classification, sentiment analysis, transformer models, fuzzy systems, continual learning techniques, underwater wireless sensor networks, and cloud computing. He is a Reviewer of several prestigious international publishers, such as Springer, *Sensors* (MDPI), and *International Journal of Distributed Sensor Networks* (IJDSN).

**SARA LUCIA MANZONI** is an Associate Professor with the Department of Informatics, Communication and Systms (DISCo), University of Milano-Bicocca. She was the Head of the AI Laboratory, DISCo, from 2007 to 2017; and a Co-Founder of Crowdyxity srl, University of Milan-Bicocca spin-off. Her main research results are in technology transfer projects of several areas of artificial intelligence, primarily, knowledge-based reasoning in decision support systems and multi-agent systems approach to study complex systems dynamics.

**FAROOQ ZAMAN** received the M.Phil. degree in computer science from the Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan, in 2017. He is a Ph.D. Scholar with the AI Laboratory, Information Technology University, Lahore. Prior to joining Scientometrics Laboratory, he was a Visiting Faculty with Quaid-i-Azam University. His research interests include text summarization, text simplification, and machine translation.

**FATIMA ES SABERY** received the Technical University degree from the Department of Computer Science, Higher School of Technology, Casablanca, Morocco, in 2013, and the master's degree in business intelligence from the Department of Computer Sciences, Sultan Moulay Sliman University, Beni Mellal, Morocco, in 2016. She received the professional license with option IT development from the Department of Computer Sciences, Faculty of Science, Casablanca, in 2014. She has published several research papers in many international conferences and journals, i.e., *Fuzzy Information and Engineering*; *International Journal of Informatics and Communication Technology*, Third International Conference on Networking, Information Systems and Security; and 2019 International Conference on Intelligent Systems and Advanced Computing Sciences. Her general research interests include data mining area, big data field, wireless sensor networks, fuzzy systems, machine learning, deep learning, and the Internet of Things.

**FRANCESCO EPIFANIA** received the bachelor's degree in digital communication, and the double master's and Ph.D. degrees in computer science from the University of Milan. He was a Research Fellow and a Professor with the Department of Computer Science, University of Milan. He is also dedicated to the evaluation of the recommender systems. He is a CEO and a Founder of Social Things srl and Whoteach. He was a Researcher in computer science with the University of Milan. He is involved in the study of the evaluation of recommender systems based on user data. He has carried out consultancy activities in the ICT field, both in academic and academic fields. He has carried out teaching activities for the degree courses in computer science with the University of Milan, such as foundations of digital communication, systems for computer-aided design, multimedia publishing and informatics, and laboratory courses. He has published more than 30 papers in national and international conferences and has supervised more than 200 student theses in the computer science field. He has produced numerous publications for national and international conferences and magazines. His research interests include human–machine interaction, and in particular the creation of intelligent and multichannel interactive systems for the enrichment of knowledge. His research interests also include artificial intelligence, in particular the evaluation, design, and development of multimedia interactive intelligent systems for knowledge enrichment.

**ITALO FRANCESCO ZOPPIS** received the Ph.D. degree in computer science from Università degli Studi di Milano. He is currently an Associate Professor of computer science with the University of Milano-Bicocca. His research focuses on translational knowledge discovery from biological and clinical datasets, mainly applying machine learning techniques to enhance medical decisions for patient well-being.

Open Access funding provided by 'Università degli Studi di Milano Bicocca' within the CRUI CARE Agreement