Adaptive Domain-Specific Document-Level Sentiment Analysis with Meta-Learning and Hybrid Language Models

Yicheng Sun[a], Jacky Wai Keung[a], Zhen Yang[b,*], Hi Kuen Yu[a], Wenqiang Luo[a] and Yihan Liao[a]

[a]Department of Computer Science, City University of Hong Kong, Hong Kong, China
[b]School of Computer Science and Technology, Shandong University, Qingdao, China

ARTICLE INFO

Keywords:
Sentiment Analysis
Model-Agnostic Meta-Learning
Large Language Model
Hybrid Language Model

ABSTRACT

Context: Sentiment analysis, a crucial technique in Natural Language Processing (NLP), extracts emotional insights from text data, influencing decision-making across multiple domains. Document-level sentiment analysis (DLSA) aims to assess the overall sentiment of entire texts, but current models face challenges such as limited accuracy in specific domains and the need for large datasets.

Objective: To address these limitations, we propose an adaptive approach to DLSA that leverages model-agnostic meta-learning (MAML) techniques in conjunction with a hybrid language model, termed DLSA-MAML. This approach enhances domain adaptability and improves performance in sentiment classification tasks, even with limited data. DLSA-MAML is particularly effective in distinguishing between positive and negative sentiments across diverse contexts, resulting in improved sentiment analysis performance.

Method: DLSA-MAML integrates MAML with a hybrid language model, enabling the framework to quickly adapt to new tasks by optimizing model parameters through few-shot learning. This approach reduces reliance on large-scale datasets while effectively capturing both semantic and contextual information, thereby enhancing the model's robustness and accuracy in sentiment classification.

Results: Experiments on three datasets demonstrate that DLSA-MAML outperforms eight benchmark models, achieving higher accuracies of 6.73%, 6.89%, and 9.26% respective to their comparative models. We also validate the principles underlying the hybrid language model design in DLSA-MAML and explore the enhancement effect of MAML on the model, further highlighting its potential for advancing sentiment classification.

Conclusion: DLSA-MAML offers a robust and adaptable solution for document-level sentiment analysis, advancing the field by improving cross-domain generalization and reducing the dependency on large training datasets.

1. Introduction

Natural Language Processing (NLP) [6] enables machines to understand and analyze human language, facilitating a wide range of text-based applications [34], including sentiment analysis. Sentiment analysis, a computational technique aimed at identifying emotional tendencies and viewpoints within textual data, has become a vital tool in NLP [22]. It enables the interpretation of emotional tones and subjective information embedded in the text [36]. With the vast amount of online textual data continuously growing, sentiment analysis has become indispensable for extracting insights from opinions and emotions, with significant implications for businesses and governments [3].

The core of sentiment analysis involves identifying and classifying opinions expressed in a piece of text and determining whether the sentiment conveyed is positive, negative, or neutral [32]. Early methods primarily relied on lexicon-based approaches, where predefined lists of words associated with specific sentiments were used to analyze the text [22]. While these methods provided some initial success, they often struggled with the contextual nuances and complexity of natural language. As sentiment analysis has developed, it has evolved from sentence-level sentiment analysis (SLSA) [24] to paragraph-level sentiment analysis (PLSA) [11], and subsequently to the more challenging task of document-level sentiment analysis (DLSA) [38, 3]. Document-level sentiment analysis has now emerged as a pivotal

*Corresponding author
yicsun2-c@my.cityu.edu.hk (Y. Sun); Jacky.Keung@cityu.edu.hk (J.W. Keung); zhenyang@sdu.edu.cn (Z. Yang); hikuenyu2-c@my.cityu.edu.hk (H.K. Yu); wenqialuo4-c@my.cityu.edu.hk (W. Luo); yihanliao4-c@my.cityu.edu.hk (Y. Liao)
ORCID(s): 0009-0007-6555-0571 (Y. Sun); 0000-0002-3803-9600 (J.W. Keung); 0000-0003-0670-4538 (Z. Yang); 0009-0009-8451-188X (H.K. Yu); 0009-0005-4171-2025 (W. Luo); 0000-0002-8002-9190 (Y. Liao)

task in NLP, with applications ranging from social media monitoring to customer feedback analysis. Unlike sentence-level or aspect-based sentiment analysis, which focuses on localized sentiment, document-level sentiment analysis considers the overall sentiment conveyed by the entire text [30]. This holistic approach presents unique challenges, such as the difficulty of capturing long-range dependencies across sentences or paragraphs [38].

As the field of NLP has evolved, so have the techniques employed in document-level sentiment analysis [55]. The advent of machine learning, and more recently, deep learning, has enabled the development of sophisticated models capable of capturing the subtleties of human language [53]. Traditional rule-based and keyword-based methods often fall short when dealing with complex data, making the use of advanced deep learning techniques increasingly critical for improving the accuracy and robustness of sentiment analysis [21]. However, despite these advancements, document-level sentiment analysis remains challenging. Key challenges include: (1) The significant variation in sentiment-rich texts across different domains, which limits the accuracy of single machine learning and deep learning models, leaving room for improvement in domain-specific sentiment analysis tasks. (2) The large amounts of training data required by existing models, making it difficult for them to quickly adapt to new tasks. Additionally, Severe imbalances between positive and negative samples in some datasets, an issue that few studies have thoroughly addressed. To address the first challenge, recent advancements in hybrid language models, which combine the strengths of pre-trained transformers and rule-based systems, have shown promise in improving classification accuracy. These hybrid models utilize the contextual embeddings generated by large pre-trained models such as BERT [8], GPT [41], or T5 [37], while integrating domain-specific rules or supplementary models to fine-tune sentiment predictions.

In light of the second challenge, domain transfer [26] and context adaptability [48] have become essential strategies for addressing the limitations of sentiment analysis. Domain transfer refers to a model's ability to shift from one application domain (such as product review analysis) to a completely different domain (for example, opinion analysis in political forums) [17]. This transfer not only requires the model to maintain its original effectiveness in the new domain but also to quickly adapt to the specific language usage, expressions, and emotional characteristics of the new domain. Context adaptability, on the other hand, focuses on the model's capacity to rapidly adjust to new environments or data [16]. In practical applications, this means that the model can swiftly modify its learning strategies when exposed to a small amount of new domain data, adapting to the new data distribution and characteristics. Given the dynamic and diverse textual data available online, this adaptability is key to effective sentiment analysis [29]. Meta-learning [44] offers a promising solution by enabling models to quickly adapt to new tasks with minimal data. In the context of sentiment analysis, meta-learning can be employed to create models that generalize better across diverse document types and domains. By learning a set of shared parameters from multiple tasks, a meta-learning framework can rapidly adjust to new document-level sentiment analysis tasks, thereby enhancing the model's robustness and adaptability.

In this paper, we aim to explore the application of meta-learning in sentiment analysis, with a particular focus on domain transfer and context adaptability. Our goal is to provide more flexible and efficient analytical tools for the field of NLP, enabling better adaptation to the ever-changing data environment and application requirements. Specifically, we propose a novel adaptive approach to **d**ocument-**l**evel **s**entiment **a**nalysis (DLSA) that leverages **m**odel-**a**gnostic **m**eta-**l**earning (MAML) techniques in conjunction with hybrid language models, referred to as **DLSA-MAML**. Our model aims to leverage the generalization capabilities and flexibility of MAML to enhance the cross-domain transferability and context adaptability of sentiment analysis models. Given the significant variability of comments across different scenarios and the limited availability of similar samples, introducing MAML can significantly improve the model's generalization ability in low-sample scenarios, allowing it to quickly adapt and converge when faced with new sentiment samples. By optimizing the model parameters through meta-learning, MAML equips the sentiment analysis model to rapidly adjust and perform effectively on new classification tasks with minimal data. This approach has proven particularly effective in handling diverse datasets with varying contexts and samples, reducing the reliance on large-scale labeled data and thereby improving the practicality and robustness of DLSA-MAML.

Additionally, the hybrid language model is built on the architecture of RoBERTa and an attention-based Bi-LSTM. RoBERTa, a robust large-scale pre-trained language model, uses the encoder section of the Transformer model to effectively capture semantic features during pre-training, demonstrating remarkable adaptability across different datasets. It encodes the document-level text, generates context-aware representations, and inputs them into an attention-based Bi-LSTM. The attention mechanism layer highlights crucial information in the text by assigning varying weights to different parts of the context, while the Bi-LSTM adjusts its focus based on different segments of the text, leveraging both forward and backward information to capture sequential features. By integrating RoBERTa and attention-based Bi-LSTM into a hybrid language model, DLSA-MAML thoroughly understands document-level text and captures semantic and contextual information at multiple levels, thereby enhancing sentiment analysis capabilities

for better classification of positive and negative sentiments. We evaluate our approach on three benchmark datasets, demonstrating its ability to generalize across different domains and document structures. The results underscore the potential of our adaptive framework to advance sentiment analysis, providing a robust tool for analyzing sentiment in complex, real-world texts.

In summary, The contributions of this paper can be summarized in four-fold:

- We propose the groundbreaking DLSA-MAML model, which employs model-agnostic meta-learning strategies to enhance cross-domain transferability and context adaptability. This approach improves the model's generalization capability in low-sample scenarios, allowing it to quickly adapt and converge when faced with new sentiment classification tasks.
- To our knowledge, we are the first to utilize model-agnostic meta-learning for document-level sentiment analysis across different scenarios and to systematically assess its effectiveness.
- DLSA-MAML integrates RoBERTa and an attention-based Bi-LSTM as a hybrid language model. This architecture provides robustness and enhances the accuracy of classifying positive and negative sentiments.
- Extensive experiments are conducted to evaluate the effectiveness of DLSA-MAML in document-level sentiment analysis across various scenarios, demonstrating the superior potential of our adaptive model in advancing the field of sentiment analysis.

The remainder of this paper is organized as follows. Section 2 presents the background of sentiment analysis. Section 3 elaborates on constructing the training set for MAML (Model-Agnostic Meta-Learning), developing the hybrid language model, fine-tuning, and conducting sentiment analysis using the fine-tuned hybrid language model. Sections 4 and 5 discuss the experimental design and results. Section 6 addresses the threats to validity. Section 7 reviews related work. Finally, Section 8 summarizes the paper.

2. Background

2.1. Sentiment Analysis Task

Sentiment analysis [43], a core component of the field of NLP, involves the in-depth analysis, processing, summarization, and reasoning of subjective texts imbued with emotional tones to ultimately determine whether the sentiment expressed in the text is positive or negative [49]. Online platforms such as Amazon, Twitter, and movie review sites have increasingly become primary venues for individuals to express their emotions and attitudes toward people, events, products, and more. Analyzing these sentiments is crucial for understanding public opinion and social trends.

The core of sentiment analysis involves identifying and classifying opinions expressed in a piece of text and determining whether the sentiment conveyed is positive, negative, or neutral [32]. Early methods primarily relied on lexicon-based approaches, where predefined lists of words associated with specific sentiments were used to analyze the text. While these methods provided some initial success, they often struggled with the contextual nuances and complexity of natural language. As sentiment analysis has developed, it has evolved from sentence-level sentiment analysis (SLSA) [24] to paragraph-level sentiment analysis (PLSA) [11], and subsequently to the more challenging task of document-level sentiment analysis (DLSA) [38, 3]. Document-level sentiment analysis has now emerged as a pivotal task in NLP, with applications ranging from social media monitoring to customer feedback analysis. Unlike sentence-level or aspect-based sentiment analysis, which focuses on localized sentiment, document-level sentiment analysis considers the overall sentiment conveyed by the entire text [30]. This holistic approach presents unique challenges, such as the difficulty of capturing long-range dependencies across sentences or paragraphs [38]. Traditional machine learning methods, though effective in specific contexts, often struggle with these challenges due to their reliance on manually engineered features and limited contextual awareness.

2.2. Text Preprocessing Techniques

Text preprocessing is a crucial step in sentiment analysis, as it directly influences the quality and performance of the subsequent models. The primary goal of preprocessing is to convert raw text into a structured format that is more suitable for computational analysis [45]. This process typically involves several stages, each aimed at cleaning and standardizing the textual data while retaining the essential information required for accurate sentiment classification.

The first stage of text preprocessing generally involves tokenization, which breaks down the text into individual units such as words or phrases. This step is essential for simplifying the analysis and allowing the model to process the text as discrete elements [35]. Traditionally, preprocessing also includes cleaning the text by removing noise like punctuation, stopwords (e.g., "and," "the," "is"), and special characters that do not contribute significant meaning to the sentiment but can clutter the analysis. As sentiment analysis techniques have evolved, more advanced methods have been developed to enhance the model's understanding of the text [1]. One of the most effective approaches in modern sentiment analysis involves transforming the text into semantic vectors, a process that allows models to capture the underlying meaning of the text rather than just its surface features. By converting text into semantic vectors, RoBERTa can generate context-aware embeddings that encapsulate the intricate relationships and meanings within the text.

2.3. Meta-Learning

Meta-learning [44], also known as "learning to learn", is a typical Few-Shot Learning method of altering search strategy in hypothesis space, which aims to learn good initialization parameters from a large-scale dataset for fine-tuning, thereby guiding the model to quickly adapt to the target task with few samples. Different from conventional transfer learning technologies, meta-learning trains the model on a variety of heterogeneous sub-tasks, and its model parameters are updated based on the query set (for the validation purpose) of each sub-task to ensure the generality of different tasks. In recent years, several representative meta-learning methods have been continuously proposed, such as Model-Agnostic Meta-Learning (MAML) [12].

By learning a set of shared parameters from multiple tasks, a meta-learning framework can rapidly adjust to new document-level sentiment analysis tasks, thereby enhancing the model's robustness and adaptability. Specifically, meta-learning plays a significant role in achieving domain transfer and context adaptability [12]. By constructing a general and flexible learning mechanism, it enables models to effectively utilize existing knowledge and quickly adjust their strategies when faced with new datasets or tasks, thereby improving the accuracy and adaptability of sentiment analysis [13]. This approach is particularly effective in handling large and diverse online textual data, significantly enhancing the practicality and flexibility of sentiment analysis across different application scenarios.

In this paper, we adopt MAML as the few-shot meta-learning method to transfer general-purpose knowledge from a data-rich training set to a domain-specific dataset. We pre-train a hybrid language model, using MAML to enhance its ability to perform sentiment analysis tasks, focusing on the accurate identification of both positive and negative sentiments. By leveraging MAML, we aim to equip the hybrid language model with a robust initialization that facilitates rapid adaptation to sentiment-specific features present in the target domain. This approach is particularly advantageous in scenarios where the domain-specific dataset is limited, as the meta-learning process enables the model to efficiently utilize the general-purpose knowledge acquired during pre-training while fine-tuning to the nuances of the new domain. As a result, the model not only generalizes well across different sentiment analysis tasks but also exhibits improved performance on the target domain, ensuring accurate sentiment classification even with minimal labeled data. Additionally, we emphasize domain transfer and context adaptability in MAML, as these capabilities are crucial for maintaining model effectiveness across diverse tasks and under varying data conditions, addressing real-world challenges in sentiment analysis.

2.3.1. Domain Transfer

Domain transfer refers to the ability of a model trained on one domain to perform well in another, different domain without requiring extensive re-training. In this study, we explore how the MAML approach enhances the cross-domain transferability of our proposed model. For example, we pre-train the model on multiple datasets, such as Rotten Tomatoes reviews, to extract general sentiment-related patterns. When fine-tuned on a new target domain, such as the IMDb dataset, the model requires only a few-shot sample (20%) to adapt efficiently. Specifically, we assess domain transfer by comparing the performance of models with and without MAML on datasets from distinct domains. This comparison quantifies the impact of MAML on transferring learned knowledge across domains, demonstrating the model's ability to generalize with minimal additional training.

2.3.2. Context Adaptability

Context adaptability refers to the model's ability to perform accurately even when trained on incomplete or less relevant data, adapting to new contexts without requiring domain-specific fine-tuning. For example, we test this capability by deliberately excluding highly relevant datasets (e.g., Rotten Tomatoes reviews) during the meta-learning phase. We then evaluate the model's performance on the IMDb dataset, which shares significant similarities with

the excluded data, to assess whether our proposed model can adapt to missing contextual information. This approach evaluates the model's ability to perform accurate sentiment classification even in the absence of optimal contextual data, demonstrating that it can generalize effectively with only partially relevant information.

2.4. RoBERTa

BERT [8] is a model consisting of Transformer encoders, distinguished by its foundation in unsupervised learning from large-scale textual data. The Key innovations of BERT include introducing three types of embeddings: Token Embedding, Segment Embedding, and Position Embedding. Token Embedding converts each word in the text into a vector representation in a high-dimensional space. Segment Embedding distinguishes different sentences in the text, aiding the model in understanding and analyzing relationships between sentences. Position Embedding provides positional information for each word in the text [20]. BERT's pre-training employs two types of unsupervised learning methods: Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM randomly masks certain vocabulary from the input text and uses "[MASK]" to replace these words, while NSP determines whether two sentences are logically connected, evaluating coherence or interdependence between the sentences.

RoBERTa [27], a robustly optimized BERT [8] pre-training method introduced by Facebook AI in 2019, represents a significant advancement over the BERT model. RoBERTa retains the multi-layer Transformer encoder structure of BERT; however, in this model, key parameters are adjusted, such as an increased number of layers, expanded size of hidden layers, and larger number of attention heads. These features enable RoBERTa to more effectively capture complex linguistic features and patterns.

In terms of pre-training, RoBERTa differs from BERT by relying solely on MLM as its pre-training task. Additionally, RoBERTa introduces Dynamic Masking mode, replicating training data and executing a series of masking strategies.

During data processing, RoBERTa employs large-scale and diversified training datasets covering a wider range of text types and larger corpora, significantly enhancing the model's understanding of different text types and its generalization capabilities. Furthermore, RoBERTa employs longer sequence lengths and larger batch sizes during training, further enhancing training efficiency and model performance.

In the context of sentiment analysis, RoBERTa excels at capturing contextual information in text sequences and comprehending complex text structures, critical for accurately identifying key information within text sequences [31]. We believe that RoBERTa, with its superior optimization strategies compared to BERT, will undoubtedly exhibit superior performance in sentiment analysis.

2.5. Bi-LSTM

The LSTM network [14], a variant of RNNs, excels in tasks involving sequential data due to its ability to capture contextual information. It belongs to the family of artificial neural networks. LSTM networks utilize LSTM units as building blocks in their hidden layers, showcasing effectiveness in capturing long-term dependencies. These units consist of three main components [25]: the input gate i_t, the forget gate f_t, and the output gate o_t. The input gate regulates the retention of candidate stage data, determining which information is incorporated into the current internal state. The forget gate decides the degree to which information from previous time steps is forgotten, crucial for managing long sequences effectively. The output gate modulates the flow of information from the current internal state to the external state, controlling the output of the LSTM unit. The process by which the forget gate determines which information will be retained in the current neural unit can be represented as:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

The input gate uses the following equation to determine whether to discard or retain information about the data:

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t) \end{aligned} \tag{2}$$

The output gate determines the output value of the LSTM unit based on the cell state and can be represented as:

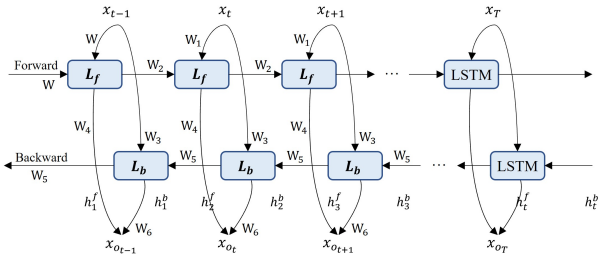

Figure 1: Bi-LSTM architecture

$$
\begin{aligned}
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
h_t &= o_t * \tanh(C_t)
\end{aligned} \tag{3}
$$

The weights for input information x are denoted by W_f, W_i, W_c and W_o, while U_i, U_f and U_o represent the weights for the previous moment $h_{(t-1)}$. The biases are represented by b_i, b_f and b_o, and t signifies time. σ is a sigmoid function ranges from 0 to 1.

In traditional LSTM, information flows unidirectionally, moving solely from the beginning to the end of a sequence. However, when handling complex linguistic structures like log sequences, there exists interdependence between preceding and subsequent information, potentially resulting in the loss of crucial contextual information within log sequences.

Bi-LSTM [18] addresses this limitation by introducing a bidirectional structure. It comprises forward and backward LSTMs, responsible for processing the input sequence in both forward (from start to end) and backward (from end to start) directions of information flow [50]. The fundamental architecture of the Bi-LSTM network is illustrated in Figure 1. The bidirectional processing strategy enables Bi-LSTM to capture contextual information from both preceding and subsequent texts. In this setup, the forward hidden layer L_f, the backward hidden layer L_b, and the output sequence X_o are employed to update the network. Iterations are performed from t to 1 for backward updates and from 1 to t for forward updates. The time step t indicates the current position within the input text sequence, where h_t^f and h_t^b are the hidden state vectors at position t during forward and backward propagations, respectively. Ultimately, these two vectors are concatenated as: $h_t = concat(h_t^f, h_t^b)$, capturing information from both forward and backward directions. The update parameters for Bi-LSTM can be represented as:

$$
\begin{aligned}
L_f &= \sigma(W_1 \cdot x_t + W_2 \cdot L_{f-1} + b_{L_f}) \\
L_b &= \sigma(W_3 \cdot x_t + W_5 \cdot L_{b-1} + b_{L_b}) \\
x_o &= W_4 \cdot L_f + W_6 \cdot L_b + b_{x_o}
\end{aligned} \tag{4}
$$

where L_f represents the forward pass, L_b represents the backward pass, and X_o denotes the final output layer; W signifies the weight coefficients, while b_{L_f}, b_{L_b}, and b_{x_o} denote the biases. σ is a sigmoid function ranges from 0 to 1. The time step t indicates the current position within the input text sequence, where h_t^f and h_t^b are the hidden state vectors at position t during forward and backward propagations, respectively. Ultimately, these two vectors are concatenated as: $h_t = concat(h_t^f, h_t^b)$, capturing information from both forward and backward directions.

3. Methodology

In this section, we elaborate on the details of our proposed sentiment analysis model.

3.1. Overview

We introduce an innovative sentiment analysis model named DLSA-MAML, as depicted in Figure 2, which includes both the MAML training phase and the classification phase. Specifically, the MAML training phase consists of four

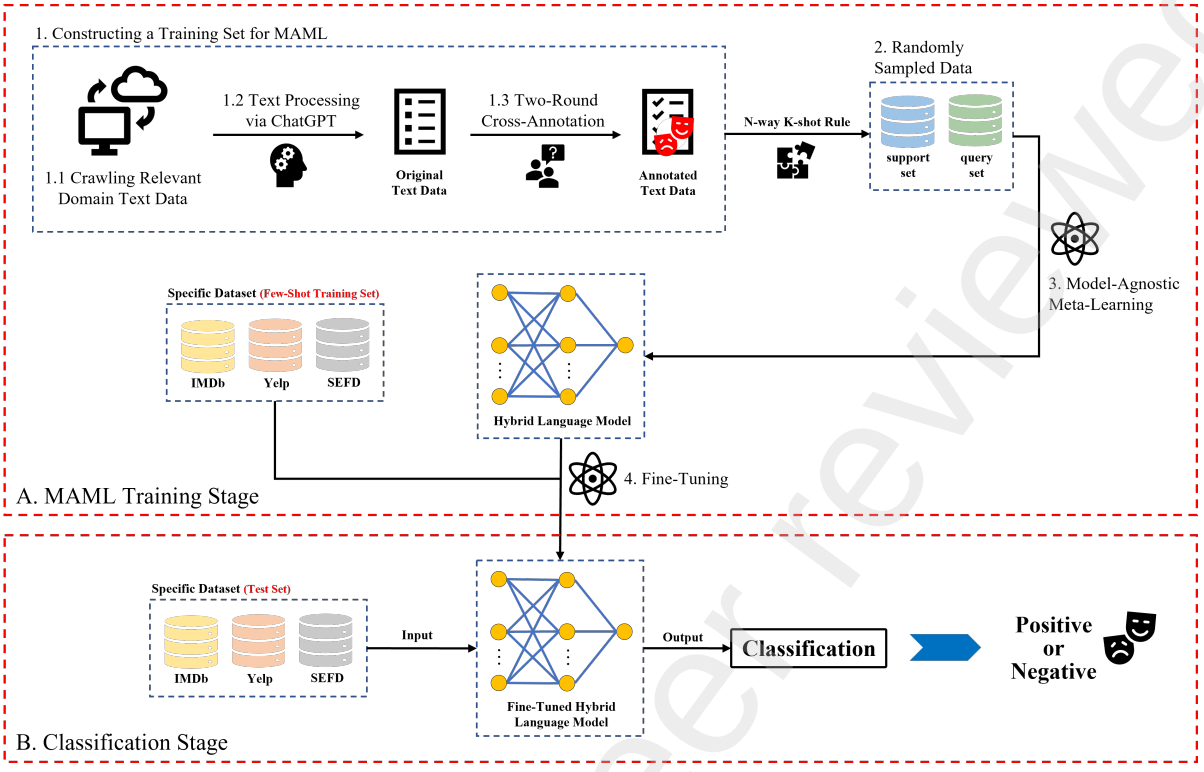

Figure 2: The overview of DLSA-MAML.

parts: (1) Constructing a dataset suitable for MAML pre-training. (2) Partitioning randomly sampled data via the n-way k-shot rule. (3) Applying MAML with a hybrid language model on the constructed dataset to learn general knowledge and features of sentiment text. (4) Fine-tuning the hybrid language model on a specific dataset (e.g., IMDb) using few-shot learning after obtaining the initialization parameters from meta-learning to achieve rapid adaptation to the target dataset's characteristics. In the classification phase, we use the fine-tuned hybrid model for sentiment analysis, classifying sentiments as positive or negative, while exploring the performance of DLSA-MAML on datasets from different domains.

It is noteworthy that, as shown in Figure 3, DLSA-MAML employs RoBERTa within the hybrid language model to encode text and generate context-related representations. These representations are then fed into an attention-based Bi-LSTM. The attention mechanism emphasizes critical information in the text by assigning different weights to various parts of the context, while the Bi-LSTM adjusts its focus based on different text segments, capturing sequential features through the fusion of forward and backward information. By integrating the strengths of each module, DLSA-MAML effectively identifies and classifies sentiments as positive or negative. The details of each phase are elaborated in the following sections.

3.2. Constructing a Training Set for MAML

In this section, we construct a training set for MAML pre-training. At the outset, we employ a web crawler to gather user comments and reviews from various websites, including Rotten Tomatoes (movie reviews), Amazon (product reviews), and Twitter (social conversations). Since this text constitutes raw data, it contains irrelevant parts (i.e., noise). To process this data, we employ ChatGPT for text cleaning. Specific prompts are designed for ChatGPT to automate the handling of the crawled raw text, removing noise such as system information, user privacy data, timestamps, punctuation marks, stopwords (e.g., "and," "the," "is"), and special characters that do not convey significant emotional meaning. Following this, we conduct two rounds of cross-annotation to filter out texts that do not meet the training requirements and to assign sentiment labels to the remaining texts.

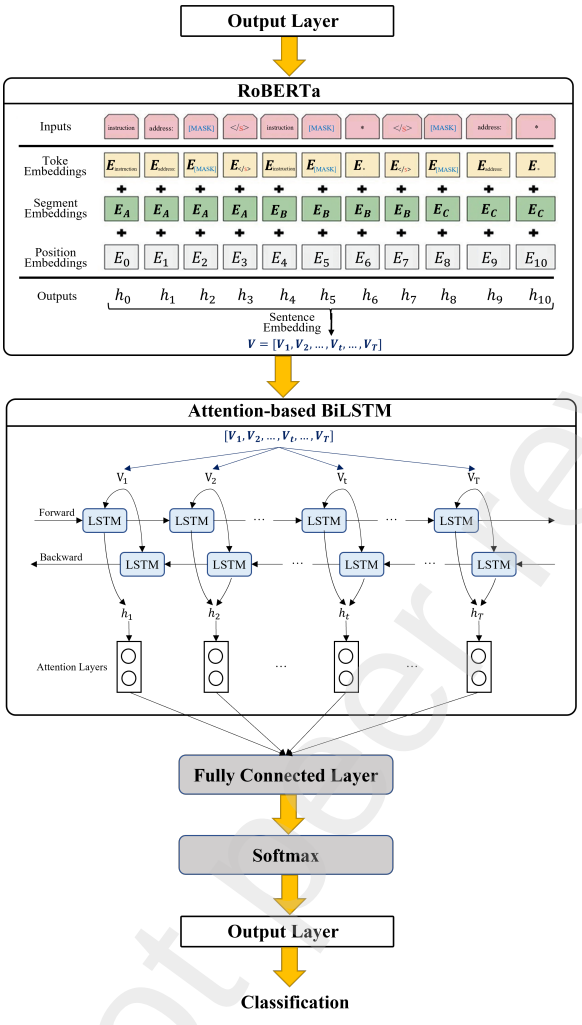

Figure 3: The architecture of proposed hybrid language model.

In detail, we initially obtain 130,183 raw data entries through the web crawler, with the data distribution as follows: 40% from Rotten Tomatoes, 40% from Amazon, and the remaining 20% from Twitter. We then apply ChatGPT for text processing. To ensure ChatGPT effectively understands the text processing task and addresses the specific types of noise, we design a comprehensive prompt template, enhanced with in-context learning, as illustrated in Figure 4, which includes: (1) Task Description, (2) Text Processing Rules, (3) Situated Learning, and (4) Input Text. Each section is tailored to the specific requirements of text processing, and we utilize ChatGPT to refine these prompts, which are detailed below:

Task Description. This section provides a detailed explanation of the text processing task assigned to ChatGPT. It includes an introduction to the task, its objectives, and the expected outcomes. Initially, we explain the text processing tasks to ChatGPT, detailing the content that needs to be processed and the desired results, specifically the final output text. To achieve these goals, we utilize two common types of ChatGPT instructions: indirect (CoT prompts) and direct instructions. Indirect instructions use phrases like "extract the main content of the comments" and "remove noise" to help ChatGPT understand the task's purpose. In contrast, direct instructions specify exactly how we want ChatGPT to generate content (e.g., "output the result directly without explanation").

Text Processing Rules. This section outlines the methodology for ChatGPT to extract and process key information from the original text content. We define specific text processing rules for ChatGPT to follow to meet the task requirements. For instance, ChatGPT is instructed to remove system information, user details, timestamps, punctuation

Prompt Template (PT)

<Task Description> \n <Text Processing Rules> \n <Situated Learning> \n <Input Text> \n

You need to process the original text of these comments by following the Text Processing Rules. Please extract the main content of the comments and remove any noise. Output the results directly without explanation.

You can refer to these text processing rules:
- Time Replacement: Replace all timestamps and specific times.
- User Details Processing: Remove all user information, including usernames, nicknames, titles, IP addresses, and similar details.
- Stop Words Processing: Remove all stop words, such as 'and,' 'the,' 'is,' etc."
- Punctuation Marks Processing: Replace or remove all punctuation marks and special characters.
(more rules…)

Here are some examples:
- Original Text: "Ed Travis (Hollywood Jesus) Apr 15, 2017: Visiting from Toronto, I was really impressed! The amount of tickets to be won was incredible. We usually got 200 tickets per play on average. Compared to Playdium here, that's awesome considering I'd win at most 10 tickets with the smallest prize costing 100 tickets. The food is great, the place is clean and the games don't feel greasy or nasty from all the food."
- Processed Text: "Visiting from Toronto I was really impressed amount tickets won incredible usually got 200 tickets per play on average compared Playdium awesome considering win at most 10 tickets smallest prize costing 100 tickets food great place clean games dont feel greasy nasty all food"
(more examples…)

Please process the text from this original text: "[TEXT]"

Figure 4: The text processing prompt template of ChatGPT.

marks, and stop words (such as “and,” “the,” and “is”). By eliminating unnecessary information, we shorten the text length, enabling the hybrid language model to more effectively capture the essential information within the text.

Situated Learning. This concept, also known as in-context learning, emphasizes the customization of prior knowledge for ChatGPT to reference while processing text, thereby enhancing its practical applicability and performance. Previous research indicates that ChatGPT performs significantly worse without any prior knowledge. Therefore, providing some prior knowledge within the prompt template is beneficial. We provide ChatGPT with examples of raw text along with their processed outputs—covering as much relevant information as possible according to the text processing rules. This approach aims to improve ChatGPT’s performance in text processing. The provision of prior knowledge prompts is referred to as "few-shot," while the absence of such knowledge constitutes a "zero-shot" setting.

Input Text. This section involves inputting the original text into ChatGPT. We sequentially feed the raw text to ChatGPT, which processes and outputs the modified text. The output text data will serve as the pre-training dataset for MAML. Due to the maximum length limit on output tokens for each request in ChatGPT, responses are terminated if the limit is exceeded. To prevent task termination due to this limitation, we implement additional settings in the program; ChatGPT skips processing such texts and provides the corresponding text identifiers.

After ChatGPT processes the text, we employ a manual annotation method to label the text data. The annotation team consists of four members, all of whom have prior experience in text annotation and possess relevant expertise, along with normal cognitive abilities. To ensure the usability of the data, we implement a two-round cross-annotation process. In this process, two members are responsible for labeling the texts. Before labeling, they verify whether ChatGPT’s processing is accurate; once confirmed, they assign positive or negative labels to the text. Upon completion of the labeling, a third member verifies the results. In cases where discrepancies arise in the annotation results, a fourth member conducts the final confirmation. We remove any ambiguous entries and duplicate text data, ultimately constructing a dataset comprising 73,565 text entries for MAML pre-training.

3.3. Model-Agnostic Meta-Learning

In this section, we adapt Model-Agnostic Meta-Learning (MAML) [12], a meta-learning method, into our DLSA-MAML model to learn general-purpose knowledge from the dataset constructed in the previous section. Initially, we construct samples using the n-way k-shot rule for random sampling. Based on the characteristics of the constructed dataset, we configure it as a 2-way 5-shot setup. This means that there are at least 5 samples from each of the 2 categories in the dataset, which we refer to as a single Task. We divide each Task into a support set and a query set. The support set contains 5 samples from each of the 2 categories, while the query set includes a larger number of samples from both categories to enhance the model’s generalization capability. The support set is used for training, while the query set is utilized for validation.

Afterward, we integrate the hybrid language model into MAML. Specifically, we use MAML to initialize the parameters of our hybrid RoBERTa and attention-based Bi-LSTM model for sentiment classification, focusing on distinguishing between positive and negative sentiments. As shown in Algorithm 1, MAML is designed to optimize the model’s capacity to quickly adapt to new sentiment analysis tasks by learning an effective initialization point through a meta-learning process that involves two key phases: the inner loop and the outer loop.

Algorithm 1 MAML for Initializing the Hybrid Language Model Parameters

1: **Input:** Task distribution $p(\mathcal{T})$, learning rate α, meta-learning rate β
2: Initialize the hybrid language model parameters θ (for RoBERTa and attention-based Bi-LSTM model)
3: **for** each iteration **do**
4: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
5: **for** each task $\mathcal{T}_i$ **do**
6: Sample K training examples $\mathcal{D}_i^{train}$ from $\mathcal{T}_i$
7: Compute task-specific loss $\mathcal{L}_{\mathcal{T}_i}(\theta)$ on $\mathcal{D}_i^{train}$
8: Compute adapted parameters with gradient descent:

$$\theta_i' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(\theta)$$

9: **end for**
10: Sample N validation examples $\mathcal{D}_i^{val}$ from each task $\mathcal{T}_i$
11: Compute meta-loss across all tasks:

$$\mathcal{L}_{meta}(\theta) = \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(\theta_i')$$

12: Update model parameters θ using meta-loss:

$$\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{meta}(\theta)$$

13: **end for**
14: **Output:** Initialized the hybrid language model parameters θ

The meta-learning process begins with the initialization of the model parameters θ. In each iteration of meta-training, a batch of tasks is sampled from the task distribution $p(\mathcal{T})$, each representing different sentiment analysis challenges. The **inner loop** of MAML focuses on fine-tuning the model for each specific task $\mathcal{T}_i$ (Line 5-9). Here, the model computes the task-specific loss $\mathcal{L}_{\mathcal{T}_i}(\theta)$ on a small set of training examples $\mathcal{D}_i^{train}$ (Line 7). Using this loss, the model parameters are adapted through gradient descent, resulting in task-specific parameters θ_i' (Line 8). This step simulates how the model would learn from a new task by focusing on the unique characteristics of each sentiment analysis task.

The **outer loop** evaluates how well the model's task-specific parameters θ_i' generalize to new data, focusing on general-purpose knowledge learning (Line 3-13). This is done by testing the adapted model on a separate validation set $\mathcal{D}_i^{val}$ for each task $\mathcal{T}_i$ (Line 10). The meta-loss $\mathcal{L}_{meta}(\theta)$ is calculated as the sum of validation losses across all tasks (Line 11), using the task-adapted parameters θ_i'. Finally, the model's original parameters θ are updated using this meta-loss (Line 12), which ensures that the learned initialization is robust and capable of quickly adapting to various sentiment analysis tasks in future iterations.

Through this iterative process of inner and outer loops, the hybrid language model is initialized in a way that significantly enhances its ability to accurately classify sentiments. This leads to improved performance in distinguishing between positive and negative sentiments, making the model more effective for real-world sentiment analysis tasks.

3.4. Hybrid Language Model

To learn the representations of text samples, we propose a Hybrid Language Model (HLM). HLM first leverages RoBERTa to comprehend the information within the text sequence. RoBERTa employs dynamic masking, generating a new masking pattern each time a text sequence is input to the model. As depicted in Figure 3, the symbol "</s>" denotes the separator for each text sequence and signifies the end of an individual text sequence. Text sequences are treated as sentence tokens to be evaluated within RoBERTa. A token represents a single input unit obtained after segmenting and adding special markers to the text. Each input word token is initially transformed into a word embedding vector W, achieved by looking up the corresponding word embedded in the vocabulary table. These word embeddings are learned during the model's pre-training, and each word possesses a unique embedding representation. Additionally, the model utilizes positional encoding to maintain the sequential order of words within the sequence.

Word embedding is created through the amalgamation of token embedding, segment embedding, and position embedding. Token embedding converts each text sequence token into a 768-dimensional vector denoted as T, segment

embedding produces vector S, and position embedding generates vector P. The fusion of these three vectors yields the embedded vector X representing the text sequence as follows:

$$X = W = T + S + P \tag{5}$$

Word embeddings produce a vector representation h for each token, and these h vectors can be integrated into a sentence-level semantic vector V_S. In RoBERTa, the CLS token is employed to encapsulate the semantic information of the entire input sequence. We use the output vector of the CLS token as the semantic vector V_S for the entire text sequence, thus obtaining the following equation:

$$V_S = [V_1, V_2, ..., V_t, ..., V_T] \tag{6}$$

DLSA-MAML uses the semantic vector sequence V_S as the model input and classifies sentiments as positive or negative using an attention-based Bi-LSTM network. The Bi-LSTM inherits the characteristics of LSTM and can capture both forward and backward contextual information flow from text sequences. Given the unique styles of each individual, the features of their comments are quite diverse. To better capture key information from negative sentiments, we introduce a "multihead" attention mechanism.

The "multihead" attention consists of eight attention heads, sequentially calculating attention scores between text sequences. To enhance the fitting ability to text sequences, three matrices are utilized within the "multihead" attention. The vector V_S is multiplied by the weight matrices W_Q, W_K, and W_V, forming three matrices, namely query matrix Q, key matrix K, and value matrix V. For each head, The semantic vector sequence is input into the self-attention function, resulting in a new vector. Their weight values are obtained using Softmax function. The specific formulas are as follows:

$$\begin{aligned} multihead &= \text{concat}(head_1, head_2, \ldots, head_n) \cdot W^O \\ head_n &= \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \end{aligned} \tag{7}$$

During the training of the hybrid language model (encompassing both the MAML training and classification stages), we use cross-entropy as the loss function, comparing the predicted outputs with the true values derived from the dataset labels. To minimize the cross-entropy loss function, we employ the Adam optimization algorithm to update the parameters of the model, which includes adjusting the weights of both the Bi-LSTM and Attention layers.

Following the attention mechanism layer, we incorporate log features via a fully connected layer. This layer consists of one or more densely connected neural network layers, primarily tasked with synthesizing the outputs from the Bi-LSTM and Attention mechanism layers. Subsequently, a Softmax layer is employed to transform the output of the fully connected layer into a probability distribution. This transformation enables the hybrid language model to make more precise predictions across positive or negative categories.

3.5. Fine-Tuning

In the fine-tuning phase, the hybrid language model is initialized with the meta-parameters θ learned during the meta-learning phase. Fine-tuning is performed on a specific dataset, such as the IMDb dataset, which contains movie reviews. Since the meta-learning phase utilized a pre-training dataset that included texts from Rotten Tomatoes—another collection of movie reviews with high textual similarity to IMDb—the model can effectively leverage the knowledge acquired during meta-learning. This similarity reduces the amount of data required for fine-tuning. Specifically, we employ a few-shot learning strategy, using only 20% of the IMDb dataset to adapt the model. This few-shot approach not only aligns the model with the unique features of the target dataset but also significantly lowers computational costs and shortens training time, enhancing efficiency.

It is important to note that MAML is not applied during the fine-tuning phase. MAML's role is to provide a robust initialization of parameters θ, optimized across multiple related tasks during the meta-learning phase. This initialization ensures that the model requires minimal fine-tuning to achieve high performance on new tasks. In this phase, the objective is to adjust the pre-trained hybrid language model to align with the specific characteristics of the target dataset without needing extensive retraining. This demonstrates the adaptability and efficiency of the DLSA-MAML framework, especially when dealing with domain-specific datasets.

3.6. Sentiment Analysis Using the Fine-Tuned Hybrid Language Model

During the classification stage, we input the text sequences from the test set into the fine-tuned hybrid language model to classify sentiments as positive or negative based on the model's output. Accurately understanding user sentiment trends can provide valuable insights for guiding the future development of both nations and enterprises.

4. Experimental Setup

4.1. Datasets

We utilize two public datasets, IMDb and Yelp Polarity, along with our proprietary dataset SEFD, to evaluate the performance of our DLSA-MAML model. These datasets cover sentiments from various domains, including movie reviews, restaurant evaluations, and student feedback. Figure 5 illustrates the sample representation from these three datasets. With the integration of MAML, we require only a few-shot sample for the fine-tuning phase; thus, we allocated 20% of the data for training and 80% for testing. The detailed descriptions of these three datasets are as follows:

Dataset	Class	Representative Sample
IMDb	Positive	One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what happened with me. The first thing that struck me about Oz was its brutality and unflinching scenes of violence, which set in right from the word GO. Trust me, this is not a show for the faint hearted or timid. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore, in the classic use of the word. It is called OZ as that is the nickname given to the Oswald Maximum Security State Penitentiary. It focuses mainly on Emerald City, an experimental section of the prison where all the cells have glass fronts and face inwards, so privacy is not high on the agenda. Em City is home to many..Aryans, Muslims, gangstas, Latinos, Christians, Italians, Irish and more....so scuffles, death stares, dodgy dealings and shady agreements are never far away. I would say the main appeal of the show is due to the fact that it goes where other shows wouldn't dare. Forget pretty pictures painted for mainstream audiences, forget charm, forget romance...OZ doesn't mess around. The first episode I ever saw struck me as so nasty it was surreal, I couldn't say I was ready for it, but as I watched more, I developed a taste for Oz, and got accustomed to the high levels of graphic violence. Not just violence, but injustice (crooked guards who'll be sold out for a nickel, inmates who'll kill on order and get away with it, well mannered, middle class inmates being turned into prison bitches due to their lack of street skills or prison experience) Watching Oz, you may become comfortable with what is uncomfortable viewing....thats if you can get in touch with your darker side.
	Negative	So im not a big fan of Boll's work but then again not many are. I enjoyed his movie Postal (maybe im the only one). Boll apparently bought the rights to use Far Cry long ago even before the game itself was even finished. People who have enjoyed killing mercs and infiltrating secret research labs located on a tropical island should be warned, that this is not Far Cry... This is something Mr Boll have schemed together along with his legion of schmucks.. Feeling loneley on the set Mr Boll invites three of his countrymen to play with. These players go by the names of Til Schweiger, Udo Kier and Ralf Moeller. Three names that actually have made them selfs pretty big in the movie biz. So the tale goes like this, Jack Carver played by Til Schweiger (yes Carver is German all hail the bratwurst eating dudes!!) However I find that Tils acting in this movie is pretty badass.. People have complained about how he's not really staying true to the whole Carver agenda but we only saw carver in a first person perspective so we don't really know what he looked like when he was kicking a**.. However, the storyline in this film is beyond demented. We see the evil mad scientist Dr. Krieger played by Udo Kier, making Genetically-Mutated-soldiers or GMS as they are called. Performing his top-secret research on an island that reminds me of "SPOILER" Vancouver for some reason. Thats right no palm trees here. Instead we got some nice rich lumberjack-woods. We haven't even gone FAR before I started to CRY (mwhehe) I cannot go on any more.. If you wanna stay true to Bolls shenanigans then go and see this movie you will not be disappointed it delivers the true Boll experience, meaning most of it will suck. There are some things worth mentioning that would imply that Boll did a good work on some areas of the film such as some nice boat and fighting scenes. Until the whole cromed/albino GMS squad enters the scene and everything just makes me laugh.. The movie Far Cry reeks of ochestee (that's poop for you simpletons) from a fa.r if you wanna take a wiff go ahead. BTW Carver gets a very annoying sidekick who makes you wanna shoot him the first three minutes he's on screen.
Yelp Polarity	Positive	Always a good time, although it's definitely not a place I would visit often. It's great for groups.... which is quite obvious since there are always several folks there for bachelorette parties and birthdays. Shut out to the piano players, who can not only play the piano, but also sing, play the drums and the sax. They are funny and know how to keep a crowd entertained. Also know some pretty obscure songs. My big criticism is the set up. If you don't have a reservation and/or come later in the evening, it's difficult to get seats. Believe me, I don't mind standing, but it does become a bit awkward when you have to keep scooting over to let waitresses and patrons past you. If you do find a seat, it is often littered with someone else's empty glasses.
	Negative	I'm writing this review to give you a heads up before you see this Doctor. The office staff and administration are very unprofessional. I left a message with multiple people regarding my bill, and no one ever called me back. I had to hound them to get an answer about my bill. Second, and most important, make sure your insurance is going to cover Dr. Goldberg's visits and blood work. He recommended to me that I get a physical, and he knew I was a student because I told him. I got the physical done. Later, I found out my health insurance doesn't pay for preventative visits. I received an $800.00 bill for the blood work. I can't pay for my bill because I'm a student and don't have any cash flow at this current time. I can't believe the Doctor wouldn't give me a heads up to make sure my insurance would cover work that wasn't necessary and was strictly preventative. The office can't do anything to help me cover the bill. In addition, the office staff said the onus is on me to make sure my insurance covers visits. Frustrating situation!
SEFD	Positive	I am extremely pleased with how the Introduction to Psychology course was conducted this semester. Dr. Emily Smith is an exceptional educator who genuinely cares about her students' understanding and growth. From the very first lecture, it was clear that she had meticulously planned the course material to ensure that each concept was presented in a logical and engaging manner. The course structure was well-organized, with each topic building on the previous one, making it easy to follow and absorb the material. Dr. Smith's teaching style is both dynamic and interactive. She incorporates a variety of teaching methods, including lectures, group discussions, and multimedia presentations, which kept the class engaged and eager to learn. I particularly appreciated the way she encouraged critical thinking and made sure that everyone felt comfortable sharing their thoughts during discussions. Her ability to connect complex psychological theories to real-life examples made the material more relatable and easier to understand. The assignments were well-designed and aligned perfectly with the course objectives. They not only reinforced what we learned in class but also challenged us to apply the concepts in practical ways. Dr. Smith provided timely and constructive feedback on all assignments, which helped me improve my understanding of the subject. Her office hours were also incredibly helpful, as she was always willing to provide additional support and clarify any doubts we had. One of the highlights of this course was the final project, which required us to conduct a small-scale psychological study. This hands-on experience was invaluable, as it allowed us to put theory into practice and gain a deeper understanding of the research process. Dr. Smith guided us through each step of the project, from formulating a research question to analyzing the data, and her expertise was evident throughout. In addition to her outstanding teaching abilities, Dr. Smith's enthusiasm for psychology is truly inspiring. Her passion for the subject is contagious, and it motivated me to delve deeper into the material and explore topics beyond what was covered in class. I feel that I have gained a solid foundation in psychology, thanks to her dedication and commitment to her students' success.
	Negative	My experience with the Introduction to Psychology course this semester has been less than satisfactory. While I recognize that psychology is a complex subject, I found the course to be disorganized and difficult to follow. Dr. Emily Smith seemed knowledgeable, but her teaching style did not resonate with me, and I struggled to keep up with the material. One of the main issues I encountered was the lack of structure in the lectures. Dr. Smith often jumped from one topic to another without clear transitions, making it challenging to understand how different concepts were connected. The lectures felt rushed, and important details were frequently glossed over. As a result, I often left class feeling confused and unsure of what was expected of us. There was also little opportunity for student interaction during lectures, which made the class feel more like a one-sided presentation rather than an engaging learning experience. The course assignments were another source of frustration. The instructions for the assignments were often vague, and I found it difficult to understand what was required. There was also a lack of alignment between the assignments and the course material, which made it hard to apply what we had learned in class. When I sought clarification from Dr. Smith, her explanations were not always helpful, and I was left feeling more confused than before. Additionally, the feedback on assignments was minimal and did not provide meaningful insights on how to improve. This made it difficult to gauge my progress in the course and left me feeling demotivated. The final project, which required us to conduct a psychological study, was particularly challenging. The project guidelines were unclear, and we were not given enough time to properly plan and execute our research. Dr. Smith's guidance during this process was limited, and I felt that I was left to navigate the project on my own. The lack of support and direction made the final project a stressful experience, and I was ultimately dissatisfied with the outcome. Another concern I had was the pace of the course. The material was covered very quickly, and there was little time to digest the information before moving on to the next topic. This made it difficult to retain what we had learned, and I often found myself falling behind. The course also lacked a clear review or summary of key concepts, which would have been helpful in reinforcing the material.

Figure 5: The representation of samples from three datasets.

IMDb. The IMDb movie reviews [9] dataset is a widely used resource for sentiment analysis, derived from the Internet Movie Database (IMDb). It includes 50,000 movie reviews, each annotated with binary sentiment labels as positive or negative. The reviews are provided as raw text strings, making the IMDb dataset a standard benchmark for binary sentiment classification tasks.

Yelp Polarity. The Yelp Reviews Polarity (Yelp Polarity) [52] dataset, which focuses on restaurant reviews, is derived from the Yelp Dataset Challenge 2015. This dataset includes a total of 598,000 reviews. Reviews with star ratings of 1 and 2 are categorized as negative polarity (Class 1), while those with star ratings of 3 and 4 are categorized as positive polarity (Class 2).

SEFD. The Student Evaluation Feedback Dataset (SEFD) comprises a total of 48,000 records, with each entry classified into either positive or negative sentiment categories. This dataset exhibits a significant class imbalance, as only 4,294 (8.95%) instances are labeled as negative, while the remaining entries are categorized as positive. The data has been preprocessed using ChatGPT to ensure consistency and quality for subsequent analysis.

4.2. Model Parameters

To implement DLSA-MAML, we adopt the version of gpt-4-turbo for the text processing and invoke their API[1] provided by OpenAI. Regarding the model structure and hyper-parameters of DLSA-MAML, we configure them as below: The hidden size for RoBERTa and input layers were both designated as 768. A learning rate of 2e-5 was selected for training, and a batch size of 32 was used. The vocabulary size was defined as 30522. Subsequently, the Bi-LSTM's hidden size was set to 256 with 4 hidden layers and 8 attention heads. Notably, the maximum sequence length was specified as 128. During the training process, the CrossEntropyLoss function was utilized in conjunction with the Adam optimizer.

[1] https://platform.openai.com/docs/introduction

To ensure robustness and mitigate randomness, we conducted each experiment five times and reported the average results. All experiments were performed on a system running Windows 11 with an Intel(R) Core(TM) i7-12700 CPU, 64GB RAM, and an NVIDIA RTX A4000 GPU.

4.3. Research Questions

This paper mainly focuses on the following four research questions and designs our experiments accordingly.

- **RQ1: What is the performance of DLSA-MAML against existing DLSA approaches?**

  In this study, we compared DLSA-MAML with eight benchmark models, which include two widely used machine learning models and six hybrid models, both those that incorporate BERT and those that do not. Due to the integration of MAML, which requires only a few-shot sample for the fine-tuning phase, we allocated 20% of the data for training and 80% for testing. It is noteworthy that the experimental results for SVM, NB, and BERT-RNN are derived from our own experiments, while the results for the other models are sourced from their respective original research.

- **RQ2: What is the contribution of each individual module within the hybrid language model?**

  In this study, we disassembled each module of the DLSA-MAML model to assess their individual contributions. RoBERTa is considered as the first module, Bi-LSTM as the second, and the attention mechanism as the third. We conducted experiments using the IMDb and SEFD datasets. The IMDb dataset is chosen for its greater data variability, while the SEFD dataset is selected due to its significant class imbalance issues. These characteristics of the datasets help in evaluating the advantages of the various modules within the hybrid language model.

- **RQ3: What improvements does MAML approach bring to the sentiment analysis performance of DLSA-MAML?**

  The generalization capabilities and flexibility of MAML enhance the cross-domain transferability and context adaptability of DLSA-MAML. In this study, we aim to explore the advantages of MAML in domain transfer and context adaptability. The entire experiment is divided into two main parts. The first part compares the performance of models with and without the MAML approach, primarily investigating the benefits of domain transfer and assessing how meta-learning improves model performance across different tasks or domains.

  The second part focuses on the meta-learning pre-training stage. In this phase, we removed training datasets that are more relevant to subsequent tasks to evaluate whether the model's performance would decline. This part primarily examines context adaptability, evaluating the model's performance in the absence of specific contextual data and its ability to adapt when faced with incomplete or suboptimal information. Specifically, we assessed the sentiment analysis performance of DLSA-MAML on the IMDb dataset after excluding the Rotten Tomatoes (movie reviews) dataset during the MAML training phase, in order to determine the model's classification accuracy for positive and negative sentiments after learning from only partially related information.

4.4. Benchmark Models

In this section, we present benchmark models for comparison with DLSA-MAML in terms of performance among various sentiment analysis models. We selected two widely used machine learning models [15, 47] and six hybrid models [4, 19, 10, 40, 5, 7], including both those that incorporate BERT [40, 5, 7] and those that do not [4, 19, 10]. All of these models have demonstrated exceptional performance across various sentiment analysis datasets.

SVM. Support Vector Machines (SVM) [15] have been widely utilized for sentiment analysis tasks, leveraging their robustness and ability to handle high-dimensional data. In this approach, Term Frequency-Inverse Document Frequency (TF-IDF) is employed as the text representation method. TF-IDF transforms text data into numerical features by quantifying the importance of each word in a document relative to a collection of documents, thereby capturing both term frequency and the inverse document frequency. The SVM model then uses these features to classify sentiment, distinguishing between positive and negative sentiments with high accuracy. However, the SVM approach with TF-IDF is not without challenges; it can be computationally intensive, and the quality of classification is highly dependent on the choice of hyperparameters and the feature representation's ability to capture nuanced sentiment expressions.

NB. Naive Bayes (NB) [47] classifiers are a popular choice for sentiment analysis due to their simplicity and efficiency. TF-IDF transforms text data into a numerical format by assessing the importance of each term within a document in relation to its occurrence across a corpus, thereby capturing both term frequency and rarity. The NB model

utilizes these TF-IDF features and applies probabilistic principles to classify sentiments as positive or negative. When combined with TF-IDF, the NB classifier benefits from its capability to handle large feature spaces while maintaining a relatively low computational cost.

CNN-LSTM. Behera et al. [4] introduced the Co-LSTM model, which integrates convolutional neural networks (CNNs) with recurrent neural networks (RNNs) to perform sentiment analysis on large-scale social media datasets. The model features a three-layer structure, beginning with an embedding layer that utilizes the Continuous Bag of Words (CBOW) approach, followed by a convolutional layer dedicated to extracting features, and an LSTM layer to capture temporal relationships. This architecture achieved high accuracy on the Twitter IMDb dataset. Nevertheless, it did not incorporate pre-trained language models such as BERT or RoBERTa, which provide more context-rich word embeddings, nor did it employ attention mechanisms to enhance the focus on relevant output vectors.

CNN-Bi-LSTM-Attention. Beakcheol Jang et al. [19] proposed a hybrid model that combines CNN and BiLSTM, augmented with an attention mechanism, to improve the accuracy of text classification. This approach uses skip-gram word embeddings to convert words into contextually based vector representations. The CNN component is utilized for extracting high-level features, while the BiLSTM is employed to capture long-term dependencies in word sequences. Leveraging these methods, the model demonstrated superior performance compared to other models, including Multi-Layer Perceptron (MLP), CNN, LSTM, and hybrid models lacking an attention mechanism.

Caps-BiLSTM. Dong et al. [10] developed Caps-BiLSTM, a model that combines capsule networks with BiLSTM, using Glove embeddings to capture the semantic nuances of words. This novel approach features a convolutional layer for extracting local features, a capsule network enhanced by an improved dynamic routing algorithm to capture global features and textual hierarchies, and a BiLSTM layer for analyzing sentiment polarity. The model demonstrated not only superior accuracy and robustness on sentiment analysis tasks but also outperformed leading existing models. However, it encountered challenges related to computational complexity, the interpretation of capsule vectors, and limitations in architectural diversity and flexibility.

DistilBERT. Sanh et al. [40] introduced DistilBERT, a more compact version of BERT created through a process called knowledge distillation. This technique transfers knowledge from a large teacher model (BERT) to a smaller student model (DistilBERT) by minimizing the discrepancy between their outputs. As a result, DistilBERT retains much of the original model's performance while significantly reducing its size and improving inference speed. It achieves an accuracy of 92.8% on the IMDb dataset, using only half the parameters and layers of BERT. However, despite these benefits, DistilBERT still demands significant training data and computational resources, and its reduced linguistic capabilities along with its dependence on the teacher model introduce new challenges.

BERT-RNN. Bello et al. [5] proposed a model that integrates BERT with Recurrent Neural Networks (RNNs) to enhance performance on text analysis tasks. The model first employs BERT to generate contextually rich word embeddings that capture extensive semantic and syntactic information. These embeddings are then fed into an RNN component to capture long-term dependencies within the text sequence. Despite its notable accuracy and robustness, BERT-RNN requires substantial computational resources for training and inference. Additionally, the model's complexity and training duration present potential challenges in practical applications.

BERT-CNN-LSTM-SVM. Cach et al. [7] proposed a hybrid approach that integrates LSTM, CNN, and SVM architectures, leveraging both BERT and Word2Vec embeddings to capture high-quality word and sentence representations. Their study emphasizes the superiority of BERT embeddings over Word2Vec in generating word vectors that effectively capture both semantic and syntactic features. The hybrid models consistently outperformed individual model approaches, with the BERT-CNN-LSTM configuration achieving a notable 93.4% accuracy on the IMDb dataset, outperforming the Word2Vec-CNN-LSTM model by 3.7%. However, these models did not incorporate attention mechanisms, such as self-attention, which could further enhance the weighting of output vectors.

4.5. Evaluation Metrics

Sentiment analysis is approached as a classification problem [46]. To evaluate the effectiveness of models in classifying sentiments as positive or negative, we employ Accuracy, Precision, Recall, and F1-score metrics. Among these, accuracy serves as the primary evaluation metric. The definitions of each metric are as follows:

Accuracy is a measure of overall performance in sentiment classification tasks. It quantifies the proportion of correctly classified texts (both positive and negative sentiments) out of all the texts evaluated by the model. A high accuracy score indicates that the model performs well in correctly identifying both positive and negative sentiments across the dataset.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Precision can be regarded as a measure of quality, particularly in binary classification tasks. It quantifies the proportion of correctly identified positive sentiment texts out of all the texts detected as positive by the model. A high precision score signifies that the model is reliable and avoids misclassifying negative sentiment texts.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall evaluates the model's capability to identify all instances of the positive sentiment class, representing the percentage of texts correctly identified as positive out of all actual positive sentiments. Recall primarily assesses whether the model can capture all positive sentiment samples, and a model with a high recall value indicates that it does not overlook many positive sentiment cases.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

The F1-score represents the harmonic mean of precision and recall. It serves as a metric for assessing the overall effectiveness of a model, particularly when dealing with datasets characterized by imbalanced positive and negative sentiment classes. The term "F1" reflects the equal weighting given to precision and recall, making it a commonly used single criterion for evaluating models in the context of sentiment analysis.

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{11}$$

Specifically, TP (True Positive) is the count of positive sentiment texts correctly detected by the model. FP (False Positive) is the count of negative sentiment texts incorrectly identified as positive. TN (True Negative) is the count of negative sentiment texts that are correctly classified. FN (False Negative) is the count of positive sentiment texts that the model failed to detect.

5. Results and Analysis

5.1. Effectiveness of DLSA-MAML (RQ1)

In this section, we conducted a comparison between DLSA-MAML and eight benchmark models, as outlined in Table 1. The benchmark models include two widely used machine learning models [15, 47] and six hybrid models [4, 19, 10, 40, 5, 7], with some incorporating BERT and others not. To summarize, the hybrid models outperform traditional machine learning models, with DLSA-MAML achieving the best results across all three datasets, demonstrating its strong generalization capability. However, due to the class imbalance in the SEFD dataset, the model's overall performance is slightly lower compared to the other two datasets.

In detail, DLSA-MAML, a hybrid language model utilizing the RoBERTa approach, outperforms the three other hybrid models using the BERT approach, namely DistilBERT, BERT-RNN, and BERT-CNN-LSTM-SVM, by an average of 4.03%, 2.13%, and 5.03% across the three datasets, respectively. Furthermore, DLSA-MAML achieves an accuracy that is, on average, 6.73%, 6.89%, and 9.26% higher than all other models across these datasets. Regarding text representation methods, while traditional approaches such as TF-IDF and CBOW are still utilized by some models like SVM and CNN-LSTM, they lag behind deep learning-based methods in terms of accuracy. The Caps-BiLSTM model using GloVe embeddings performs reasonably well on the YELP-P and SEFD datasets but does not surpass the performance of DLSA-MAML and other BERT-based models. Additionally, models with attention-based Bi-LSTM significantly outperform those using LSTM alone. Hybrid models employing BERT show a more pronounced improvement compared to those that do not use BERT; models based on BERT embeddings are more adept at understanding textual features and making more accurate classifications. It is worth noting that the class imbalance issue in the SEFD dataset significantly increases the difficulty of learning negative sentiments. While other models utilized

Table 1
The performance of different models on three datasets.

Model	Text Representations	Accuracy (%) on Each Dataset		
		IMDB	YELP-P	SEFD
SVM	TF-IDF	83.2	81.8	79.2
NB	TF-IDF	82.6	76.2	72.5
CNN-LSTM	CBOW	88.1	92.9	85.1
CNN-Bi-LSTM-Attention	Skip-Gram Embeddings	91.4	95.8	89.4
Caps-BiLSTM	Glove Embeddings	92.0	95.1	89.8
DistilBERT	BERT Embeddings	92.8	95.6	91.2
BERT-RNN	BERT Embeddings	92.1	95.7	91.9
BERT-CNN-LSTM-SVM	BERT and Word2Vec Embeddings	93.4	96.6	92.8
DLSA-MAML	RoBERTa Embeddings	**96.8**	**98.1**	**97.0**

Note: The highest accuracy for each dataset is highlighted in **bold**.

Table 2
The performance of three modules on two datasets.

Module	Accuracy (%) on Each Dataset	
	IMDB	SEFD
RoBERTa +		
(1) Bi-LSTM + Attention	96.8	97.0
(2) Attention	92.6 (↓ 4.1)	91.6 (↓ 5.4)
(3) Bi-LSTM	95.7 (↓ 1.1)	94.5 (↓ 2.5)
BERT +		
(4) Bi-LSTM + Attention	95.9 (↓ 0.9)	95.8 (↓ 1.2)
(5) Bi-LSTM + Attention	93.0 (↓ 3.8)	91.8 (↓ 5.2)

70% of the dataset for training, DLSA-MAML employed only 20% of few-shot samples. Despite this, it still achieved an accuracy that is 9.26% higher than that of other models, further demonstrating DLSA-MAML's effectiveness in handling imbalanced data and its robustness in sentiment classification tasks. MAML equips the hybrid language model with a strong initialization, enabling rapid adaptation to sentiment-specific features in the target domain. This is particularly advantageous in scenarios where the domain-specific dataset is limited.

Answer to RQ1: DLSA-MAML consistently outperforms current state-of-the-art DLAD models across diverse datasets, underscoring the superior domain transfer capability of MAML and the exceptional sentiment classification performance of the hybrid language model.

5.2. Contribution of Each Module within the hybrid language model (RQ2)

We disassembled each module within the hybrid language model of DLSA-MAML to evaluate its individual contribution. As detailed in Table 2, Line 1 represents the full set of modules in the hybrid language model. Lines 4 and 5 explore the enhancement effects of the RoBERTa module, Line 2 assesses the improvements from the Bi-LSTM module, and Line 3 investigates the enhancement effects of the attention mechanism. We conducted experiments using the IMDb and SEFD datasets, with the IMDb dataset chosen for its greater data variability and the SEFD dataset selected due to its significant class imbalance issues. These selections allowed for more extensive experimentation in this section.

Overall, the hybrid language model we designed is robust, as each module positively contributes to the model's performance in sentiment classification. Across all configurations, we observed that the introduction of the Bi-LSTM module resulted in the most significant improvement to the model's performance, followed by the incorporation of the RoBERTa module, which also provided a noticeable enhancement. Although the improvement from integrating

Table 3
The performance of models with and without MAML on two datasets.

Model	Class	IMDB				SEFD			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
DLSA-HLM	Positive	92.7 (↓ 4.1)	93 (↓ 3)	94 (↓ 4)	93 (↓ 4)	93.4 (↓ 3.6)	96	92 (↓ 6)	95 (↓ 3)
	Negative		89 (↓ 6)	91 (↓ 4)	90 (↓ 5)		87 (↓ 6)	89 (↓ 5)	88 (↓ 6)
DLSA-MAML	Positive	96.8	96	98	97	97.0	96	98	98
	Negative		95	95	95		93	94	94

the attention mechanism was not as pronounced as that of the first two, it still contributed to improving the model's sentiment classification performance to some extent. Additionally, the performance degradation observed in Line 5 compared to the RoBERTa and BERT-based configurations underscores the superior capabilities of large-scale pre-trained language models in capturing contextual information and semantic relationships within the data.

When comparing the use of the RoBERTa module to replacing the first module with BERT or removing it entirely, we observed average accuracy improvements of 1.05% and 4.5% across the two datasets, respectively. Introducing the Bi-LSTM module results in an average improvement of 4.75%, and adding an attention layer leads to an average improvement of 1.8% across the two datasets. Although DLSA-MAML achieved very similar accuracy on the two datasets, differing by only 0.2%, the class imbalance characteristics of the SEFD dataset led to a more pronounced decline in accuracy after removing its module. Specifically, the accuracy on the SEFD dataset dropped an average of 4.4% more than that on the IMDb dataset. This further highlights the challenges faced by the model in the absence of sufficient samples.

✎ **Answer to RQ2:** In summary, each module of DLSA-MAML contributes to improving sentiment classification performance, with the RoBERTa and Bi-LSTM modules showing the most significant improvements.

5.3. Contribution of MAML approach (RQ3)

5.3.1. Investigation into MAML's domain transfer effectiveness

In this section, we compare the performance of models with and without the MAML approach, focusing on the benefits of domain transfer and evaluating how meta-learning enhances model performance across different tasks or domains. Table 3 highlights the performance differences between the DLSA-HLM (without MAML) and DLSA-MAML (with MAML) models on the IMDb and SEFD datasets.

To summarize, in the context of diverse domain reviews, integrating MAML into the DLSA model significantly improves performance on both the IMDb and SEFD datasets, under both balanced and imbalanced data scenarios. This integration notably contributes to achieving more balanced metrics, such as Precision and Recall. On the IMDb dataset, the DLSA-MAML model achieves an accuracy of 96.8%, which is 4.1% higher than the DLSA-HLM model's 92.7%. Specifically, for the positive class, the recall of DLSA-MAML reaches 98%, surpassing DLSA-HLM's 94% by 4%. This improvement suggests that MAML effectively enhances the model's ability to generalize to unseen data, particularly in identifying positive sentiment.

Similarly, on the SEFD dataset, characterized by data imbalance, DLSA-MAML continues to demonstrate superior performance. The model achieves an accuracy 3.6% higher than DLSA-HLM (97.0% vs. 93.4%). For the recall metric, DLSA-MAML exhibits a particularly strong advantage in the negative class, with a recall of 94%, which is 6% higher than DLSA-HLM's 88%. The positive class also benefits, with DLSA-MAML achieving a recall of 98%, compared to DLSA-HLM's 92%, reflecting a 6% improvement. Notably, the larger decline in recall compared to precision for the DLSA-HLM model suggests that it struggles more with correctly identifying all relevant instances in the presence of data imbalance, leading to a higher rate of false negatives.

Overall, these results highlight that the MAML-enhanced model (DLSA-MAML) is more effective at adapting to new tasks and handling imbalanced data. MAML's ability to optimize model parameters for better generalization results in consistently higher recall and overall accuracy, making it particularly beneficial for datasets with challenging characteristics like the SEFD dataset.

Table 4
The impact of MAML pre-training with and without movie reviews on model performance

Model	Class	IMDB			
		Accuracy	Precision	Recall	F1-score
DLSA-MAML (Without Rotten Tomatoes)	Positive	94.4 (↓ 2.4)	96	94 (↓ 4)	95 (↓ 2)
	Negative		92 (↓ 3)	91 (↓ 4)	91 (↓ 4)
DLSA-MAML	Positive	96.8	96	98	97
	Negative		95	95	95

5.3.2. Evaluation of MAML's Proficiency in Context Adaptability

In this section, we specifically focus on evaluating the performance of the meta-learning pre-training stage and its impact on subsequent sentiment classification. By excluding the training data from Rotten Tomatoes (which consists of movie reviews) and assessing the model's sentiment classification performance on IMDb (also consisting of movie reviews), we evaluate whether the model's performance deteriorates. This approach effectively assesses MAML's domain transfer capability, specifically the model's ability to perform well in the absence of specific contextual data and its capacity to adapt to incomplete or suboptimal information. Table 4 evaluates the impact of MAML pre-training on model performance with and without the inclusion of Rotten Tomatoes movie reviews.

To summarize, the results indicate that the DLSA-MAML model, when pre-trained without Rotten Tomatoes data, shows a decrease in performance across all metrics on the IMDb dataset compared to the DLSA-MAML model pre-trained with Rotten Tomatoes data. The unchanged precision for both classes between the models suggests that while the inclusion of Rotten Tomatoes data enhances the model's recall, it does not significantly affect its ability to avoid false positives. Conversely, the DLSA-MAML model pre-trained with Rotten Tomatoes data achieves a significantly higher accuracy of 96.8%, compared to 94.4% for the model pre-trained without Rotten Tomatoes, reflecting a 2.4% drop in accuracy when Rotten Tomatoes data is excluded. This decline in performance highlights that the diverse and contextually rich pre-training data from Rotten Tomatoes helps the model generalize better across various sentiment contexts, which is crucial for maintaining high accuracy.

The recall metrics further emphasize the impact of using Rotten Tomatoes data. For the positive class, the DLSA-MAML model pre-trained with Rotten Tomatoes achieves a recall of 98%, which is 4% higher than the 94% recall of the model pre-trained without Rotten Tomatoes. This suggests that the model trained with Rotten Tomatoes data is better at identifying positive sentiments, likely due to its exposure to a broader range of movie review contexts during pre-training. Similarly, for the negative class, the recall is 95% with Rotten Tomatoes data, compared to 91% without, indicating that the diverse data helps the model capture negative sentiments more accurately.

Overall, these results underscore the effectiveness of including Rotten Tomatoes data in pre-training, as it enhances the model's ability to adapt to sentiment classification tasks, particularly in terms of recall. The drop in performance when excluding this data highlights MAML's dependence on diverse and contextually rich pre-training data for maintaining high performance.

✎ **Answer to RQ3:** Systematic experiments indicate that MAML enhances the generalization capabilities of DLSA-MAML, improving its cross-domain transferability and context adaptability. These results demonstrate the advantages and necessity of incorporating MAML into sentiment analysis.

6. Threats to Validity

This section clarifies the threats to computational complexity and scalability, privacy issue, and inconsistency in manual annotations, respectively.

6.1. Computational Complexity and Scalability

DLSA-MAML leverages both meta-learning and hybrid language models, which may require substantial computational resources, especially for large-scale datasets or real-time applications. This limitation can be mitigated through the use of more efficient hardware, such as GPUs or TPUs, and optimized parallelization techniques. Furthermore, the continuous advancement of computing technologies ensures that these challenges will be more easily addressed with newer hardware in future implementations.

6.2. Privacy Issue

From an enterprise perspective, many review datasets contain sensitive information, including extensive customer and product data, which raises concerns about privacy and confidentiality. However, DLSA-MAML is a general framework designed to support multiple languages and domains. Organizations can locally train the model on their own datasets, ensuring that sensitive data remains secure and is not shared externally, thereby mitigating privacy risks.

6.3. Inconsistency in Manual Annotations

Individual annotators may interpret and categorize text data differently due to subjective variations, leading to inconsistencies in labeling. To address this issue, we implemented a two-round cross-annotation process, allowing discrepancies to be identified, reviewed, and resolved, thereby significantly reducing annotation errors. Additionally, five domain experts were invited to evaluate the annotated dataset, and the accuracy of our labeled data received unanimous approval from these evaluators.

7. Related Work

7.1. ChatGPT for Text Preprocessing

The advent of Large Language Models (LLMs) like ChatGPT has showcased considerable promise across diverse domains, spanning conversational interactions, language understanding, and text extraction. In recent years, ChatGPT has emerged as a powerful tool for text preprocessing, leveraging its advanced natural language understanding capabilities to perform tasks such as tokenization, lemmatization, and stopword removal with high accuracy and efficiency [33]. Unlike traditional rule-based methods, which often require extensive customization and manual intervention, ChatGPT adapts to diverse text inputs, effectively handling nuances like slang, dialects, and contextual ambiguities [39]. Its ability to generate contextually appropriate corrections and normalizations significantly enhances the quality of processed text, making it an invaluable asset in sentiment analysis, machine translation, and other NLP tasks.

Moreover, ChatGPT's integration into text preprocessing workflows offers distinct advantages in terms of scalability and consistency [2]. It processes large volumes of text data rapidly, ensuring uniform preprocessing across datasets, which is critical for downstream tasks like feature extraction and model training. Studies [23] have shown that incorporating ChatGPT into preprocessing pipelines improves performance in text classification and sentiment analysis models, particularly when dealing with noisy or unstructured data. For instance, Sudirjo et al. [41] demonstrated that ChatGPT effectively normalizes informal language and corrects grammatical errors, significantly enhancing the quality of input data for subsequent natural language processing tasks. Similarly, Zhang et al. [51] explored the use of ChatGPT for text augmentation, finding that it improved model robustness by generating diverse paraphrases, which reduced the risk of overfitting in sentiment analysis models. These studies underscore the growing role of ChatGPT in refining textual data, making it an essential tool for preprocessing in various NLP applications.

However, despite its strengths, using ChatGPT for text preprocessing is not without challenges. Issues such as computational overhead and the need for large amounts of training data can limit its applicability in resource-constrained environments [28]. Furthermore, while ChatGPT excels in handling standard language forms, it may struggle with domain-specific jargon or highly technical texts without additional fine-tuning. Ongoing research is therefore focused on optimizing its efficiency and expanding its applicability across diverse domains.

7.2. Sentiment Analysis Models

Previous sentiment analysis studies have extensively utilized both traditional machine learning (ML) [15, 47] and deep learning (DL) [4, 19, 10, 40, 5, 7] algorithms across various datasets and languages. Among these approaches, Support Vector Machines (SVM) [15] have consistently demonstrated high accuracy, especially when combined with techniques like unigrams and TF-IDF. Other ML algorithms, such as Naive Bayes (NB) and Random Forest (RF), have been frequently applied as well, though their performance often falls behind that of SVM and more recent DL models.

In contrast, DL-based approaches like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Recurrent Neural Networks (RNN) have gained popularity due to their ability to handle more complex text representations, resulting in higher accuracy rates [3]. Studies by Symeonidis et al. [42], and Zhao et al. [54] demonstrate the effectiveness of CNN and LSTM across various sentiment analysis tasks, with accuracy rates reaching up to 96.81% on datasets such as Yelp and Amazon product reviews. Additionally, advanced word embedding techniques like BERT, word2vec, and GloVe have further enhanced the performance of both ML and DL models

in sentiment analysis, allowing them to capture deeper semantic relationships. Furthermore, Transformer-based models have dramatically transformed sentiment analysis by effectively capturing long-range dependencies and subtle contextual meanings [1], as exemplified by influential models like BERT, XLNet, and RoBERTa. Although these models have established new standards in various NLP tasks, their high computational demands and complexity introduce challenges in terms of efficiency and scalability. In response, models like DistilBERT have been introduced to mitigate these challenges by reducing model size and accelerating inference times, all while maintaining robust performance [30]. Nevertheless, the quest for more efficient and interpretable models continues, which has driven us to develop a novel approach.

With the advancement of NLP, hybrid models [7] have made significant strides in sentiment analysis by combining various neural network architectures, thereby enhancing the capacity of models to capture intricate linguistic patterns. Notable examples include the Co-LSTM model by Behera et al. [4], which integrates convolutional and recurrent layers for large-scale social data sentiment analysis, and the work by Jang et al. [19], which incorporates CNN, BiLSTM, and attention mechanisms to improve text classification accuracy. These models have achieved high performance on datasets like IMDb and Twitter, though many still lack integration with more advanced pre-trained language models and sophisticated attention mechanisms, limiting their potential.

Recent studies have sought to address these limitations by incorporating techniques such as capsule networks, BERT embeddings, and Transformer models into hybrid architectures. For instance, Dong et al. [10] propose a caps-BiLSTM model, while Cach et al. [7] introduce a BERT-CNN-LSTM hybrid, both of which have demonstrated notable accuracy improvements by leveraging advanced word embeddings and innovative network structures. However, challenges such as computational complexity and the need for more sophisticated feature extraction methods persist. To tackle these challenges, our research proposes a dynamic hybrid model that integrates RoBERTa with BiLSTM and self-attention mechanisms, aiming to enhance both accuracy and efficiency in sentiment analysis, improve feature extraction, and increase the interpretability of sentiment analysis models.

8. Conclusion

This paper introduces DLSA-MAML, an innovative sentiment analysis model that integrates both MAML training and classification phases. The MAML training phase includes constructing a suitable dataset, partitioning data for meta-learning, applying MAML with a hybrid language model to learn general sentiment features, and fine-tuning on specific datasets. In the classification phase, the fine-tuned hybrid language model, which uses RoBERTa for encoding and an attention-based Bi-LSTM for context-aware sentiment classification, is evaluated across various datasets. Extensive experiments demonstrate the effectiveness of DLSA-MAML, validating the principles of the hybrid language model design and confirming the enhancement effects of MAML, further highlighting its advantages in advancing sentiment classification.

In the future, sentiment classification research is poised to benefit from continued advancements in deep learning and meta-learning techniques. As models like DLSA-MAML demonstrate significant improvements in handling diverse and imbalanced datasets, future research will likely explore further enhancements in meta-learning algorithms to improve model adaptability across a broader range of domains. Incorporating more sophisticated context-aware mechanisms and leveraging emerging technologies such as large-scale pre-trained language models and advanced attention mechanisms could drive further progress in sentiment analysis. Additionally, integrating sentiment classification with real-time data streams and multimodal inputs, such as text combined with audio and visual data, may enhance the accuracy and contextual understanding of sentiment models. Emphasis on efficient model training and deployment in resource-constrained environments will also be crucial, enabling more practical applications of sentiment analysis in various industries. Overall, the future of sentiment classification will likely see a convergence of cutting-edge techniques and practical applications, pushing the boundaries of what can be achieved in understanding and interpreting human emotions.

Acknowledgment

This work is supported in part by the General Research Fund of the Research Grants Council of Hong Kong and the research funds of the City University of Hong Kong (6000796, 9229109, 9229098, 9220103, 9229029).

References

[1] Acheampong, F.A., Nunoo-Mensah, H., Chen, W., 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. Artificial Intelligence Review 54, 5789–5829.

[2] Al-Gaphari, G., AL-Hagree, S., Al-Helali, B., 2023. Investigating the impact of utilizing the chatgpt for arabic sentiment analysis, in: International Conference of Reliable Information and Communication Technology, Springer. pp. 93–107.

[3] Atandoh, P., Zhang, F., Adu-Gyamfi, D., Atandoh, P.H., Nuhoho, R.E., 2023. Integrated deep learning paradigm for document-based sentiment analysis. Journal of King Saud University-Computer and Information Sciences 35, 101578.

[4] Behera, R.K., Jena, M., Rath, S.K., Misra, S., 2021. Co-lstm: Convolutional lstm model for sentiment analysis in social big data. Information Processing & Management 58, 102435.

[5] Bello, A., Ng, S.C., Leung, M.F., 2023. A bert framework to sentiment analysis of tweets. Sensors 23, 506.

[6] Chowdhary, K., Chowdhary, K., 2020. Natural language processing. Fundamentals of artificial intelligence , 603–649.

[7] Dang, C.N., Moreno-García, M.N., De la Prieta, F., 2021. Hybrid deep learning models for sentiment analysis. Complexity 2021, 9986920.

[8] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

[9] Dodds, K., 2006. Popular geopolitics and audience dispositions: James bond and the internet movie database (imdb). Transactions of the Institute of British Geographers 31, 116–130.

[10] Dong, Y., Fu, Y., Wang, L., Chen, Y., Dong, Y., Li, J., 2020. A sentiment analysis method of capsule network based on bilstm. IEEE access 8, 37014–37020.

[11] Ferguson, P., O'Hare, N., Davy, M., Bermingham, A., Sheridan, P., Gurrin, C., Smeaton, A.F., 2009. Exploring the use of paragraph-level annotations for sentiment analysis of financial blogs .

[12] Finn, C., Abbeel, P., Levine, S., 2017. Model-agnostic meta-learning for fast adaptation of deep networks, in: International conference on machine learning, PMLR. pp. 1126–1135.

[13] Fu, Y., Hao, J.X., Li, X., Hsu, C.H., 2019. Predictive accuracy of sentiment analytics for tourism: A metalearning perspective on chinese travel news. Journal of Travel Research 58, 666–679.

[14] Graves, A., Graves, A., 2012. Long short-term memory. Supervised sequence labelling with recurrent neural networks , 37–45.

[15] Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intelligent Systems and their applications 13, 18–28.

[16] Huang, B., Guo, R., Zhu, Y., Fang, Z., Zeng, G., Liu, J., Wang, Y., Fujita, H., Shi, Z., 2022. Aspect-level sentiment analysis with aspect-specific context position information. Knowledge-Based Systems 243, 108473.

[17] Huang, X., Li, J., Wu, J., Chang, J., Liu, D., 2023. Transfer learning with document-level data augmentation for aspect-level sentiment classification. IEEE Transactions on Big Data .

[18] Huang, Z., Xu, W., Yu, K., 2015. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 .

[19] Jang, B., Kim, M., Harerimana, G., Kang, S.u., Kim, J.W., 2020. Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism. Applied Sciences 10, 5841.

[20] Jawahar, G., Sagot, B., Seddah, D., 2019. What does bert learn about the structure of language?, in: ACL 2019-57th Annual Meeting of the Association for Computational Linguistics.

[21] Jia, X., Li, C., Zeng, M., Wang, L., Mi, Q., 2023. An improved unified domain adversarial category-wise alignment network for unsupervised cross-domain sentiment classification. Engineering Applications of Artificial Intelligence 126, 107108.

[22] Jin, W., Zhao, B., Zhang, Y., Huang, J., Yu, H., 2024. Wordtransabsa: enhancing aspect-based sentiment analysis with masked language modeling for affective token prediction. Expert Systems with Applications 238, 122289.

[23] Katib, I., Assiri, F.Y., Abdushkour, H.A., Hamed, D., Ragab, M., 2023. Differentiating chat generative pretrained transformer from humans: detecting chatgpt-generated text and human text using machine learning. Mathematics 11, 3400.

[24] Khan, A., Baharudin, B., Khan, K., 2011. Sentiment classification using sentence-level lexical based. Trends in Applied Sciences Research 6, 1141–1157.

[25] Kulshrestha, A., Krishnaswamy, V., Sharma, M., 2020. Bayesian bilstm approach for tourism demand forecasting. Annals of tourism research 83, 102925.

[26] Lei, Y., Li, Y., 2023. A novel scheme of domain transfer in document-level cross-domain sentiment classification. Journal of Information Science 49, 567–581.

[27] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 .

[28] Lossio-Ventura, J.A., Weger, R., Lee, A.Y., Guinee, E.P., Chung, J., Atlas, L., Linos, E., Pereira, F., 2024. A comparison of chatgpt and fine-tuned open pre-trained transformers (opt) against widely used sentiment analysis tools: sentiment analysis of covid-19 survey data. JMIR Mental Health 11, e50150.

[29] Lv, Y., Wei, F., Cao, L., Peng, S., Niu, J., Yu, S., Wang, C., 2021. Aspect-level sentiment analysis using context and aspect memory network. Neurocomputing 428, 195–205.

[30] Mao, Y., Zhang, Y., Jiao, L., Zhang, H., 2022. Document-level sentiment analysis using attention-based bi-directional long short-term memory network and two-dimensional convolutional neural network. Electronics 11, 1906.

[31] Mengi, R., Ghorpade, H., Kakade, A., . Fine-tuning t5 and roberta models for enhanced text summarization and sentiment analysis .

[32] Mercha, E.M., Benbrahim, H., 2023. Machine learning and deep learning for sentiment analysis across languages: A survey. Neurocomputing 531, 195–216.

[33] Mujahid, M., Rustam, F., Shafique, R., Chunduri, V., Villar, M.G., Ballester, J.B., Diez, I.d.l.T., Ashraf, I., 2023. Analyzing sentiments regarding chatgpt using novel bert: A machine learning approach. Information 14, 474.

[34] Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W., 2011. Natural language processing: an introduction. Journal of the American Medical Informatics Association 18, 544–551.
[35] Park, J., 2023. Combined text-mining/dea method for measuring level of customer satisfaction from online reviews. Expert Systems with Applications 232, 120767.
[36] Ping, Z., Sang, G., Liu, Z., Zhang, Y., 2024. Aspect category sentiment analysis based on prompt-based learning with attention mechanism. Neurocomputing 565, 126994.
[37] Pipalia, K., Bhadja, R., Shukla, M., 2020. Comparative analysis of different transformer based architectures used in sentiment analysis, in: 2020 9th international conference system modeling and advancement in research trends (SMART), IEEE. pp. 411–415.
[38] Rhanoui, M., Mikram, M., Yousfi, S., Barzali, S., 2019. A cnn-bilstm model for document-level sentiment analysis. Machine Learning and Knowledge Extraction 1, 832–847.
[39] Roumeliotis, K.I., Tselikas, N.D., 2023. Chatgpt and open-ai models: A preliminary review. Future Internet 15, 192.
[40] Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 .
[41] Sudirjo, F., Diantoro, K., Al-Gasawneh, J.A., Azzaakiyyah, H.K., Ausat, A.M.A., 2023. Application of chatgpt in improving customer sentiment analysis for businesses. Jurnal Teknologi Dan Sistem Informasi Bisnis 5, 283–288.
[42] Symeonidis, S., Effrosynidis, D., Arampatzis, A., 2018. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Systems with Applications 110, 298–310.
[43] Taboada, M., 2016. Sentiment analysis: An overview from linguistics. Annual Review of Linguistics 2, 325–347.
[44] Vanschoren, J., 2019. Meta-learning. Automated machine learning: methods, systems, challenges , 35–61.
[45] Venugopalan, M., Gupta, D., 2022. A reinforced active learning approach for optimal sampling in aspect term extraction for sentiment analysis. Expert Systems with Applications 209, 118228.
[46] Wankhade, M., Rao, A.C.S., Kulkarni, C., 2022. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review 55, 5731–5780.
[47] Webb, G.I., Keogh, E., Miikkulainen, R., 2010. Naïve bayes. Encyclopedia of machine learning 15, 713–714.
[48] Wu, F., Wu, S., Huang, Y., Huang, S., Qin, Y., 2016. Sentiment domain adaptation with multi-level contextual sentiment knowledge, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 949–958.
[49] Yan, C., Liu, J., Liu, W., Liu, X., 2022. Research on public opinion sentiment classification based on attention parallel dual-channel deep learning hybrid model. Engineering Applications of Artificial Intelligence 116, 105448.
[50] Yildirim, Ö., 2018. A novel wavelet sequence based on deep bidirectional lstm network model for ecg signal classification. Computers in biology and medicine 96, 189–202.
[51] Zhang, B., Yang, H., Zhou, T., Ali Babar, M., Liu, X.Y., 2023. Enhancing financial sentiment analysis via retrieval augmented large language models, in: Proceedings of the fourth ACM international conference on AI in finance, pp. 349–356.
[52] Zhang, X., Zhao, J., LeCun, Y., 2015. Character-level convolutional networks for text classification. Advances in neural information processing systems 28.
[53] Zhang, Y., Wang, J., Zhang, X., 2021. Conciseness is better: Recurrent attention lstm model for document-level sentiment analysis. Neurocomputing 462, 101–112.
[54] Zhao, W., Guan, Z., Chen, L., He, X., Cai, D., Wang, B., Wang, Q., 2017. Weakly-supervised deep embedding for product review sentiment analysis. IEEE Transactions on Knowledge and Data Engineering 30, 185–197.
[55] Zhu, W., Qiu, J., Yu, Z., Luo, W., 2024. A survey on personalized document-level sentiment analysis. Neurocomputing , 128449.