information

MDPI

Article

Prompt-Based Few-Shot Text Classification with Multi-Granularity Label Augmentation and Adaptive Verbalizer

Deling Huang [1,*], Zanxiong Li [1], Jian Yu [2,*] and Yulong Zhou [1]

[1] School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; s231231037@stu.cqupt.edu.cn (Z.L.); zhouyl@cqupt.edu.cn (Y.Z.)
[2] Department of Computer Science, Auckland University of Technology, Auckland 1010, New Zealand
* Correspondence: huangdl@cqupt.edu.cn (D.H.); jian.yu@aut.ac.nz (J.Y.)

Abstract

Few-Shot Text Classification (FSTC) aims to classify text accurately into predefined categories using minimal training samples. Recently, prompt-tuning-based methods have achieved promising results by constructing verbalizers that map input data to the label space, thereby maximizing the utilization of pre-trained model features. However, existing verbalizer construction methods often rely on external knowledge bases, which require complex noise filtering and manual refinement, making the process time-consuming and labor-intensive, while approaches based on pre-trained language models (PLMs) frequently overlook inherent prediction biases. Furthermore, conventional data augmentation methods focus on modifying input instances while overlooking the integral role of label semantics in prompt tuning. This disconnection often leads to a trade-off where increased sample diversity comes at the cost of semantic consistency, resulting in marginal improvements. To address these limitations, this paper first proposes a novel Bayesian Mutual Information-based method that optimizes label mapping to retain general PLM features while reducing reliance on irrelevant or unfair attributes to mitigate latent biases. Based on this method, we propose two synergistic generators that synthesize semantically consistent samples by integrating label word information from the verbalizer to effectively enrich data distribution and alleviate sparsity. To guarantee the reliability of the augmented set, we propose a Low-Entropy Selector that serves as a semantic filter, retaining only high-confidence samples to safeguard the model against ambiguous supervision signals. Furthermore, we propose a Difficulty-Aware Adversarial Training framework that fosters generalized feature learning, enabling the model to withstand subtle input perturbations. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods on most few-shot and full-data splits, with F1 score improvements of up to +2.8% on the standard AG's News benchmark and +1.0% on the challenging DBPedia benchmark.

Keywords: few-shot learning; text classification; prompt tuning; data augmentation; adversarial training

Academic Editors: Diego Reforgiato Recupero and Thomas Mandl

Received: 26 November 2025
Revised: 25 December 2025
Accepted: 3 January 2026
Published: 8 January 2026

1. Introduction

Despite the rapidly increasing diversity of online textual content, the scarcity of high-quality training samples is still a major challenge for natural language understanding (NLU) tasks. This issue primarily stems from the high costs of acquiring annotated data, which often requires expert assessment to avoid noisy or mislabeled instances [1]. Consequently, the lack of sufficient labeled data limits model robustness and generalization.

In response to this challenge, prompt tuning has emerged as a new paradigm, leveraging PLMs to adapt to downstream tasks by recalling patterns from their vast pre-training corpora [2,3]. Recent research demonstrates that employing natural language prompts to guide models in recalling similar patterns from pre-training corpora can significantly improve prediction accuracy [4–7]. However, the verbalizer as a core component of prompt tuning is typically handcrafted or constructed via gradient descent, which has limited coverage and a high level of variance [4,8–10]. To address this issue, Ref. [4] proposed a method to leverage external knowledge bases (KBs) for verbalizer construction. However, KB-based methods often struggle with complex contexts and are limited by the static nature of the knowledge source. Alternatively, Ref. [11] proposed a method that uses the intrinsic knowledge of PLMs to estimate class probability distributions. While promising, this method generates a global candidate set for the entire dataset, turning the selection of class-specific label words into a computationally expensive combinatorial problem. Furthermore, automated selection methods based on non-parametric statistics [12] rely heavily on the prior probabilities of label words predicted by PLMs. These priors are often skewed, leading to significant prediction bias in the final verbalizer [9,13,14].

Another efficient strategy for FSTC is data augmentation. Ref. [15] proposed Easy Data Augmentation (EDA), a technique including synonym replacement, random insertion, swapping, and deletion. While generally effective, Refs. [16,17] highlight that applying EDA indiscriminately can degrade performance, as operations like random swapping may corrupt label-critical information. For instance, in sentiment analysis, altering "The scenery was breathtaking, but the hotel service was terrible" via swapping could result in "The hotel service was breathtaking, but the scenery was terrible," thereby inadvertently flipping the sentiment label. Consequently, Ref. [18] proposed transforming data using category-specific identification information to preserve label integrity. However, Ref. [19] demonstrated that conventional augmentation yields marginal gains in prompt-based settings. This limitation essentially arises because traditional methods focus on modifying instances while overlooking the crucial role of label semantics [20,21]. Furthermore, these methods often fail to balance the diversity of augmented samples with their semantic consistency; overemphasizing diversity introduces noise, while excessive consistency limits generalization. Incorporating label semantics while optimizing this trade-off is critical for success in prompt-based learning [10,22,23].

To address these limitations, this paper proposes a novel framework named Adaptive Multi-granularity Label Data Augmentation (AMLDA) for single-label FSTC. Notably, while the downstream task involves assigning a single class label to each input, our augmentation strategy exploits multiple label words (tokens) associated with that class. By incorporating a diverse set of label words derived from our adaptive verbalizer, we enrich the semantic representation of the single ground-truth category, thereby mitigating data sparsity without confusing the model with incorrect class signals. First, to generate a label set covering diverse granularities while correcting prior distribution shifts, we propose a Bayesian Mutual Information (BMI)-Driven Adaptive Verbalizer Construction method. Specifically, rather than relying solely on pre-trained weights, we treat the PLM's zero-shot prediction probability as a prior. We then refine the joint probability of label words and classes via Bayes' rule, incorporating their empirical co-occurrence in the candidate set. This approach subsequently employs BMI to evaluate the quality of candidate words retrieved by the PLM. Second, building on the finding that incorporating class labels into sequences is an effective adaptation strategy [24], we design a Prompt-Template-Guided Multi-Granularity Label Data Augmentation method. This module integrates original texts with multi-granularity label sets via templates, generating semantically consistent samples that minimize noise. Furthermore, we introduce a Difficulty-Aware Adversarial

Training framework. Unlike traditional adversarial strategies that use fixed perturbation sizes—which can harm generalization [25,26]—our approach dynamically generates adversarial samples based on sample difficulty (entropy). This balances diversity and robustness, creating a dynamic learning environment suitable for data-scarce scenarios.

In this paper, we make the following key contributions to the field of FSTC:

(1) We propose a BMI-Driven Adaptive Verbalizer Construction method that combines Bayesian-inspired re-weighting strategies with mutual information. By formalizing the probability refinement process as $P_{refined}(w, c) \propto P_{PLM}(w|c) \cdot P_{freq}(w, c)$, our method dynamically integrates the general linguistic knowledge of PLMs with the specific data distribution of the few-shot task. This effectively alleviates biases in prior knowledge and constructs a robust adaptive verbalizer.

(2) We developed a prompt-template-guided multi-granularity label data augmentation approach, aiming to generate high-quality semantic-coherent samples. Additionally, we introduced an innovative data augmentation training framework that integrates adversarial training strategies and Difficulty-Aware Learning to address the inherent complexity of FSTC tasks and enhance model robustness. This method demonstrates great potential in enhancing other fine-grained natural language understanding tasks.

(3) Extensive experiments on four benchmark datasets demonstrate that our method outperforms existing FSTC models and fine-tuning strategies, achieving state-of-the-art results.

The remainder of this paper is organized as follows: Section 2 reviews related work on few-shot learning and prompt tuning. Section 3 details the proposed AMLDA framework, focusing on the adaptive verbalizer and multi-dimensional augmentation strategies. Section 4 presents comprehensive experimental evaluations, including comparative analyses and ablation studies. Finally, Section 5 concludes the paper and discusses future directions.

2. Related Work

This section examines multiple research areas pertinent to this study, primarily covering few-shot text classification and prompt tuning for PLMs.

2.1. Few-Shot Text Classification

Few-shot text classification entails training or fine-tuning the model using only a limited set of examples before performing classification [27–29]. This task is generally associated with the following paradigms.

Metric-based and model-based meta-learning methods for FSTC enable models to quickly adjust to new tasks with scarce data, using general knowledge acquired from a range of tasks. For instance, Ref. [22] introduces the metric-based meta-learner parameter generator network, which uses external knowledge to generate relational network parameters. Additionally, Ref. [30] enhances the semantic representation of samples by utilizing self-supervised tasks in the inner loop. Although effective in specific settings, these meta-learning approaches fundamentally rely on the availability of a large number of auxiliary tasks for pre-training. As noted by [31], this requirement necessitates significant computational resources, and, more critically, their performance often degrades significantly when the target domain distribution shifts away from the source tasks. Unlike these methods, our AMLDA is a task-agnostic framework that does not require extensive meta-training on auxiliary tasks.

Semi-supervised learning for FSTC utilizes the hidden information in unlabeled data to support the learning process of a small labeled dataset [32,33]. For instance, Ref. [34] presents an approach that predicts negative pseudo-labels for unlabeled data using an itera-

tive exclusion method. However, a primary limitation of such pseudo-labeling strategies is the error accumulation problem. The effectiveness of these methods heavily depends on the initial quality of the classifier; incorrect pseudo-labels can reinforce model biases, leading to performance deterioration—a risk that is exacerbated in few-shot scenarios where the initial supervision is extremely weak. In contrast, our approach enriches data via controlled augmentation rather than relying on potentially noisy unlabeled data.

Weakly supervised learning for FSTC focuses on learning using noisy, incomplete, or imperfectly labeled data, including rough or indirect annotation information [35]. For instance, Ref. [36] uses language models to generate pseudo-labels and refines them iteratively, while Ref. [37] proposes ranking potential class words using statistical measures for open-world classification. Nevertheless, these weakly supervised methods inherently suffer from the "noise propagation" issue. The initial signals (e.g., seed words or rough rules) are often imprecise, and without sufficient labeled data for correction, the iterative refinement process may converge to suboptimal representations or overfit to the initial noise. In contrast, our AMLDA mitigates this uncertainty by grounding the verbalizer construction in the robust, pre-trained knowledge of PLMs, refined by Bayesian formalism to ensure alignment with the specific task semantics.

2.2. Prompt Tuning for PLMs

Despite numerous PLMs achieving outstanding performance in diverse natural language processing tasks, in various downstream tasks based on fine-tuning PLMs, the prior knowledge within PLMs is not fully utilized [38]. Prompt tuning offers an effective approach to solving this problem. In this paradigm, rather than adapting PLMs to downstream tasks through objective engineering, downstream tasks are restructured to resemble those encountered during the initial language model training, utilizing a textual prompt [10,39]. A series of studies suggest that selecting the right prompts can modify model behavior, allowing the PLMs to generate the desired output without requiring extra task-specific training [40–42]. This effectiveness is primarily influenced by the design of prompt templates and the verbalizer construction, as these elements determine the extent to which the PLMs can be guided to achieve the desired output.

In the existing prompt tuning methods, templates typically involve soft templates and hard templates. Soft templates are continuous, learnable representations that adapt to different tasks during training. For example, the SKP [43] method employs BiLSTM and other neural networks to train soft prompts. However, introducing such learnable parameters requires gradient updates and sufficient data, which poses a significant risk of overfitting in extreme few-shot scenarios. On the other hand, hard templates are discrete, pre-defined text prompts composed of specific words or phrases. For example, Ref. [44] introduced a method named KPT, which argues that hard templates are simpler to construct, contain rich expert knowledge, and offer better interpretability. Aligning with KPT's insight, we adopt hard templates in our framework. This choice avoids the complexity of training soft tokens and ensures that the templates directly convey clear semantic instructions, which is crucial for maximizing the utilization of PLM's pre-trained knowledge with limited samples.

In addition to template design, prompt verbalizer engineering seeks to identify a label space and establish a mapping to the original output. The typical methods used to construct the verbalizer are manual design, discrete answer search, and continuous answer search. For instance, Ref. [45] manually constructed word lists associated with relevant topics. However, such manually designed verbalizers may not be optimal for leveraging the language model's full predictive potential [10]. To automate this, methods like discrete search [46] and continuous search [47] have been proposed. Furthermore, KPT [44] en-

hances verbalizer construction by incorporating external knowledge bases (KBs). Although incorporating external knowledge expands the candidate set, KPT suffers from the static nature of KBs. The retrieved words may not always align with the specific semantic context of the input text, and maintaining or querying large-scale KBs introduces additional computational overhead. These methods typically require complex filtering operations to remove noise. In contrast, to reduce computational costs while ensuring context sensitivity, our approach constructs an adaptive verbalizer without external dependencies. We utilize a Bayesian-Mutual-Information-based algorithm that dynamically refines label mappings based on the PLM's intrinsic probability distribution.

The theoretical foundation of our metric, Mutual Information (MI) has long been established as a standard criterion for feature selection in text classification [48]. It quantifies the expected information gain about a category provided by the presence of a specific word. Classical approaches typically estimate MI based on empirical frequency counts in large corpora or employ non-parametric estimators like k-Nearest Neighbors (k-NN) to approximate joint distributions [49]. However, these traditional estimations rely heavily on large-scale statistics to function reliably; in few-shot settings, sparse data leads to high variance and inaccurate MI estimates. Our work bridges this gap by integrating Bayesian-inspired refinement with MI. By treating the PLM's zero-shot predictions as a prior, we enable a robust estimation of word-class dependencies even when empirical evidence is extremely scarce.

3. Methodology

In this section, we provide a detailed introduction to the structure of the proposed AMLDA framework, as shown in Figure 1. This framework consists of three primary components: (1) BAVC, which constructs the verbalizer; (2) PTMD-LDA, which generates and filters augmented samples (incorporating the Low-Entropy Selector, LES); and (3) Difficulty-Aware Adversarial Training (DAAT). By leveraging Bayesian refinement and mutual information calculations, our BAVC constructs a verbalizer for each dataset. Subsequently, two synergistically working generators, LEG and PGG, are employed to generate semantically consistent samples, which are then filtered through LES. Finally, DAAT is developed to better address challenges related to model robustness and the enhancement of sample diversity.

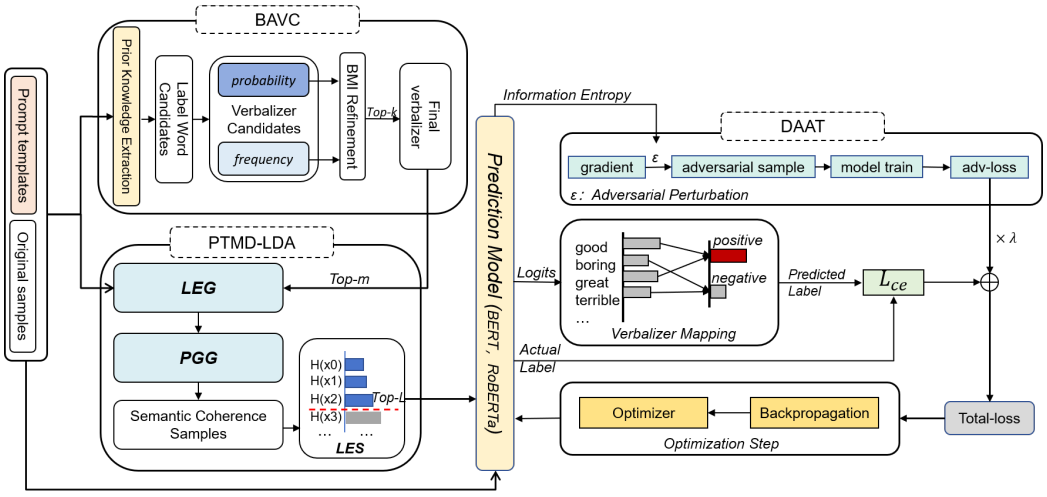


Figure 1. The framework of the proposed AMLDA.

3.1. BMI-Driven Adaptive Verbalizer Construction

In prompt tuning, the verbalizer refers to the process of mapping label words to their respective categories. This approach has been shown to effectively bridge the gap between text and label spaces, thereby enhancing the performance of downstream tasks [44]. The proposed BAVC method involves three coherent stages, as overviewed in Figure 1 and detailed in Figure 2. First, it leverages the prior knowledge of PLMs to search potential candidate words for each category. The [CLS] and [SEP] tokens in the input sample are two special markers in PLMs, where [CLS] denotes the first token of the input sequence, and [SEP] is used to separate different sentences or paragraphs. Second, it applies a Bayesian Mutual Information (BMI)-based refinement strategy that incorporates the prior probabilities provided by the PLMs while capturing the dependency relationships between candidate words and their respective categories. Finally, it employs a dynamic adaptive k-value selection mechanism, analyzing BMI score sequences to automatically determine the optimal verbalizer size for each category, ensuring semantic consistency and efficiency.

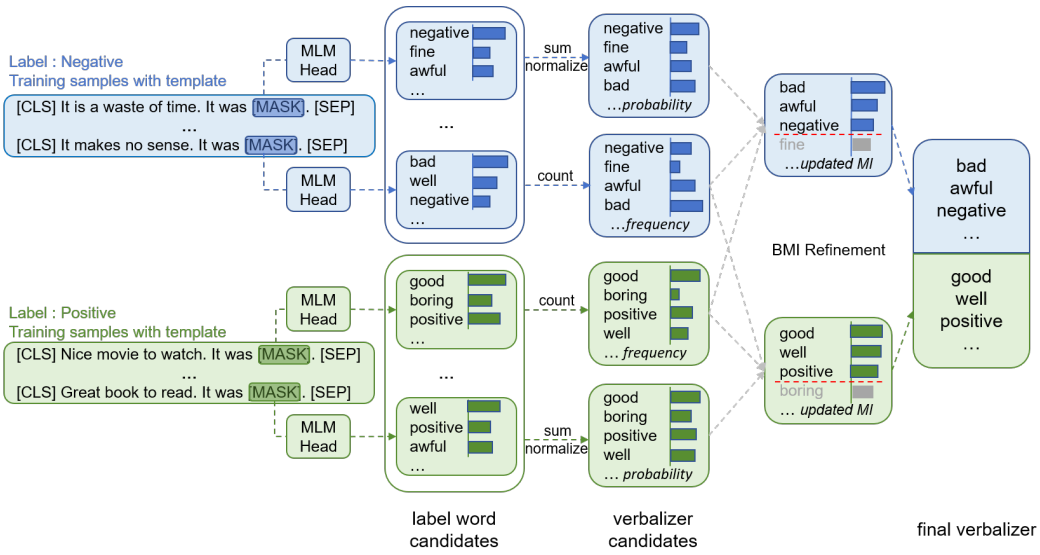


Figure 2. The proposed verbalizer construction approach.

3.1.1. PLM-Based Candidate Word Search

Manually constructed verbalizers lack coverage and are time-consuming, so we leverage the prior knowledge of PLMs to assist in verbalizer construction, a process illustrated as Prior Knowledge Extraction in Figure 1. Define the set of classification classes $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$, where each C_i corresponds to a class, such as "positive" or "negative". Let $F : C \rightarrow V_C$ denote the one-to-multiple verbalizer that maps each class $C_i \in C$ to a set of label words in the vocabulary $V_{C_i} = \{v_1^{C_i}, v_2^{C_i}, \ldots, v_{k_{C_i}}^{C_i}\} \subset V_C$, where $k_{C_i} = |V_{C_i}|$ denotes the number of selected label words for each class. We aim to search for a set of label word candidates $\tilde{V}_C$ that cover different granularity and perspectives. Let D_C^{train} represent the subset of training data pertaining to the class C_i. $T(x)$ represents applying a fixed template T to the input x. The position of the token [MASK] in the input x is denoted by $Po([MASK])$. The corresponding probability score for each token in the vocabulary to fill in $Po([MASK])$ during PLMs inference can be expressed as

$$P(v \mid T(x)) = P(\text{Po}([\text{MASK}]) = v \mid T(x)) \quad (1)$$

The set of label word candidates $\tilde{V}_C$ contains candidate label words for each class, where each v is associated with a predicted probability $P(v \mid T(x))$.

3.1.2. Bayesian Mutual Information Refinement

PLMs exhibit a prediction bias for certain label words, while some label words are more difficult to predict. To address this, AMLDA introduces a novel refinement method: BMI Refinement. Specifically, the prior probabilities provided by the PLMs, denoted as $\Pr(v \mid C_i)$, are treated as the initial estimates of the word probability distribution. These estimates serve as a prior distribution over each label word v in the candidate set $\tilde{V}_C$.

By using a Bayesian-inspired re-weighting mechanism, we optimize the prior probabilities of these label words. It is worth clarifying that in our framework, both the label words v and class categories C_i are formally treated as discrete random variables. To ensure rigorous calculation, we estimate the required probabilities via Maximum Likelihood Estimation (MLE) based on the frequency statistics of the generated candidate sets for each class. The empirical probabilities are calculated as follows:

$$P(v, C_i) = \frac{\text{count}(v, C_i)}{N_{\text{total}}} \tag{2}$$

$$P(C_i) = \sum_{v \in \tilde{V}_C \cup V_{\text{other}}} P(v, C_i) \tag{3}$$

$$P(v) = \sum_{C_k \in \mathcal{C}} P(v, C_k) \tag{4}$$

where $\text{count}(v, C_i)$ denotes the frequency with which the candidate word v appears in the candidate set generated for class C_i, and $N_{\text{total}} = \sum_{v,C_i} \text{count}(v, C_i)$ is the total number of candidate word occurrences across all classes. Note that $P(C_i)$ and $P(v)$ represent the marginal probabilities derived from the joint distribution.

Subsequently, inspired by information-theoretic concepts, mutual information is used to measure the dependency between two random variables. The refined mutual information $MI_u(v, C_i)$ between a word v and class C_i can be computed as follows:

$$MI_{\text{u}}(v, C_i) = P_u(v, C_i) \log \frac{P_u(v, C_i)}{P_u(v) P_u(C_i)} \tag{5}$$

Here, $P_u(v, C_i)$ denotes the re-weighted joint probability distribution. Direct application of frequency statistics is unreliable in few-shot settings. Therefore, we employ a Bayesian-inspired re-weighting strategy that treats the PLM's zero-shot prediction probability $Pr(v \mid C_i)$ as prior knowledge to re-weight the empirical joint probability $P(v, C_i)$. The re-weighted joint probability implies a fusion of prior belief and empirical evidence:

$$P_u(v, C_i) = \frac{Pr(v \mid C_i) \cdot P(v, C_i)}{Z} \tag{6}$$

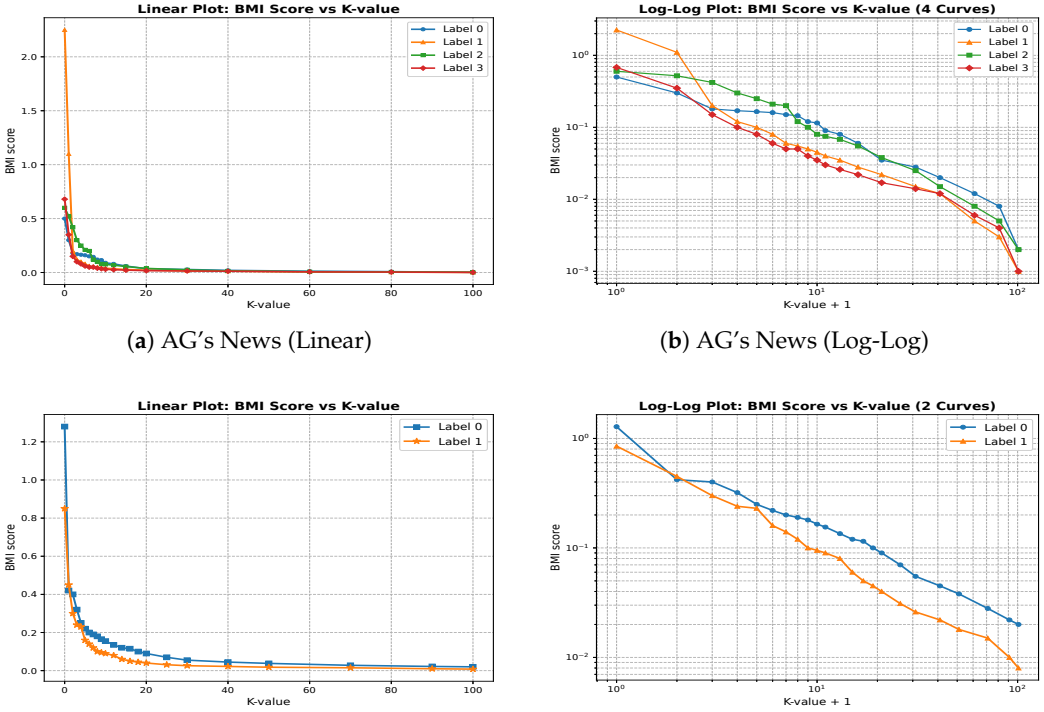
where Z is the global normalization factor (partition function) ensuring that P_u constitutes a valid joint probability distribution over the entire candidate space. It is calculated by summing over all classes and all candidate words:

$$Z = \sum_{v' \in \tilde{V}_C} \sum_{k=1}^{|\mathcal{C}|} Pr(v' \mid C_k) \cdot P(v', C_k) \tag{7}$$

This refinement mechanism effectively mitigates the inherent prediction bias of PLMs while compensating for the data sparsity in few-shot settings. By integrating prior knowledge with empirical frequency, it allows us to compute a refined mutual information score that robustly quantifies the word's relevance to the class.

Analysis of the BMI score distribution is presented in Figure 3 to validate the verbalizer construction process. Specifically, we examine the distribution characteristics on both the

AG's News and IMDB datasets using linear and log–log scales. As observed in the linear plots, the scores reveal a steep initial decline, indicating that the most discriminative semantic information is concentrated in the top-ranked words. This observation serves as the theoretical basis for our choice of setting $k = 20$, as the BMI scores flatten out rapidly after this threshold, representing diminishing returns. This choice is further corroborated by our empirical sensitivity analysis in Section 4.8 , which confirms that $k = 20$ yields optimal classification performance. Furthermore, to verify the statistical nature of the distribution, we visualize the data using log–log plots as suggested by Ref. [50]. The resulting curves exhibit an approximate linear trend, particularly evident in the IMDB dataset. This indicates that the refined BMI scores follow a Zipfian (or Pareto-like) distribution, which is consistent with the statistical laws of natural language usage. This finding confirms that our method effectively prioritizes high-relevance label words while filtering out the long tail of noise.



(**a**) AG's News (Linear) (**b**) AG's News (Log-Log)

(**c**) IMDB (Linear) (**d**) IMDB (Log-Log)

Figure 3. Analysis of BMI score distributions. (**a**,**c**) display the sorted BMI scores in linear scale for AG's News and IMDB, showing the steep decline. (**b**,**d**) present the log-log plots. The approximate linear trend in (**b**,**d**) indicates a Zipfian (power-law-like) distribution, validating the feature selection strategy.

3.2. Prompt-Template-Guided Multi-Granularity Label Data Augmentation

To generate high-quality semantic-coherent samples for each class, we adopt a Prompt-Template-Guided Multi-Granularity Label Data Augmentation method and design two synergistic generators to produce sentences with a predefined structural format that is as shown in Figure 4. This process involves two main steps: First, the Label-Enhanced Generator (LEG) generates a set of label-enhanced templates that incorporate class-related label words, providing a foundation for subsequent data augmentation tasks. Next, the Prompt-Guided Generator (PGG) integrates the label-enhanced templates produced by LEG with the original samples and adjusts the positions of the original samples according to the rules defined by the prompt template, ensuring sentence coherence is maintained.

Label-Enhanced Generator: Let the original dataset be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i represents the i-th input text, and $c_i \in \mathcal{C}$ denotes its corresponding class label, with $\mathcal{C}$ being the set of all classes. For each class $c \in \mathcal{C}$, a verbalizer set $\mathcal{V}_c = \{w_{c,j}\}_{j=1}^{K}$ is constructed

using a PLM and an improved BAVC strategy. Based on the BMI scores $MI_{\mathrm{u}}(v_{c,j}, c)$, the top m (with m being a hyperparameter) label words are selected to form the subset $\mathcal{L}_c$.

A predefined template set $\mathcal{T} = \{\tau_k\}_{k=1}^{M}$ consists of a group of prompt phrases with placeholders $[MASK]$. By replacing the $[MASK]$ position with the top-m label words $v_{c,j} \in \mathcal{L}_c$, augmented texts are generated. For a given input sample x_i, the label-augmented text is expressed as follows:

$$x_i^{\mathrm{aug}} = \tau_k(x_i, v_{c,j}) \tag{8}$$

where $v_{c,j}$ represents a label word from $\mathcal{L}_c$, and τ_k is a selected template. By iterating over the top-m label words in $\mathcal{L}_c$ and the template set $\mathcal{T}$, multiple augmented texts are generated for each input x_i, enriching the training data and enhancing the model's class awareness.

Prompt-Guided Generator: Given the label-augmented text $\mathcal{X}_i^{\mathrm{aug}}$ generated by the LEG, the PGG synthesizes these texts into semantically coherent samples, referred to as $\mathcal{X}_i^{\mathrm{coh}}$. To achieve this, PGG employs a set of predefined prompt templates $\mathcal{T}' = \{\tau'_k\}_{k=1}^{M'}$, each designed to incorporate the original input x_i and one or more label-augmented texts $x_i^{\mathrm{aug},(k,j)}$ into a single coherent sequence.

For a given input x_i, PGG constructs a semantic-coherent sample by concatenating x_i with $x_i^{\mathrm{aug},(k,j)}$ at specific positions within the selected template τ'_k:

$$x_i^{\mathrm{coh}} = \tau'_k(x_i, x_i^{\mathrm{aug},(k,j)}), \tag{9}$$

where $\tau'_k(x_i, x_i^{\mathrm{aug},(k,j)})$ represents the mapping function of template τ'_k that ensures the semantic coherence of the combined sequence.

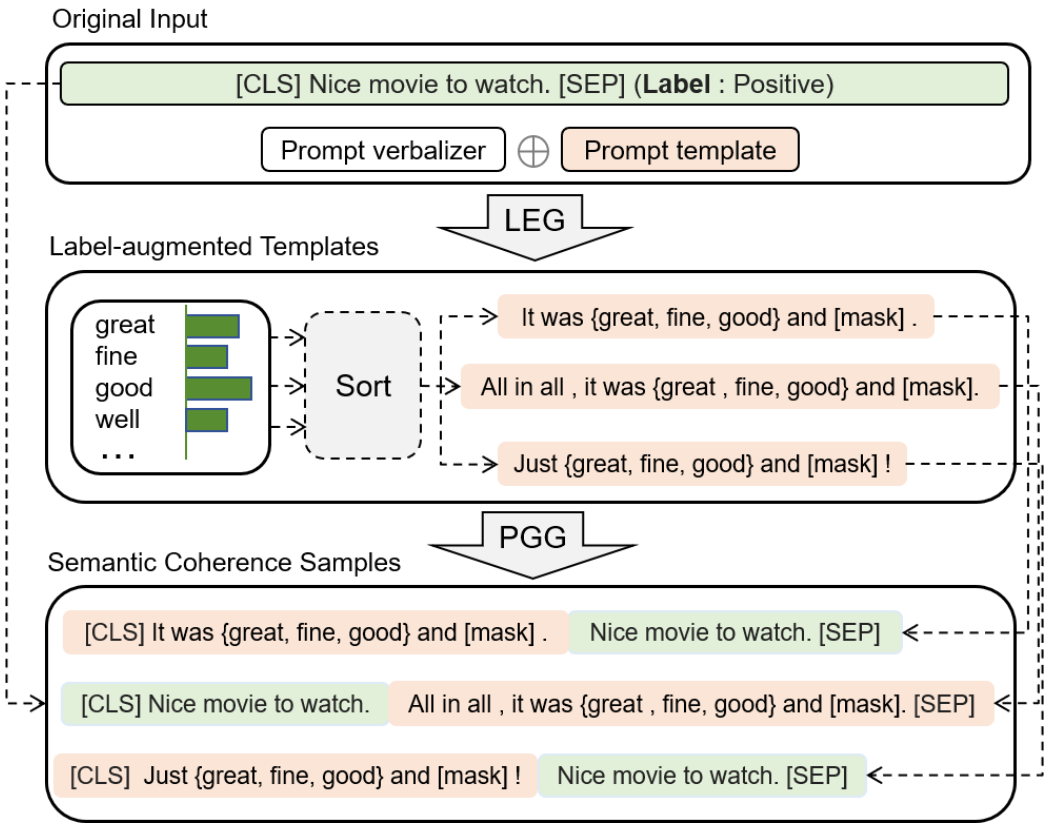


Figure 4. The process of data augmentation in AMLDA.

By iterating over all templates $\mathcal{T}'$ and label-augmented texts $\mathcal{X}_i^{\text{aug}}$, PGG generates a set of semantic-coherent samples:

$$\mathcal{X}_i^{\text{coh}} = \{x_i^{\text{coh},(k',j)} \mid k' = 1, \ldots, M',\ j = 1, \ldots, 4\}. \tag{10}$$

This process not only preserves the original semantic context of x_i but also ensures the structural and linguistic alignment of the augmented components, improving the diversity and coherence of the training data.

3.3. Low-Entropy Selector

It is essential to filter the semantic-coherent samples generated by the two generators to reduce noise and ensure diversity, removing overly ambiguous samples to prevent biases during model training and maintain a balanced data distribution.

For a set of semantic-coherent samples $\mathcal{X}_i^{\text{coh}}$, where each sample is generated from an original sample using two generators (e.g., LEG and PGG), our method aims to filter out noisy samples or those whose predicted categories are inconsistent with the original category, ultimately retaining the most informative augmented samples. Specifically, for each sample $x_i \in \mathcal{X}_i^{\text{coh}}$, the model predicts the logits distribution h_i, from which the class probabilities $p(x_i)$ are computed using the softmax function:

$$p(x_i) = \text{softmax}(h_i W_s + b_s) \tag{11}$$

Here, W_s and b_s are the learnable parameters of the classifier. Based on these probabilities, the information entropy $H(x_i)$ of each sample is calculated to measure the uncertainty of the prediction:

$$H(x_i) = -\sum_{c=1}^{C} p(x_{i,c}) \log p(x_{i,c}), \tag{12}$$

where C denotes the total number of classes, and $p(x_{i,c})$ is the predicted probability of sample x_i belonging to class c. Lower entropy indicates higher confidence in the model's prediction. Accordingly, we rank the augmented candidates in ascending order of their label entropy and select the L samples with the lowest entropy for each class independently. This class-wise selection strategy ensures that the augmented dataset maintains strict class balance, preventing the model from biasing towards "easier" classes. Regarding the determination of the hyperparameter L, we analyze its impact on model performance in the Parameter Sensitivity section (see Section 4.8) , balancing the trade-off between sample quantity and semantic quality.

3.4. Difficulty-Aware Adversarial Training

Our objective is to dynamically adjust the perturbation intensity during training based on the predicted difficulty of each sample, generating adversarial samples with varying strengths to refine the model's text recognition capability. Unlike image data, text inputs consist of discrete tokens that are non-differentiable, making it impossible to directly compute gradients with respect to the input indices. To address this, following the standard practice in textual adversarial training (Ref. [51]), we apply perturbations to the continuous embedding space rather than the discrete tokens.

Let $\boldsymbol{x}_i \in \mathbb{R}^{L \times d}$ denote the input embedding vectors for the i-th sample (where L is sequence length and d is embedding dimension) and $H(\boldsymbol{x}_i)$ be the sample-wise entropy. During each epoch, we compute the global mean entropy $\bar{H} = \frac{1}{m} \sum_{i=1}^{m} H(\boldsymbol{x}_i)$. The adaptive perturbation strength ϵ_i for each sample is determined by its difficulty (entropy) as follows:

$$\epsilon_i = \epsilon_{\text{base}} + \frac{\bar{H}}{|H(\boldsymbol{x}_i) - \bar{H}| + \delta} \cdot \text{sign}(\bar{H} - H(\boldsymbol{x}_i)) \tag{13}$$

where ϵ_{base} is a scaling factor, and δ is a small constant for numerical stability. Note that Equation (13) ensures that samples with higher entropy (difficult samples, $H(\boldsymbol{x}_i) > \bar{H}$) receive smaller perturbations to preserve their semantic structure, while lower entropy samples (easy samples) are assigned larger perturbations to enhance robustness.

The adversarial perturbation $r_{\text{adv},i}$ is generated via gradient ascent on the embedding vectors to maximize the loss:

$$r_{\text{adv},i} = \epsilon_i \cdot \frac{\nabla_{\boldsymbol{x}} L(f(\theta, \boldsymbol{x}_i), y_i)}{\|\nabla_{\boldsymbol{x}} L(f(\theta, \boldsymbol{x}_i), y_i)\|_2} \tag{14}$$

By injecting this perturbation into the embedding layer ($\boldsymbol{x}_{\text{adv}} = \boldsymbol{x} + r_{\text{adv}}$), we compel the model to learn generalized representations that are robust to distributional shifts in the feature space. The composite loss combines standard and adversarial components:

$$\begin{aligned} L(\theta) = & \underbrace{-\frac{1}{N}\sum_{n=1}^{N} \log P(y_n|\boldsymbol{x}_n, \theta)}_{L_{\text{NLL}}} \\ & + \lambda \underbrace{\left[-\frac{1}{N}\sum_{n=1}^{N} \log P(y_n|\boldsymbol{x}_n + r_{\text{adv},n}, \theta)\right]}_{L_{\text{adv}}} \end{aligned} \tag{15}$$

where λ is a hyperparameter balancing the two objectives. Finally, as illustrated in the "Optimization Step" of Figure 1, the model parameters θ are updated by minimizing this composite loss function via backpropagation. This completes the end-to-end training loop of the proposed AMLDA framework.

4. Experiments

In this section, we conduct experiments on four benchmark datasets to evaluate the effectiveness of our AMLDA approach. First, we describe the datasets and templates used. Subsequently, we investigate the compared methodologies and the detailed experimental setup used in our study. Finally, we analyze our findings and examine the influence of hyper-parameters on the outcomes.

4.1. Datasets and Templates

In our experiments, we utilize four public text classification datasets to assess our method: AG's News [52], DBPedia [53], IMDB [54], and Amazon [55]. The statistics of the training and testing sets listed in Table 1 strictly adhere to the standard splits defined in their respective original papers [52–55]. For example, AG's News utilizes the standard split of 120,000 training and 7600 testing samples, while IMDB employs a balanced split of 25,000 samples each for training and testing. It is important to note that for the few-shot learning settings, we do not use the full training set; instead, we sample $N \times K$ instances from the original training set (where K is the shot number) to construct the few-shot support set, while the evaluation is conducted on the complete standard test set to ensure robust performance metrics. Additionally, considering that [47] has demonstrated that manually crafted templates are competitive with or superior to automatically generated templates and that they are easier to construct, we utilize manual templates in our experiments. At the same time, to mitigate the influence of varying templates on the experiments,

we follow [4]'s approach to design templates and customize the configuration for each experiment to fit the dataset.

Table 1. The statistics strictly follow the standard splits of the original benchmark datasets. In our few-shot experiments, we sample K training instances per class from the original training set while evaluating on the full test set to ensure consistent comparison.

Dataset	Type	Classes	Train Set	Test Set
AG's News	Topic classification	4	120,000	7600
DBPedia	Topic classification	14	560,000	70,000
Amazon	Sentiment classification	2	20,000	10,000
IMDB	Sentiment classification	2	25,000	25,000

4.2. Baselines

We detail the baseline models used in our experiments to benchmark the performance of our proposed method. The chosen baselines include fine-tuning, prompt tuning, EDA, KPT, and SKP. Below, we provide a comprehensive description of each baseline and its implementation details.

- **Fine-tuning**: As a traditional transfer learning baseline, it adapts pre-trained language models to downstream tasks by directly adjusting their weights [56]. We include it to demonstrate the limitations of standard fine-tuning methods in few-shot settings, particularly their sensitivity to data sparsity and overfitting.
- **Prompt tuning**: As a representative prompt-based method, it reformulates classification tasks as masked language modeling problems using manually crafted templates [57]. We incorporate it to validate the advantages of prompt-based approaches over fine-tuning in few-shot scenarios while highlighting the limitations of manual prompts.
- **EDA**: As a classical data augmentation technique, it generates synthetic samples through lexical-level transformations [15]. We select it to compare traditional augmentation strategies with our proposed semantic-consistent augmentation approach, particularly emphasizing EDA's shortcomings in preserving label semantics and ensuring sample quality.
- **KPT**: As a knowledge-enhanced prompt-tuning method, it enriches prompt content by incorporating external knowledge bases [44]. We use it as a baseline to verify the costs of external knowledge-based methods in few-shot settings and to demonstrate the advantages of our proposed external-knowledge-free adaptive approach.
- **SKP**: As a state-of-the-art soft prompt-tuning method, it constructs verbalizers using learnable soft tokens [43]. We choose this advanced baseline to demonstrate the efficacy of our method in verbalizer construction and its theoretical advantages over soft prompts in label mapping.

4.3. Experimental Setting

In our primary experiments, we used the RoBERTa-large model as the core PLM, leveraging the Hugging Face Transformer Library's implementation. Considering the balanced nature of the datasets, we adopted the Micro-F1 score as our evaluation metric to ensure a thorough assessment of classification performance across all trials.

To gauge the effectiveness of each prompt-based method, we utilize four distinct templates and five unique random seeds, with the reported outcomes representing the mean of 20 iterations. This multi-run strategy mitigates the impact of variability on our results. In the case of N-shot experiments, we extract N samples per class from the original training corpus to assemble the few-shot training set and similarly draw N samples per

class to establish the validation set, with N values being 1, 5, 10, and 20 across different few-shot scenarios. On the hyper-parameter front, the model undergoes five epochs of training, and the checkpoint demonstrating optimal validation performance is chosen for the final evaluation. The maximum input lengths are set at 128 for AG's News and DBPedia and 512 for the Amazon and IMDB datasets. The optimizer's learning rate is maintained at 3×10^{-5}.

4.4. Experimental Results

In this subsection, the comparison results between the AMLDA model and baseline models across all four datasets are presented in Table 2. The results demonstrate that AMLDA consistently outperforms the other methods in terms of both F1 scores and robustness. Based on these results, several important conclusions can be drawn:

(1) In comparison to the other methods, across different shot settings (1/5/10/20-shot), our method achieved the highest Micro-F1 scores on almost all datasets. For example, it achieved 87.1 Micro-F1 on the AG's News dataset (5-shot), an impressive 98.1 Micro-F1 on the DBPedia dataset (5-shot), and 93.7 Micro-F1 on the Amazon dataset (5-shot). These results demonstrate the outstanding performance of AMLDA in few-shot text classification tasks, particularly in scenarios with extremely limited training samples.

(2) Compared to the other data augmentation models, such as PT+EDA, EDA indiscriminately deletes or inserts tokens, introducing more noise into the prediction model and affecting its decision-making process. In contrast, AMLDA consistently outperformed EDA across all baselines, highlighting its effectiveness in improving data quality and reducing noise.

(3) In most cases, our approach outperforms other state-of-the-art prompt-tuning methods like KPT and SKP. The results demonstrate that the AMLDA model improves the accuracy of prompt-based FSTC methods. However, it performs slightly worse than SKP on the IMDB dataset. Notably, our method not only achieves better average performance but also exhibits superior stability. For instance, on the Amazon dataset with 1-shot learning, our method achieves a standard deviation of 1.2 compared to SKP's 2.1, which is significantly lower. This indicates that, with very few samples available, our method enhances the model's robustness to variations in the input data.

We also observe notable performance variations across the four benchmark datasets. While AMLDA consistently outperforms baselines, the absolute F1 scores vary significantly (e.g., DBpedia vs. AG's News). This cross-dataset variance is primarily driven by three factors:

(1) Semantic Distinctiveness: Despite having the highest number of classes (14), DBpedia yields the highest performance. This is because its categories are semantically well separated (e.g., Artist vs. OfficeHolder), which aligns effectively with our BMI-driven verbalizer. Conversely, AG's News involves topics with higher lexical overlap (e.g., Business and Politics), making the few-shot boundary inherently more difficult to define.

(2) Domain-Knowledge Alignment: The performance gap also stems from the varying degrees of alignment between the PLM's pre-trained knowledge and the target domain. RoBERTa exhibits stronger zero-shot priors for the global facts found in DBpedia than for the specific linguistic patterns in localized news or sentiment datasets.

(3) Label Granularity: In binary sentiment tasks like Amazon and IMDB, the model focuses on polar semantics. The higher baseline variance compared to DBpedia suggests that while the task is simpler (fewer classes), the model's sensitivity to the

specific nuances of "positive/negative" phrasing in few-shot samples is higher, a challenge that our DAAT module specifically aims to mitigate.

Table 2. Results of 1/5/10/20-shot text classification based on RoBERTa-large. We report their mean ± standard deviation of 20 runs (four templates and five random seeds), bold values indicate the best performance for each dataset under various shot settings.

Shot	Method	AG's News	DBPedia	Amazon	IMDB
1	Fine-tuning	19.8 ± 10.4	8.6 ± 4.5	49.9 ± 0.2	50.0 ± 0.0
	Prompt-tuning	80.0 ± 6.0	92.2 ± 2.5	91.9 ± 2.7	91.2 ± 3.7
	PT + EDA	79.3 ± 7.6	90.3 ± 1.9	88.6 ± 3.5	89.5 ± 2.4
	KPT	83.7 ± 3.5	93.7 ± 1.8	**93.2 ± 1.3**	92.2 ± 3.0
	SKP	84.3 ± 2.9	/	92.6 ± 2.1	**92.9 ± 1.7**
	ours	**86.1 ± 2.5**	**94.5 ± 1.7**	92.4 ± 1.2	91.2 ± 1.6
5	Fine-tuning	37.9 ± 10.0	95.8 ± 1.3	52.1 ± 1.3	51.4 ± 1.4
	Prompt-tuning	82.7 ± 2.7	97.0 ± 0.6	92.2 ± 3.3	91.9 ± 3.1
	PT + EDA	80.6 ± 4.1	92.7 ± 2.9	89.5 ± 3.1	89.8 ± 3.6
	KPT	85.0 ± 1.2	97.1 ± 0.4	93.4 ± 1.9	92.7 ± 1.5
	SKP	84.4 ± 1.8	/	93.3 ± 1.7	**93.1 ± 1.6**
	ours	**87.1 ± 1.4**	**98.1 ± 0.6**	**93.7 ± 1.3**	92.8 ± 1.4
10	Fine-tuning	75.9 ± 8.4	93.8 ± 2.2	83.0 ± 7.0	76.2 ± 8.7
	Prompt-tuning	84.9 ± 2.4	97.6 ± 0.4	93.9 ± 1.3	93.0 ± 1.7
	PT + EDA	81.4 ± 3.6	93.5 ± 2.4	92.7 ± 1.8	91.1 ± 3.1
	KPT	86.3 ± 1.6	98.0 ± 0.2	93.8 ± 1.2	92.9 ± 1.8
	SKP	86.6 ± 1.2	/	94.1 ± 1.2	**94.1 ± 1.5**
	ours	**89.1 ± 1.1**	**98.3 ± 0.2**	**94.6 ± 0.8**	93.6 ± 1.3
20	Fine-tuning	85.4 ± 1.8	97.9 ± 0.2	71.4 ± 4.3	78.5 ± 10.1
	Prompt-tuning	86.5 ± 1.6	97.7 ± 0.3	93.5 ± 1.0	93.0 ± 1.1
	PT + EDA	84.1 ± 1.4	93.6 ± 2.7	92.9 ± 1.9	92.8 ± 2.0
	KPT	87.2 ± 0.8	98.1 ± 0.3	93.7 ± 1.6	93.1 ± 1.1
	SKP	88.0 ± 1.1	/	94.8 ± 2.2	95.0 ± 1.8
	ours	**89.8 ± 0.6**	**98.8 ± 0.4**	**95.7 ± 1.3**	**95.1 ± 1.4**

Then we conducted Friedman tests on four datasets as significance tests. In these tests, different methods were used as grouping variables and compared against all baseline methods on each dataset. The test statistics on the respective datasets were as follows: AG's News (0.0018), DBPedia (0.0056), Amazon (0.0039), and IMDB (0.0028). By comparing the test statistics with the critical value of the theoretical distribution at a significance level of $\alpha = 0.01$, we conclude that the AMLDA method demonstrates a significant advantage over all competing methods.

In addition, to further evaluate the robustness of the model in handling few-shot scenarios, we tested its classification stability under different shot settings. Specifically, we conducted experiments on the recall metric. Recall is defined as the model's ability to accurately identify positive samples, calculated as follows:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{16}$$

This metric reflects the model's capacity to retrieve positive samples effectively. As shown in Figure 5, the AMLDA method achieves consistently high and stable classification performance under different shot settings, further illustrating the robustness of our method in few-shot settings.

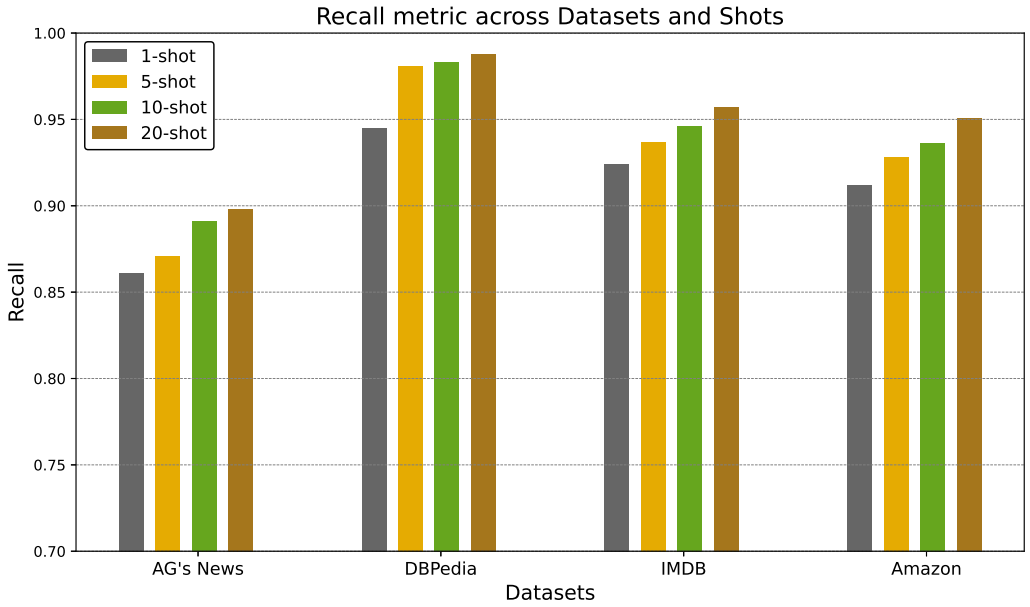


Figure 5. The results of the recall metric.

4.5. Computational Complexity Analysis

We further evaluate the computational overhead of our proposed AMLDA framework. Since Difficulty-Aware Adversarial Training (DAAT) requires an additional gradient computation step to generate perturbations in the embedding space, the training time per epoch is theoretically expected to be higher than standard prompt-tuning. Empirically, the average training time per epoch for standard prompt-tuning is defined as a base unit T_{base}. Our experiments on the AG's News dataset (using an NVIDIA RTX 3090 GPU) show that AMLDA requires approximately $1.8 \times T_{base}$ seconds per epoch. However, it is crucial to note that this overhead is strictly limited to the training phase. The inference speed remains identical to the base model because the auxiliary modules—specifically, the augmented sample generators and adversarial perturbation mechanisms—are deactivated during prediction. Given the significant performance gains (+F1 points) and improved robustness in few-shot scenarios, we consider this training cost to be a justifiable trade-off.

4.6. Ablations

To systematically investigate the effectiveness of key components within the AMLDA framework, including BMI-Driven Adaptive Verbalizer Construction (BAVC), Prompt-Template-Guided Multi-Granularity Label Data Augmentation (PTMD-LDA), and Difficulty-Aware Adversarial Training (DAAT), we conducted comprehensive comparisons against each component across four datasets. The primary objective was to assess how modifying or removing these components affects the model performance, thereby underscoring the contribution of each module to the overall efficacy of the framework. The results are delineated in Table 3. In this experiment, RoBERTa-Large is used as the prediction model, and w/o denotes the removal of the corresponding module.

Table 3. Ablation study of AMLDA, bold values indicate the best performance in each column.

Method	AGNews	DBPedia	Amazon	IMDB
AMLDA	**89.8 ± 0.6**	**98.8 ± 0.4**	**95.7 ± 1.3**	**95.1 ± 1.4**
AMLDA w/o all	86.5 ± 1.6	97.7 ± 0.3	93.5 ± 1.0	93.0 ± 1.1
AMLDA w/o DAAT	89.1 ± 1.3	98.2 ± 0.6	94.5 ± 1.4	94.3 ± 2.1
AMLDA w/o PTMD-LDA	87.3 ± 0.9	97.9 ± 0.5	94.3 ± 0.8	93.2 ± 1.3
AMLDA w/o Classical MI	88.9 ± 1.1	98.3 ± 0.5	94.8 ± 1.2	93.8 ± 1.5
AMLDA w/o BAVC	88.6 ± 0.8	98.0 ± 0.6	94.5 ± 1.3	93.5 ± 1.6

First, removing all three modules and reverting to standard RoBERTa-Large for prompt tuning leads to a noticeable degradation in model performance. Specifically, the F1 scores on the AG's News, DBPedia, Amazon, and IMDB datasets decrease by 3.3%, 1.1%, 2.2%, and 2.1%, respectively. This clearly demonstrates that our AMLDA framework enhances the model's generalization ability and learning efficiency.

Next, regarding the *w/o DAAT* setting (i.e., removing adversarial training), a slight drop in F1 scores is observed, but the standard deviation increases significantly. This suggests that the adversarial samples introduced by DAAT simulate small perturbations and boundary examples that the model may encounter in real-world scenarios, thereby enhancing robustness to input variations. Additionally, we infer that DAAT compensates for sample limitations; while PTMD-LDA generates semantically consistent samples, adversarial training further enriches diversity, forcing the model to learn more generalized representations.

Similarly, for the *w/o PTMD-LDA* variant (without multi-dimensional label data augmentation), the results show a substantial decline in F1 scores across all four datasets, with a particularly significant drop of 1.9% on IMDB. This indicates that the semantically consistent samples generated by PTMD-LDA are of high quality and crucial for helping the model capture class semantics.

Then, the *w/o BAVC* setting implies that the final verbalizer is composed of a manually selected set of label words. As shown in Table 3, the removal of BAVC leads to a significant decrease in F1 scores across all four datasets. Although the drop is less pronounced than removing PTMD-LDA, it confirms that BAVC constructs a higher-quality verbalizer than manual selection. Manual selection is susceptible to subjective bias and may result in insufficient label coverage. In contrast, the automated BAVC strategy, incorporating mutual information optimization and dynamic thresholding, ensures semantic consistency and better expressiveness.

More importantly, to demonstrate the methodological advantage of our Bayesian approach, we implemented a variant named AMLDA *w/ClassicalMI*. In this setting, we constructed the verbalizer using classical Mutual Information based solely on the empirical word frequency distribution within the generated candidate set, without the PLM-based Bayesian re-weighting. The results indicate that *w/ClassicalMI* consistently underperforms the proposed AMLDA (with BMI). The performance gap is particularly noticeable on the AG's News and IMDB datasets. This degradation occurs because classical MI relies heavily on statistical co-occurrences, which are prone to high variance and noise in few-shot scenarios (e.g., overfitting to incidental words). By contrast, our BMI integrates the PLM's zero-shot predictions as a prior, effectively regularizing the selection process and filtering out statistically frequent but semantically irrelevant noise.

Collectively, the ablation study confirms PTMD-LDA as the performance-critical module: its removal causes the most severe degradation (1.9% on IMDB), surpassing BAVC's impact. As the central processing hub, PTMD-LDA transforms BAVC's semantic mappings into actionable augmented samples while providing the foundation for DAAT's diversity enhancement—making this dual-function component indispensable for AMLDA's robustness.

4.7. AMLDA and Conventional Data Augmentation

Conventional data augmentation techniques provide only minimal enhancements or even have negative effects on to prompt-based few-shot learning [19]. We hypothesize the reason for this is that such data augmentation approaches are primarily focused on modifying instances while neglecting the importance of label semantics, which may lead to changes in the original categories and introduce additional noise. Here, we compare

the performance of conventional data augmentation methods with our AMLDA in a few-shot setting and explore the potential of combining AMLDA with conventional data augmentation techniques.

The experimental setup is consistent with the main experiments, using 20 samples per class. AMLDA retains two semantic-coherent samples with minimal entropy for each original sample, while the most representative conventional data augmentation methods—Synonym Replacement and Random Insertion—each generate two augmented samples for every original sample. The results are shown in Table 4.

Table 4. AMLDA vs. Conventional DA.

Method	AGNews	DBPedia	Amazon	IMDB
PT	86.5 ± 1.6	97.7 ± 0.3	93.5 ± 1.0	93.0 ± 1.1
PT + ConvDA	84.1 ± 1.4	93.6 ± 2.7	92.9 ± 1.9	92.8 ± 2.0
AMLDA	89.8 ± 0.6	98.8 ± 0.4	95.7 ± 1.3	95.1 ± 1.4
AMLDA + ConvDA	**89.9 ± 0.5**	**98.8 ± 0.2**	**96.2 ± 1.1**	**95.4 ± 0.8**
AMLDA + ConvDA (*w*/*o* LES)	88.6 ± 1.8	97.9 ± 0.6	94.5 ± 1.3	93.2 ± 2.1

4.7.1. Comparison with Conventional Data Augmentation

By analyzing the results in Table 4, we observe that the combination of conventional DA with prompt tuning reduces the performance of the FSTC model, but our AMLDA, which is also based on prompt tuning, shows significantly better performance compared to conventional DA. Specifically, on the IMDB dataset, AMLDA increases the F1 score from 93.0% to 95.1%, an improvement of approximately 2.1%, while conventional DA decreases the F1 score from 93.0% to 92.8%. It is plausible that while conventional DA attempts to address the problem of data scarcity by modifying instances, it primarily focuses on leveraging the semantic information of the training instances themselves. In contrast, AMLDA enhances the data under the guidance of label words, effectively strengthening the model's understanding of other label-related words while preserving the original sentence information.

4.7.2. Combination with Conventional Data Augmentation

We observed that the combination of AMLDA and Conventional DA outperformed other combinations. While traditional data augmentation methods such as Synonym Replacement and Random Insertion increase sample diversity, they may introduce low-quality or ambiguous samples, which may not effectively improve the model's learning performance. In contrast, our method leverages LES to remove high-ambiguity samples based on an entropy-based selection mechanism, thereby reducing the interference of noisy samples and improving the quality of the training data.

To further verify the role of LES, we performed a comparison experiment by removing LES. The results showed a decrease in F1 scores across all four datasets, with the F1 scores of the datasets being lower than those achieved by AMLDA alone. This result strongly demonstrates the crucial role of LES in filtering high-quality samples and reducing noise, and it also indicates that Conventional DA serves as a complementary method to AMLDA.

4.8. Parameter Sensitivity

In this subsection, we explore the influencing factors of AMLDA from three perspectives: the size of the data augmentation, the number and order of label words, and the verbalizer size k. To isolate the impact of each factor, we adopt a univariate sensitivity analysis approach: when varying the target parameter, the remaining hyperparameters are fully fixed at their empirically determined optimal values (i.e., $L = 2$, $m = 3$, and $k = 20$).

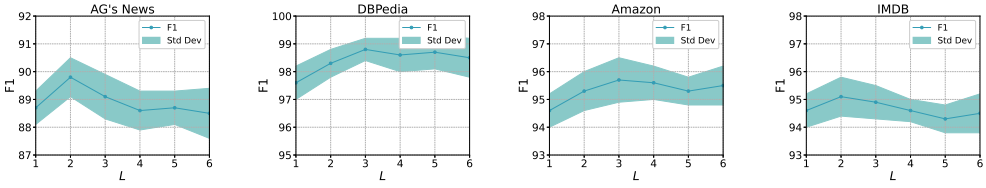
The size of data augmentation L**.** Across all datasets, we investigate how the quantity L of augmented samples generated for each example influences the classification performance of AMLDA. For each category, there are 20 examples in the training set, and we generate between 1 and 6 augmented samples for each example. The results from five trials over ten epochs are presented in Figure 6. The best performance consistently occurs when $L = 2$ or 3.

However, an excessive quantity of augmented data may lead to worse final performance. We analyze the reasons from two perspectives. Firstly, L is a critical hyperparameter in LES that determines the number of augmented samples retained. Each augmented sample is designed to preserve semantic information consistent with the original sample. When the number of augmented samples is too large, information redundancy arises, meaning that newly generated samples fail to provide additional semantic features useful for classification. This redundancy may cause the model to overfit the training data. Secondly, in LES, the entropy of augmented samples tends to increase as their order progresses, leading to higher noise levels in later samples. When the number of augmented samples exceeds a certain threshold, the model may tend to learn noise rather than meaningful features. This shift in learning focus degrades the model's performance on real test data, resulting in a decline in final performance.

The number and order of label words m**.** We investigated the impact of the number m and the order of label words on model performance in PTMD-LDA when generating augmented samples. Specifically, we evaluated the performance of AMLDA under two conditions: label words in random order and label words sorted by BMI scores, with $m \in \{1, 2, 3, 4, 5, 6\}$. The results are shown in Figure 7.

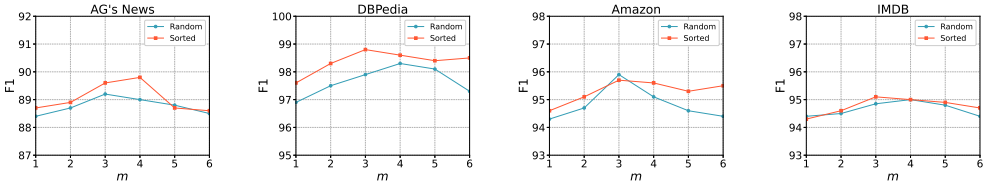
First, under the condition where label words are sorted by BMI scores, increasing the number of label words led to consistent improvements in F1 scores across the four different datasets. However, when the number of label words exceeded 3 or 4, the F1 scores started to decrease. This indicates that label words with lower BMI scores tend to have weaker associations within the same class, and the inclusion of lower-ranked label words is more likely to introduce noise. Second, under the condition of randomly ordered label words, the model performance was slightly worse than that of the sorted approach. We argue that sorting label words by BMI scores encapsulates the relationships between labels, which helps the model learn the classification task more effectively.

The verbalizer size k. To empirically verify the optimal number of label words in the verbalizer, we conducted a sensitivity analysis on the parameter k, ranging from 5 to 50. The results, as shown in Figure 8, indicate a clear trend across all four datasets. Initially, as k increases from 5 to 20, the model's performance (Micro-F1 score) improves significantly. This suggests that expanding the verbalizer incorporates more semantic information relevant to the classes, thereby enhancing prediction accuracy. However, performance peaks consistently around $k = 20$. Beyond this threshold (e.g., $k = 30, 40, 50$), the F1 score either plateaus or noticeably declines, while the standard deviation tends to increase (as seen in the AG's News and Amazon datasets). This performance drop aligns with our observation of the power-law distribution in Figure 3: the top-ranked words contain the most discriminative information. Including lower-ranked words (the "long tail") introduces semantic noise rather than useful signals, confusing the model. Therefore, $k = 20$ is selected as the optimal trade-off between semantic coverage and noise reduction.



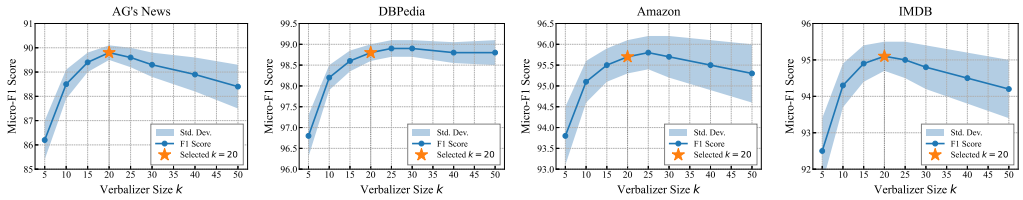
(**a**) (**b**) (**c**) (**d**)

Figure 6. The impact of the size of augmentation parameter L on F1 score and standard deviation across four text classification datasets: (**a**) AG's News; (**b**) DBPedia; (**c**) Amazon; (**d**) IMDB.



(**a**) (**b**) (**c**) (**d**)

Figure 7. Impact of the number and order of label words m on model performance across four datasets: (**a**) AG's News; (**b**) DBPedia; (**c**) Amazon; (**d**) IMDB.



(**a**) (**b**) (**c**) (**d**)

Figure 8. Parameter Sensitivity analysis of verbalizer size k across four datasets: (**a**) AG's News; (**b**) DBPedia; (**c**) Amazon; (**d**) IMDB. The red star denotes the selected value $k = 20$.

4.9. Error Analysis and Case Studies

To gain a deeper understanding of the model's prediction behavior and error distribution, we first visualize the confusion matrices for AG's News, IMDB, and Amazon datasets under the 20-shot setting, as shown in Figure 9. As illustrated in Figure 9, the confusion matrices reveal that our AMLDA framework maintains a balanced performance on binary sentiment tasks (IMDB and Amazon), effectively mitigating the class bias often found in few-shot scenarios. However, for the multi-class AG's News dataset, while the "Sports" category exhibits distinct boundaries, a noticeable portion of errors exists between the "Business" and "Politics" (or "Technology") categories. Specifically, the model tends to misclassify ambiguously phrased Business news as Politics due to overlapping contexts (e.g., government regulations on industries).

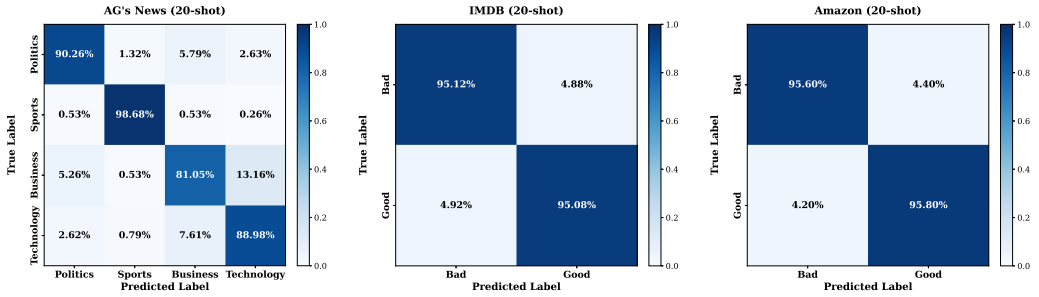
To better understand how AMLDA processes such ambiguous inputs and reduces noise during augmentation, we present a step-by-step qualitative analysis of the intermediate results in Table 5. By tracking a representative sample from the AG's News dataset, we observe the specific contributions of each module. In Phase 1 (BAVC), distinct from generic manual labels, the adaptive verbalizer extracts granular terms like "match" and "athlete," meaningfully extending the semantic scope of the "Sports" category. Subsequently, in Phase 3 (PGG), the generator synthesizes diverse candidates by varying the structural position of the templates (e.g., placing the prompt before or after the input), which enhances the model's adaptability to different linguistic patterns. Most crucially, Phase 4 (LES) acts as a semantic filter. As clearly shown in Table 5, while Candidates 1 and 2 are retained due to high predictive confidence (low entropy), Candidate 3—which combined the template

with the input to form an ambiguous context—exhibits significantly higher entropy (0.92). Consistent with our retention setting ($L = 2$), this noisy sample is explicitly discarded. This granular visualization demonstrates that AMLDA not only enriches data diversity but also possesses an internal mechanism to rigorously prevent noise propagation.

While the confusion matrix reveals these inherent semantic challenges, our method significantly ameliorates them compared to traditional approaches. To investigate the root causes of these specific confusions observed in the matrix, we selected representative samples that are particularly prone to confusion for detailed analysis. Table 6 presents four specific instances where the prompt-tuning prediction fails, but the AMLDA prediction is accurate. "Wrong case" refers to the original input sample; "Label" denotes the true label; and "Prediction" indicates the result from the prompt-tuning prediction. Consistent with the confusion patterns observed in Figure 9, the first example misclassifies a Business sample as Politics. This error stems from prompt tuning's heavy reliance on PLMs prior knowledge under few-shot learning conditions. Such knowledge carries domain-generalization bias, and prompt tuning's use of generic label words for verbalizer-based single-category mapping leads to excessive focus on limited keywords. This observation aligns with findings in KPT++ [58], where the model exhibits biased scoring towards Business and Politics label words while neglecting other category indicators (e.g., "international", "subsidiary", and "official").

Analyzing the failure patterns, we find that these problematic samples generally have complex structures and contain multiple keywords closely associated with different categories. Such high-confusion samples are more likely to lead to the failure of prompt-tuning generalization. AMLDA addresses these challenges through several strategies that correspond to the phases illustrated above: (1) BAVC is used to extract domain-specific label mappings, extending the original "Business" to include terms like "acquisition," "subsidiary," and "merger," thereby eliminating biases through posterior frequency analysis. (2) PTMD-LDA explicitly constructs causal chains between business terminology and policy language, decoupling compound semantics. (3) DAAT dynamically injects controlled perturbations for high-confusion samples, forcing the model to distinguish between policy statements and business facts.

Furthermore, statistical comparison reinforces this advantage: in the Amazon test set, AMLDA accurately predicted 306 samples that the prompt-tuning method misclassified, whereas prompt tuning correctly identified only 42 samples that AMLDA misclassified. This comparison underscores the effectiveness of the AMLDA method in FSTC.



(**a**) AG's News (**b**) IMDB (**c**) Amazon

Figure 9. Confusion matrices of AMLDA on AG's News, IMDB, and Amazon datasets under the 20-shot setting. The matrices illustrate the model's prediction distribution, with diagonal elements representing correct predictions and off-diagonal elements indicating misclassifications. The values are row-normalized to show the recall for each true category.

Table 5. An illustrative case study of the data evolution across the AMLDA phases. We track a specific sample from the *AG's News* dataset (Class: Sports). In Phase 3 (PGG), we demonstrate that templates can be integrated at different positions (before or after the input). In Phase 4 (LES), with the augmentation size set to $L = 2$, the high-entropy candidate is filtered out to ensure semantic consistency.

Phase	Operation	Sample/Intermediate Result
Input (Original)	Raw input text x_i with Ground Truth Label	**Text:** "Michael Phelps won his eighth gold medal at the Beijing Games, breaking Mark Spitz's record." **Label:** *Sports*
Phase 1: BAVC	Bayesian Mutual Information extracts top-m label words ($v_{c,j}$)	**Adaptive Verbalizer (Top-3):** { *match, athlete, game* }
Phase 2: LEG	**Label-Enhanced Generator:** Constructs label-enriched templates (semantic components)	**Template A:** "News about { *match, athlete, game* } and [MASK]." **Template B:** "This outlines { *match, athlete, game* } in [MASK]." **Template C:** "Regarding the { *match, athlete, game* } [MASK]."
Phase 3: PGG	**Prompt-Guided Generator:** Synthesizes samples with diverse structural positions (Front/Back) relative to input x_i	**Candidate 1:** "[CLS] News about { *match, athlete, game* } and [MASK]. [SEP] Michael Phelps won... [SEP]" **Candidate 2:** "[CLS] Michael Phelps won... [SEP] This outlines { *match, athlete, game* } in [MASK]. [SEP]" **Candidate 3:** "[CLS] Regarding the { *match, athlete, game* } [MASK]. [SEP] Michael Phelps won... [SEP]"
Phase 4: LES	**Low-Entropy Selector:** Sorts candidates by Entropy $H(x)$. Set retention size $L = 2$.	**Candidate 1:** Entropy = 0.08 (Pred: Sports) $\rightarrow$ **Rank 1 (Keep)** **Candidate 2:** Entropy = 0.15 (Pred: Sports) $\rightarrow$ **Rank 2 (Keep)** **Candidate 3:** Entropy = **0.92** (Pred: Ambiguous) $\rightarrow$ **Rank 3 (Discard)** *(Reason: High entropy indicates the template "Regarding..." introduced semantic ambiguity.)*
Final Sample	Final Augmented Training Set for this instance	{ Original Input, Candidate 1, Candidate 2 }

Table 6. Cases where prompt tuning predicted wrongly but AMLDA predicted correctly, bold text in the table identifies the source dataset for each sample.

Misclassified Sample	Label	Prediction
Briefly: China interest in key Yukos unit China is interested in participating in the bidding for Yuganskneftegaz, the top oil-producing subsidiary of the Russian oil giant Yukos, a Chinese economic official was quoted as saying in a report Thursday by the Russian news agency Interfax. (**AG's News**)	Business	Politics
Security scare as intruder dives in A CANADIAN husband's love for his wife has led to a tightening of security at all Olympic venues in Athens. (**AG's News**)	Sports	Politics
Disturbing readings.... This collection is terribly read, especially the woman's voice, the strange crying tune she had bothered me so much that none of the words registered. If you like to buy an audio reading of poems, I highly recommend the collection producted by BBC, it is so far the best. (**Amazon**)	Negative	Positive
It was. This is a book about adoption. The subject description that I see listed for this book is WRONG at this time! This book is actually about: "Mary Bradford Clark, the author, is an adoptee. She gives birth to her daughter at 18, and places her for adoption. The daughter, Kathy, searches for Mary and they are reunited. Unfortunately, Kathy's adoptive parents were not the greatest people, and her life was difficult. It all ends up with Mary (her birthmom) ADOPTING Kathy!" (**Amazon**)	Positive	Negative

5. Conclusions

In this paper, we presented AMLDA, a novel framework designed to address the challenges of data scarcity and feature reuse sensitivity in FSTC. Through a Bayesian-Mutual-Information-Driven Adaptive Verbalizer Construction, we explicitly mitigated the bias in the prior knowledge of PLMs, effectively mitigating their feature reuse sensitivity. Building on this, the combination of Prompt-Template-Guided Multi-Granularity Label Data Augmentation and Difficulty-Aware Adversarial Training balances sample diversity with semantic consistency, significantly enhancing the model's generalization ability against subtle input perturbations.

Empirical evaluations across four benchmark datasets validate the efficacy of our approach. Specifically, AMLDA achieves significant F1 score improvements, such as up to +2.8% on AG's News and +1.0% on DBPedia compared to state-of-the-art baselines, demonstrating its superiority in data-deficient scenarios.

Despite these promising results, our current study has limitations. First, the experiments were primarily conducted using the RoBERTa-large backbone on English datasets. Future work will extend the AMLDA framework to other pre-trained language models (e.g., DeBERTa and BERT) and multilingual settings to assess its generalizability. Second, while our method employs multi-granularity augmentation specifically for single-label classification, applying this verbalizer-driven strategy to intrinsic multi-label classification tasks remains a valuable direction for exploration.

Author Contributions: Conceptualization, D.H.; methodology, D.H. and Z.L.; software, D.H.; validation, D.H.; formal analysis, Z.L.; investigation, Z.L.; resources, Z.L., J.Y. and Y.Z.; data curation, J.Y. and Y.Z.; writing—original draft preparation, D.H.; writing—review and editing, D.H., Z.L., J.Y. and Y.Z.; visualization, Y.Z.; supervision, D.H.; project administration, D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chongqing Municipal Education Commission Key Project of Higher Education in Chongqing, grant number CQZSKS2025012.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. The code is not publicly available due to ongoing research.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Alzubaidi, L.; Bai, J.; Al-Sabaawi, A.; Santamaría, J.; Albahri, A.S.; Al-Dabbagh, B.S.; Fadhel, M.A.; Manoufali, M.; Zhang, J.; Al-Timemy, A.H.; et al. A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *J. Big Data* **2023**, *10*, 46. [CrossRef]
2. Kotei, E.; Thirunavukarasu, R. A systematic review of transformer-based pre-trained language models through self-supervised learning. *Information* **2023**, *14*, 187. [CrossRef]
3. Zhu, K.; Wang, J.; Zhou, J.; Wang, Z.; Chen, H.; Wang, Y.; Yang, L.; Ye, W.; Zhang, Y.; Gong, N.; et al. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis, Salt Lake City, UT, USA, 14–18 October 2024; pp. 57–68.
4. Schick, T.; Schütze, H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Online, 19–23 April 2021; pp. 255–269. [CrossRef]

5. Ding, N.; Hu, S.; Zhao, W.; Chen, Y.; Liu, Z.; Zheng, H.; Sun, M. OpenPrompt: An Open-source Framework for Prompt-learning. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Dublin, Ireland, 22–27 May 2022; pp. 105–113. [CrossRef]
6. Ju, T.; Zheng, Y.; Wang, H.; Zhao, H.; Liu, G. Is continuous prompt a combination of discrete prompts? towards a novel view for interpreting continuous prompts. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 7804–7819.
7. Chen, Y.; Yang, G.; Wang, D.; Li, D. Eliciting knowledge from language models with automatically generated continuous prompts. *Expert Syst. Appl.* **2024**, *239*, 122327. [CrossRef]
8. Hambardzumyan, K.; Khachatrian, H.; May, J. WARP: Word-level Adversarial ReProgramming. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 4921–4933. [CrossRef]
9. Zhao, Z.; Wallace, E.; Feng, S.; Klein, D.; Singh, S. Calibrate Before Use: Improving Few-shot Performance of Language Models. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; Volume 139, pp. 12697–12706.
10. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [CrossRef]
11. Chen, C.; Shu, K. PromptDA: Label-guided Data Augmentation for Prompt-based Few Shot Learners. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics 2023, Dubrovnik, Croatia, 2–6 May 2023; pp. 562–574. [CrossRef]
12. Schick, T.; Schmid, H.; Schütze, H. Automatically Identifying Words That Can Serve as Labels for Few-Shot Text Classification. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 5569–5578. [CrossRef]
13. Holtzman, A.; West, P.; Shwartz, V.; Choi, Y.; Zettlemoyer, L. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 7038–7051. [CrossRef]
14. Guo, Y.; Guo, M.; Su, J.; Yang, Z.; Zhu, M.; Li, H.; Qiu, M.; Liu, S.S. Bias in large language models: Origin, evaluation, and mitigation. *arXiv* **2024**, arXiv:2411.10915. [CrossRef]
15. Wei, J.; Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 6382–6388. [CrossRef]
16. Bayer, M.; Kaufhold, M.A.; Buchhold, B.; Keller, M.; Dallmeyer, J.; Reuter, C. Data augmentation in natural language processing: A novel text generation approach for long and short text classifiers. *Int. J. Mach. Learn. Cybern.* **2023**, *14*, 135–150. [CrossRef]
17. Bayer, M.; Kaufhold, M.A.; Reuter, C. A survey on data augmentation for text classification. *ACM Comput. Surv.* **2022**, *55*, 1–39. [CrossRef]
18. Zhou, J.; Zheng, Y.; Tang, J.; Jian, L.; Yang, Z. FlipDA: Effective and Robust Data Augmentation for Few-Shot Learning. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 8646–8665. [CrossRef]
19. Abaskohi, A.; Rothe, S.; Yaghoobzadeh, Y. LM-CPPF: Paraphrasing-Guided Data Augmentation for Contrastive Prompt-Based Few-Shot Fine-Tuning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 670–681. [CrossRef]
20. Luo, Q.; Liu, L.; Lin, Y.; Zhang, W. Don't Miss the Labels: Label-semantic Augmented Meta-Learner for Few-Shot Text Classification. In Proceedings of the Findings of Association for Computational Linguistics, Online, 1–6 August 2021; pp. 2773–2782.
21. Cao, C.; Zhou, F.; Dai, Y.; Wang, J.; Zhang, K. A survey of mix-based data augmentation: Taxonomy, methods, applications, and explainability. *ACM Comput. Surv.* **2024**, *57*, 1–38. [CrossRef]
22. Sui, D.; Chen, Y.; Mao, B.; Qiu, D.; Liu, K.; Zhao, J. Knowledge Guided Metric Learning for Few-Shot Text Classification. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 3266–3271. [CrossRef]
23. Zhuo, L.; Wang, Z.; Fu, Y.; Qian, T. Prompt as free lunch: Enhancing diversity in source-free cross-domain few-shot learning through semantic-guided prompting. *arXiv* **2024**, arXiv:2412.00767.
24. Kumar, V.; Choudhary, A.; Cho, E. Data Augmentation using Pre-trained Transformer Models. In Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems, Suzhou, China, 21 September–9 October 2020; pp. 18–26.
25. Ma, L.; Liang, L. Adaptive adversarial training to improve adversarial robustness of DNNs for medical image segmentation and detection. *arXiv* **2022**, arXiv:2206.01736. [CrossRef]
26. Fang, H.; Kong, J.; Yu, W.; Chen, B.; Li, J.; Wu, H.; Xia, S.; Xu, K. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. *arXiv* **2024**, arXiv:2406.05491.

27. Zheng, H.; Zhong, Q.; Ding, L.; Tian, Z.; Niu, X.; Wang, C.; Li, D.; Tao, D. Self-Evolution Learning for Mixup: Enhance Data Augmentation on Few-Shot Text Classification Tasks. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 8964–8974. [CrossRef]
28. Aljehani, A.; Hasan, S.H.; Khan, U.A. Advancing text classification: A systematic review of few-shot learning approaches. *Int. J. Comput. Digit. Syst.* **2024**, *16*, 1–14. [CrossRef]
29. Chae, Y.; Davidson, T. Large language models for text classification: From zero-shot learning to instruction-tuning. *Sociol. Methods Res.* **2025**. [CrossRef]
30. Lei, T.; Hu, H.; Luo, Q.; Peng, D.; Wang, X. Adaptive Meta-learner via Gradient Similarity for Few-shot Text Classification. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 4873–4882.
31. Vettoruzzo, A.; Bouguelia, M.R.; Rögnvaldsson, T. Multimodal meta-learning through meta-learned task representations. *Neural Comput. Appl.* **2024**, *36*, 8519–8529. [CrossRef]
32. Huang, Z.; Shen, L.; Yu, J.; Han, B.; Liu, T. Flatmatch: Bridging labeled data and unlabeled data with cross-sharpness for semi-supervised learning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 18474–18494.
33. Chen, Y.; Mancini, M.; Zhu, X.; Akata, Z. Semi-supervised and unsupervised deep visual learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *46*, 1327–1347. [CrossRef]
34. Wei, X.S.; Xu, H.Y.; Zhang, F.; Peng, Y.; Zhou, W. An embarrassingly simple approach to semi-supervised few-shot learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 14489–14500.
35. Zhu, D.; Shen, X.; Mosbach, M.; Stephan, A.; Klakow, D. Weaker than you think: A critical look at weakly supervised learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 14229–14253.
36. Park, S.; Lee, J. LIME: Weakly-Supervised Text Classification without Seeds. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 1083–1088.
37. Wang, T.; Wang, Z.; Liu, W.; Shang, J. WOT-Class: Weakly Supervised Open-world Text Classification. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, UK, 21–25 October 2023; pp. 2666–2675.
38. Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; Tang, J. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Dublin, Ireland, 22–27 May 2022; pp. 61–68.
39. Shi, Z. Optimising Language Models for Downstream Tasks: A Post-Training Perspective. *arXiv* **2025**, arXiv:2506.20917. [CrossRef]
40. Schick, T.; Schütze, H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 2339–2352. [CrossRef]
41. Chang, K.; Xu, S.; Wang, C.; Luo, Y.; Xiao, T.; Zhu, J. Efficient Prompting Methods for Large Language Models: A Survey. *arXiv* **2024**, arXiv:2404.01077. [CrossRef]
42. Cohen, Y.; Aperstein, Y. A Review of Generative Pretrained Multi-step Prompting Schemes –and a New Multi-step Prompting Framework. *Preprints* **2024**.
43. Zhu, Y.; Wang, Y.; Mu, J.; Li, Y.; Qiang, J.; Yuan, Y.; Wu, X. Short text classification with Soft Knowledgeable Prompt-tuning. *Expert Syst. Appl.* **2024**, *246*, 123248. [CrossRef]
44. Hu, S.; Ding, N.; Wang, H.; Liu, Z.; Wang, J.; Li, J.; Wu, W.; Sun, M. Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; pp. 2225–2240. [CrossRef]
45. Yin, W.; Hay, J.; Roth, D. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3914–3923. [CrossRef]
46. Ling, T.; Chen, L.; Lai, Y.; Liu, H.L. Evolutionary Verbalizer Search for Prompt-Based Few Shot Text Classification. In Proceedings of the International Conference on Knowledge Science, Engineering and Management, Guangzhou, China, 16–18 August 2023; Springer: Cham, Switzerland, 2023; pp. 279–290.
47. Gao, T.; Fisch, A.; Chen, D. Making Pre-trained Language Models Better Few-shot Learners. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Virtual Event, 1–6 August 2021; pp. 3816–3830. [CrossRef]
48. Yang, Y.; Pedersen, J.O. A comparative study on feature selection in text categorization. In Proceedings of the 14th International Conference on Machine Learning (ICML), Nashville, TN, USA, 8–12 July 1997; Volune 97, pp. 412–420.

49. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef]
50. Newman, M.E. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **2005**, *46*, 323–351. [CrossRef]
51. Kim, M.; Tack, J.; Shin, J.; Hwang, S.J. Entropy weighted adversarial training. In Proceedings of the ICML Workshop, Online, 18–24 July 2021.
52. Zhang, X.; Zhao, J.; LeCun, Y. Character-level Convolutional Networks for Text Classification. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
53. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; Van Kleef, P.; Auer, S.; et al. Dbpedia—A large-scale, multilingual knowledge base extracted from wikipedia. *Semant. Web* **2015**, *6*, 167–195. [CrossRef]
54. Maas, A.L.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 142–150.
55. McAuley, J.; Leskovec, J. Hidden factors and hidden topics: Understanding rating dimensions with review text. In Proceedings of the 7th ACM Conference on Recommender Systems, Hong Kong, China, 12–16 October 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 165–172.
56. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 2790–2799.
57. Lester, B.; Al-Rfou, R.; Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3045–3059.
58. Ni, S.; Kao, H.Y. KPT++: Refined knowledgeable prompt tuning for few-shot text classification. *Knowl.-Based Syst.* **2023**, *274*, 110647. [CrossRef]