ISSN 2616-5775 Vol. 9, Issue 1: 63-69, DOI: 10.25236/AJCIS.2026.090108

Model-Guided Test-Time Augmentation for Robust Text Classification

Yonghao Wu[1,a], Yu Xiang[1,b,*], Wei Wang[1,c], Tongzhu Zhao[1,d], Tiancai Zhu[1,e]

[1]School of Information Science and Technology, Yunnan Normal University, Kunming, China
[a]2324100051@ynnu.edu.cn, [b]xiangyu@ynnu.edu.cn, [c]2324100048@ynnu.edu.cn,
[d]2324100063@ynnu.edu.cn, [e]2324100067@ynnu.edu.cn
[*]Corresponding author

Abstract: *Test-Time Augmentation (TTA) has the potential to improve robustness in text classification, yet its effectiveness is often limited by semantic drift introduced by indiscriminate text transformations. We propose Model-Guided Test-Time Augmentation (MG-TTA), a selective and training-free algorithm that adaptively chooses augmentation strategies for each test instance. MG-TTA leverages a pre-trained labeling model to evaluate label consistency of candidate augmented samples and selects only those most aligned with the model's prediction. The selected augmentations are combined with the original input to form a compact test-time ensemble, and final predictions are obtained by simple probability averaging. Experiments on multiple text classification benchmarks demonstrate that MG-TTA consistently outperforms fixed and random augmentation baselines, highlighting the importance of model-guided augmentation selection at inference time.*

Keywords: *Test-Time Augmentation, Selective Augmentation, Text Classification, Transformer Models, Model-Guided Inference*

1. Introduction

Transformer-based language models have achieved remarkable success in text classification and other natural language processing (NLP) tasks, largely due to large-scale pretraining and contextualized representations. Models such as BERT and its variants have become standard backbones for a wide range of applications[1,2]. Despite these advances, inference-time robustness remains a persistent challenge, as model predictions can be sensitive to minor lexical or syntactic variations in the input text[3,4].

Test-Time Augmentation (TTA) is a well-established technique for improving robustness by aggregating predictions over multiple transformed versions of the same input. In computer vision, TTA has been widely adopted and shown to provide consistent performance gains[5,6]. However, extending TTA to NLP is substantially more difficult. Unlike image transformations such as rotation or flipping, text transformations operate in a discrete space and may easily introduce semantic drift, potentially altering the underlying label of the input[7,8].

Existing text-based TTA approaches typically rely on fixed or randomly selected augmentation strategies, including synonym replacement, back-translation, and token-level perturbations[9–11]. While these methods can improve robustness in some cases, they implicitly assume that augmentation strategies are equally reliable across all samples. In practice, this assumption is often violated: different sentences exhibit varying sensitivity to specific transformations, and indiscriminate augmentation may even degrade performance for certain instances[12,13].

Recent studies have begun to explore selective or adaptive augmentation strategies for NLP, aiming to mitigate semantic drift by filtering unreliable transformations[14,15]. A key insight emerging from this line of work is that instance-level information is crucial: augmentation strategies that preserve semantic consistency for one input may not do so for another. However, many existing selective approaches require additional training stages or auxiliary models, increasing system complexity and limiting reproducibility.

In this work, we propose Model-Guided Test-Time Augmentation (MG-TTA), a simple yet effective algorithm that performs instance-wise augmentation strategy selection without introducing additional trainable components. MG-TTA leverages a pre-trained labeling model to estimate label consistency of candidate augmented samples at test time, and adaptively selects only those augmentations that best align with the model's prediction. By reusing an existing classifier as a guidance signal, MG-TTA avoids the

ISSN 2616-5775 Vol. 9, Issue 1: 63-69, DOI: 10.25236/AJCIS.2026.090108

need for explicit policy learning while retaining the benefits of selective augmentation.

MG-TTA is designed to be modular and easily integrated into standard Transformer-based classification pipelines. The method operates entirely at inference time and requires no modification to the downstream classifier architecture. Through extensive experiments on multiple text classification benchmarks, we demonstrate that MG-TTA consistently outperforms traditional fixed and random TTA baselines, highlighting the importance of model-guided augmentation selection for robust NLP inference.

The main contributions of this work are summarized as follows:

We identify and address the limitations of indiscriminate test-time augmentation in text classification by emphasizing the role of instance-level semantic consistency.

We propose MG-TTA, a training-free, model-guided framework that adaptively selects augmentation strategies at inference time using label-consistency scoring.

We conduct comprehensive experiments across multiple benchmarks, showing that MG-TTA provides stable and consistent improvements over conventional TTA baselines.

2. Related Work

2.1. Transformer-based Text Classification

Pre-trained Transformer models have become the dominant paradigm for text classification. BERT introduced bidirectional masked language modeling and achieved strong performance across a wide range of NLP tasks[1]. Subsequent work, such as RoBERTa, demonstrated that careful optimization of pretraining strategies can further improve downstream performance without architectural changes[2]. These models provide powerful contextual representations and serve as strong backbones for both training-time and inference-time enhancements.

Despite their success, Transformer-based classifiers are known to be sensitive to small input perturbations, including synonym substitutions and syntactic variations[3,4]. This sensitivity motivates the exploration of robustness-enhancing techniques beyond standard training procedures, particularly at inference time.

2.2. Text Data Augmentation

Text data augmentation has been extensively studied as a means to improve generalization during training. Early approaches focus on simple lexical operations such as synonym replacement, random insertion, deletion, and swapping, exemplified by the EDA framework[9]. Back-translation, originally proposed for leveraging monolingual data in neural machine translation, has also been widely adopted as a semantic-level augmentation technique for NLP tasks[10,11].

Morerecentworkleveragesmaskedlanguagemodelstogeneratecontext-awaresubstitutions, enabling more fluent and diverse augmentations[12]. In addition, statistical methods such as TF-IDF weighting have been used to identify salient terms for targeted perturbations[13]. While these methods can enrich training data, their effectiveness depends heavily on preserving the original label semantics.

Most existing studies on text augmentation focus on the training phase. Applying similar transformations at test time is less straightforward, as semantic drift introduced by augmentation may directly harm prediction reliability[7,8].

2.3. Test-Time Augmentation and Selective Strategies

Test-Time Augmentation (TTA) has been widely adopted in computer vision, where label-preserving transformations such as rotation and flipping are readily available[5,6]. By aggregating predictions over augmented inputs, TTA can reduce prediction variance and improve robustness. Inspired by this success, several recent works have explored TTA for NLP tasks[14,15].

However, unlike images, text transformations operate in a discrete and highly structured space, making it difficult to guarantee label preservation. As a result, naïvely applying all available augmentations or randomly sampling them often leads to unstable performance[12,16]. To address this issue, selective TTA approaches have been proposed, aiming to filter unreliable augmented samples based on heuristic rules or confidence measures[15,17].

ISSN 2616-5775 Vol. 9, Issue 1: 63-69, DOI: 10.25236/AJCIS.2026.090108

In parallel, pseudo-labeling and self-training techniques have been studied in semi-supervised learning, where model predictions are used as surrogate labels to guide further learning[18]. While pseudolabels are typically employed to expand training data, their use as a test-time guidance signal remains relatively underexplored.

MG-TTA differs from existing approaches in that it performs instance-wise augmentation selection at test time, guided by a pre-trained labeling model, without introducing additional trainable components. By directly evaluating label consistency of augmented samples under the same model, MG-TTA provides a simple and effective mechanism for selective test-time augmentation tailored to each input instance.

3. Methods

3.1. MG-TTA Algorithm

We propose Model-Guided Test-Time Augmentation (MG-TTA), a selective inference-time algorithm designed to improve robustness for text classification. The overall workflow of MG-TTA consists of three sequential stages: (1) labeling model training and pseudo-label generation, (2) model-guided augmentation selection, and (3) test-time prediction aggregation.

First, a Transformer-based classifier is fine-tuned on the training data and serves as a labeling model. This model is used to generate pseudo-labels for test instances, which define the reference labels for subsequent selection. Second, given a fixed pool of text augmentation operators, multiple augmented candidates are generated for each test instance. The labeling model evaluates these candidates by measuring their label-consistency with respect to the pseudo-label. Only the most label-consistent augmentations are retained. Finally, predictions from the original input and the selected augmented variants are aggregated to produce the final output.

This design enables instance-wise adaptive augmentation selection at test time, without introducing additional trainable modules or modifying the downstream classifier. MG-TTA Algorithm is shown in Figure 1:

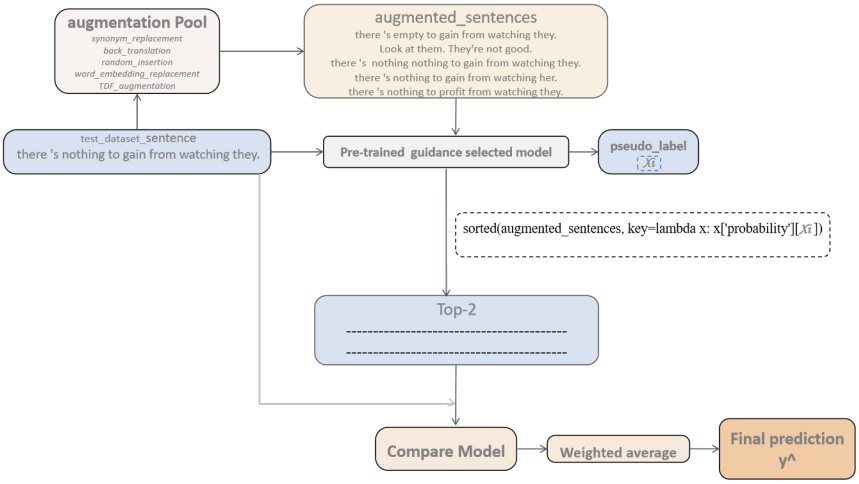

Figure 1: MG-TTA Architecture

3.2. Model-Guided Augmentation Selection and Aggregation

Let $D_{train} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the training set, where x_i is a text input and $y_i \in \{1, \ldots, C\}$ is its ground-truth label. Let $D_{test} = \{x_t\}_{t=1}^{T}$ denote the test set.

We first fine-tune a Transformer-based classifier as a labeling model, denoted by f_θ. In our implementation, f_θ is initialized from RoBERTa-base. To balance stability and task adaptability, all embedding layers are frozen, while only the last three encoder layers and the classification head are updated during fine-tuning.

Given an input sentence x, the labeling model produces a predictive distribution:

$$f_\theta(y|x) = softmax(g_\theta(x)) \quad (1)$$

ISSN 2616-5775 Vol. 9, Issue 1: 63-69, DOI: 10.25236/AJCIS.2026.090108

where $g_\theta(x)$ denotes the logits of the classifier head.

For each test instance $x_t \in D_{test}$, we obtain a pseudo-label:

$$\hat{y}_t = arg\max_{y \in \mathcal{Y}} f_\theta\,(y \mid x_t) \tag{2}$$

These pseudo-labels represent the labeling model's own decisions and are used exclusively as reference labels for augmentation selection, rather than as supervision for additional training.

Let $A = \{a_1, a_2, a_3, \ldots, a_M\}$denote a fixed pool of text augmentation operators. For each test instance x_t , we generate a set of augmented candidates:

$$\tilde{X}_t = \{\tilde{x}_{t,m} | \tilde{x}_{t,m} = a_m(x_t), a_m \in A\} \tag{3}$$

To assess whether an augmented sample preserves the semantic label of the original input, we define a label-consistency score induced by the labeling model:

$$s_{t,m} = f_\theta(\hat{y}_t | \tilde{x}_{t,m}) \tag{4}$$

Where y_t is the pseudo-label of x_t.

MG-TTA selects the top−K augmentation operators with the highest label-consistency scores:

$$A_t^* = Top - k_{a_m \in s_{t,m}} \tag{5}$$

In this work, we set K=2.For each test instance, the final inference set is constructed as:

$$X_t^* = \{x_t\} \cup \{\tilde{x}_{t,m} | a_m \in A_t^*\} \tag{6}$$

A downstream classifier h_ϕ produces predictive distributions for all elements in X_t^*. The final prediction is obtained via simple probability averaging:

$$\bar{p}(y|x_t) = \frac{1}{|X_t^*|} \sum_{x \in X_t^*} h_\phi(y|x) \tag{7}$$

$$\hat{y}_t = arg\max_{y \in \mathcal{Y}} \bar{p}\,(y \mid x_t) \tag{8}$$

3.3. Algorithm Description

For clarity and reproducibility, Algorithm 1 summarizes the complete MG-TTA procedure.

Algorithm 1 Model-Guided Test-Time Augmentation (MG-TTA)

Input: Training set D_{train} ; test set D_{test}; augmentation pool $A = \{a_1, a_2, a_3, \ldots, a_M\}$; pre-trained model M; downstream classifier h_ϕ

Output: Predicted labels

Train labeling model: fine-tune M on D_{train} to obtain f_θ
Pseudo-label generation:
for all $x_t \in D_{test}$ **do**
    $\hat{y}_t \leftarrow \arg\max_{y \in \mathcal{Y}} f_\theta\,(y \mid x_t)$
end for
Model-guided selection and inference:
for all $x_t \in D_{test}$ **do**
    for all $a_m \in A$ **do**
        $\tilde{x}_{t,m} \leftarrow a_m(x_t)$
        $s_{t,m} \leftarrow f_\theta(\hat{y}_t | \tilde{x}_{t,m})$
    end for
    Select top-K augmentation operators A_t^* according to $\{s_{t,m}\}_{m=1}^M$
    $X_t^* \leftarrow \{x_t\} \cup \{\tilde{x}_{t,m} | a_m \in A_t^*\}$
    $\hat{y}_t = arg\max_{y \in \mathcal{Y}} \frac{1}{|X_t^*|} \sum_{x \in X_t^*} h_\phi(y|x)$
end for
return $\hat{Y}$

ISSN 2616-5775 Vol. 9, Issue 1: 63-69, DOI: 10.25236/AJCIS.2026.090108

4. Experiments

4.1. Experimental Setup

We evaluate the proposed MG-TTA on five widely used text classification benchmarks: AG-News, RTE, SST-2, SST-5, and SUBJ. These datasets cover diverse classification scenarios, ranging from coarsegrained topic classification to fine-grained sentiment analysis.

For all experiments, a Transformer-based classifier is trained using the standard training and validation splits. At inference time, different test-time augmentation strategies are applied according to the compared methods. Performance is evaluated using Accuracy, Precision, Recall, and weighted-F1, as implemented in our evaluation pipeline. All reported results are obtained from the same trained downstream model to ensure fair comparison across inference strategies.

4.2. Compared Methods

We compare MG-TTA with the following inference-time baselines:

Simple: Standard inference without test-time augmentation.

T-TTA (Traditional): Fixed augmentation strategy applied uniformly to all test instances.

R-TTA (Random): Randomly selects two distinct augmentation operators for each test instance.

MG-TTA (Ours): Model-guided, instance-wise selection of the top-K label-consistent augmentations.

For all TTA-based methods, each test instance is expanded into a triplet consisting of the original input and two augmented variants. Predictions are aggregated using simple probability averaging.

4.3. Main Results

Tables 1, 2 report the Accuracy and F1-scores on all benchmarks. Overall, MG-TTA consistently outperforms both fixed and random test-time augmentation strategies across all datasets. These results demonstrate that selectively choosing label-consistent augmentations at the instance level is crucial for reliable test-time augmentation in text classification.

Notably, fixed TTA (T-TTA) occasionally underperforms the Simple baseline, highlighting the risk of semantic drift introduced by indiscriminate augmentation. In contrast, MG-TTA effectively mitigates this issue by leveraging model-guided selection, resulting in more stable and robust inference.

Table 1: Accuracy comparison across multiple datasets.

Dataset	Simple	T-TTA	R-TTA	MG-TTA
AG-News	0.9413	0.9411	0.9416	**0.9430**
RTE	0.5632	0.5596	0.5704	**0.5812**
SST-2	0.8924	0.8984	0.8995	**0.9039**
SST-5	0.5181	0.5149	0.5145	**0.5240**
SUBJ	0.9645	0.9615	0.9635	**0.9650**

Table 2: F1-scores comparison across multiple datasets.

Dataset	Simple	T-TTA	R-TTA	MG-TTA
AG-News	0.9414	0.9411	0.9416	**0.9431**
RTE	0.5507	0.5334	0.5474	**0.5547**
SST-2	0.8924	0.8984	0.8995	**0.9039**
SST-5	0.4931	0.4900	0.4858	**0.4994**
SUBJ	0.9645	0.9615	0.9635	**0.9650**

4.4. Ablation Study

To isolate the contribution of model-guided augmentation selection, we perform an ablation study to evaluate the effect of different augmentation selection strategies. Specifically, we compare MG-TTA with two baseline methods that use the same augmentation pool and the same number of augmented samples, but differ in how augmentation operators are selected.

ISSN 2616-5775 Vol. 9, Issue 1: 63-69, DOI: 10.25236/AJCIS.2026.090108

Ablation Settings. We consider the following three settings:

Fixed Selection (T-TTA): Two fixed augmentation operators are applied to all test instances.

Random Selection (R-TTA): Two augmentation operators are randomly selected for each test instance.

Model-Guided Selection (MG-TTA): Two augmentation operators are selected per test instance based on label-consistency with the pseudo-label generated by the pre-trained labeling model.

All methods use the same downstream classifier and the same probability averaging scheme for aggregation of predictions.

The results in Table 3 show the performance of each method on the SST-5 dataset.

Table 3: Ablation study on SST-5. Effect of augmentation selection strategy.

Method	Accuracy	Precision	Recall	Weighted-F1
Fixed Selection (T-TTA)	0.5632	0.5613	0.5597	0.5334
Random Selection (R-TTA)	0.5704	0.5727	0.5681	0.5474
Model-Guided Selection (MG-TTA)	**0.5812**	**0.5883**	**0.5798**	**0.5547**

Analysis. The results in Table 3 highlight the importance of model-guided augmentation selection. Although random selection (R-TTA) improves over fixed augmentation (T-TTA) by introducing diversity, it still leads to inconsistent performance due to the inclusion of semantically incompatible transformations. MG-TTA further improves performance by explicitly selecting augmentations that are consistent with the model's predicted label, leading to more stable and reliable test-time predictions.

4.5. Discussion

MG-TTA's model-guided selection enables better utilization of the augmentation pool, ensuring that only the most reliable augmentations are used during inference. This allows the method to overcome the instability introduced by traditional fixed and random augmentation strategies. The consistency in performance across various datasets also suggests that MG-TTA can be generalized to different text classification tasks, providing a robust and scalable solution for test-time augmentation.

4.6. Result Analysis

Across all datasets, MG-TTA achieves consistent improvements over both Simple inference and baseline TTA strategies. The gains are particularly pronounced on SST-5, a fine-grained sentiment classification task, where semantic drift introduced by augmentation is more likely to affect model predictions. These results indicate that instance-wise augmentation selection is especially beneficial in settings with subtle class boundaries.

On datasets such as AG-News and SUBJ, where baseline performance is already strong, MG-TTA still yields modest but consistent improvements, demonstrating that model-guided selection does not degrade performance even when augmentation benefits are limited. Overall, the experimental results confirm that MG-TTA provides a robust and general inference-time enhancement for text classification.

5. Conclusion

In this work, we propose Model-Guided Test-Time Augmentation (MG-TTA), a simple and effective method for improving test-time inference in text classification tasks. By leveraging a pre-trained labeling model to guide the selection of label-consistent augmentations, MG-TTA adapts to each test instance, ensuring that only the most reliable augmentations are used during inference.

Our experiments on multiple text classification benchmarks, including AG-News, RTE, SST-2, SST5, and SUBJ, demonstrate that MG-TTA consistently outperforms traditional fixed and random test-time augmentation strategies. Notably, MG-TTA excels in tasks like SST-5, where semantic drift introduced by indiscriminate augmentations can significantly degrade performance. By focusing on label consistency, MG-TTA provides a more stable and robust augmentation strategy that can be applied to a wide range of tasks without requiring additional training.

We further show through ablation studies that model-guided augmentation selection is the key factor in achieving improved performance. Fixed and random augmentation strategies, while simple, often

ISSN 2616-5775 Vol. 9, Issue 1: 63-69, DOI: 10.25236/AJCIS.2026.090108

suffer from instability due to the inclusion of irrelevant or semantically inconsistent augmentations. MG-TTA addresses this issue by selecting augmentations that align with the model's predicted label, thus reducing variability in predictions.

While MG-TTA performs well across several datasets, there are still opportunities for improvement. In particular, the method could benefit from exploring larger augmentation pools and more advanced selection strategies, potentially incorporating uncertainty estimation to further refine the selection process. Additionally, the computational cost of generating augmentations and performing multiple inference steps could be optimized, particularly for large-scale datasets.

In conclusion, MG-TTA offers a simple, efficient, and powerful method for improving the robustness of text classifiers at test time. Future work will explore further extensions of the approach, including the integration of additional augmentation strategies and the application of MG-TTA to more complex NLP tasks such as sequence labeling.

References

[1] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. NAACL, 2019.
[2] Liu, Y., Ott, M., Goyal, N., et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach [J]. arXiv:1907.11692, 2019.
[3] Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention Is All You Need [J]. NeurIPS, 2017.
[4] Wei, J., Zou, K.: EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks [J]. EMNLP-IJCNLP, 2019.
[5] Sennrich, R., Haddow, B., Birch, A.: Improving Neural Machine Translation Models with Monolingual Data [J]. ACL, 2016.
[6] Miller, G.A.: WordNet: A Lexical Database for English [J]. Communications of the ACM, 1995.
[7] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval [J]. Information Processing & Management, 1988.
[8] Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine Learning in Python [J]. JMLR, 2011.
[9] Wolf, T., Debut, L., Sanh, V., et al.: Transformers: State-of-the-Art Natural Language Processing [J]. EMNLP: System Demonstrations, 2020.
[10] Paszke, A., Gross, S., Massa, F., et al.: PyTorch: An Imperative Style, High-Performance Deep Learning Library [J]. NeurIPS, 2019.
[11] Lu, H., Shanmugam, D., Suresh, H., Guttag, J.: Improved Text Classification via Test-Time Augmentation [J]. arXiv:2206.13607, 2022.
[12] Wang, G., Li, W., Aertsen, M., et al.: Test-time augmentation with uncertainty estimation for deep learning-based medical image segmentation [J]. arXiv:1807.07356, 2019.
[13] Feng, S.Y., Gangal, V., Wei, J., et al.: A Survey of Data Augmentation Approaches for NLP [J]. Findings of ACL, 2021.
[14] Socher, R., Perelygin, A., Wu, J., et al.: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank [J]. EMNLP, 2013.
[15] Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization [J]. ICLR, 2019.
[16] Lee, D.-H.: Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks [J]. Workshop/Technical report, 2013.
[17] Ashukha, A., Lyzhov, A., Molchanov, D., Vetrov, D.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles [J]. NeurIPS, 2020.
[18] Sharma, M., Borisov, L., Hassani, S., Skoglund, A.: Unsupervised Learning for Robust Image Classification with Test-Time Augmentation [J]. CVPR, 2018.