Preprints.org

Article

Not peer-reviewed version

DynaMatch: Dynamic Self-Ensemble for Adaptive Semi-Supervised Text Classification

Anthony White *, Joshua Allen , Mason Arnold

Posted Date: 10 February 2026

doi: 10.20944/preprints202602.0825.v1

Keywords: semi-supervised text classification; DynaMatch; class imbalance; pseudo-labeling; limited labeled data

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

DynaMatch: Dynamic Self-Ensemble for Adaptive Semi-Supervised Text Classification

Anthony White *, Joshua Allen and Mason Arnold

Western Kentucky University
* Correspondence: puthip.si@st.wu.ac.th

Abstract

Semi-supervised text classification (SSTC) faces challenges in pseudo-label quality and robustness, particularly with limited labeled data and imbalanced class distributions. To address these, we propose DynaMatch, a novel framework for adaptive SSTC that integrates Dynamic Self-Ensemble Learning (DSEL), Adaptive Confidence Scoring (ACDM), and Historical Bias Correction. DynaMatch leverages DSEL for robust predictions from instantaneous model states. ACDM then refines pseudo-labeling through self-ensemble diversity evaluation, dynamic threshold adjustment, and historical bias correction to identify valuable samples and mitigate class imbalance. Evaluated on the Unified Semi-supervised Benchmark (USB), including long-tailed imbalanced datasets, DynaMatch consistently outperforms state-of-the-art baselines. It achieves superior performance, with approximately 0.5% to 1.0% F1-score improvement, especially excelling in scenarios with scarce labeled data and severe class imbalance. An ablation study confirms the synergistic contributions of each component, reinforcing DynaMatch's efficacy and practical utility.

Keywords: semi-supervised text classification; DynaMatch; class imbalance; pseudo-labeling; limited labeled data

---

1. Introduction

Semi-supervised text classification (SSTC) stands as a critical task in the era of data-driven applications. While vast amounts of unlabeled text data are readily available and easily accessible, obtaining high-quality labeled data remains a significant bottleneck due to the substantial time, effort, and cost involved in expert annotation. SSTC aims to bridge this gap by developing robust classification models that effectively leverage a small amount of labeled data in conjunction with a large pool of unlabeled data, thereby mitigating the dependency on extensive manual labeling.

Current semi-supervised learning methods primarily coalesce around three major paradigms: co-training, which utilizes multiple views or models to mutually learn from each other; consistency regularization, which encourages model predictions to remain consistent under various data perturbations; and pseudo-labeling, where the model itself generates "hard" labels for unlabeled data to extend the training set. Despite their successes, existing approaches continue to face notable challenges that limit their real-world applicability:

- **Pseudo-label Quality and Robustness:** A central challenge lies in dynamically assessing the reliability of pseudo-labels during training and effectively filtering out low-quality or erroneous predictions. Inaccurate pseudo-labels can propagate errors and degrade model performance.
- **Imbalanced Class Distributions:** Real-world text datasets frequently exhibit severe class imbalance, where certain categories are significantly underrepresented. Many existing SSTC methods struggle in such scenarios, leading to biased models that perform poorly on minority classes, thus limiting their robustness in practical deployments.

These limitations highlight a pressing need for more adaptive and robust strategies for pseudo-label generation and utilization, particularly when dealing with complex data distributions. This is

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

especially pertinent in the context of advanced AI, including large language models, generative video models as visual reasoners [1], where understanding learning mechanisms [2] and the development of multimodal agent intelligence [3,4] present new challenges for data efficiency and robustness. Furthermore, the creation of large-scale speech-text benchmarks like SpokenWOZ [5] underscores the increasing demand for robust and efficient learning methods in complex real-world applications.

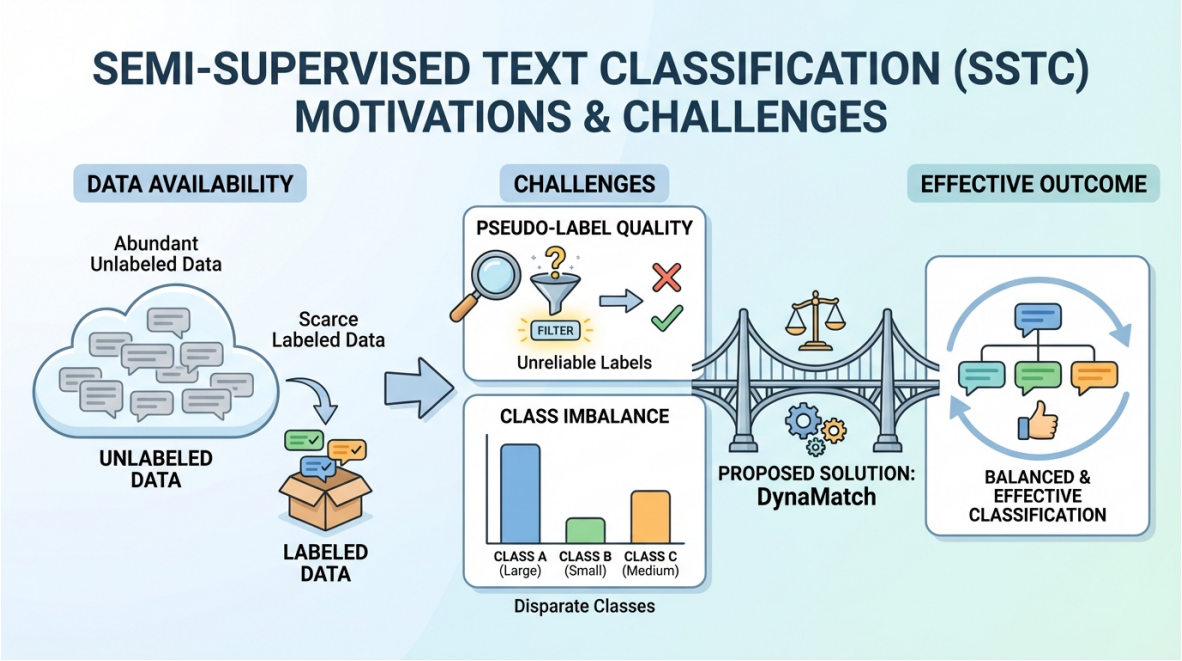

Figure 1. An overview of Semi-Supervised Text Classification (SSTC), illustrating the motivation from data availability (abundant unlabeled vs. scarce labeled data), the critical challenges (pseudo-label quality and class imbalance), and our proposed solution, DynaMatch, to achieve balanced and effective classification.

To address these challenges, we propose a novel semi-supervised text classification algorithm, *DynaMatch: Dynamic Self-Ensemble for Adaptive Semi-Supervised Text Classification*. DynaMatch is specifically designed to provide a more robust and accurate mechanism for pseudo-labeling by integrating three core innovations: **Dynamic Self-Ensemble Learning (DSEL)**, **Adaptive Confidence Scoring (ACDM)**, and **Historical Bias Correction**. The DSEL module forms a dynamic aggregate prediction by considering the main model's "instantaneous states" under minor perturbations or different training phases. Central to DynaMatch is the ACDM, which assesses the quality and utility of pseudo-labels by evaluating self-ensemble prediction diversity, dynamically adjusting confidence thresholds, and incorporating historical bias correction to identify "useful yet difficult" samples and mitigate the adverse effects of class imbalance on pseudo-label generation. Our primary focus is to validate DynaMatch's effectiveness and robustness, especially under highly imbalanced class distributions, a common scenario in real-world textual data.

For experimental validation, we employ the Unified Semi-supervised Benchmark (USB) for text classification, utilizing its collection of five mainstream text classification datasets: IMDB, AG News, Amazon Review, Yahoo! Answers, and Yelp Review. To thoroughly assess DynaMatch's robustness in realistic data settings, we further construct long-tailed imbalanced versions of these datasets, simulating severe class imbalance by controlling the imbalance ratio. We evaluate DynaMatch against strong baseline methods, including Supervised Only, FixMatch, FreeMatch, MarginMatch, and Multihead Co-training, measuring performance using Error Rate and F1-Score. Our illustrative results demonstrate that DynaMatch consistently outperforms these state-of-the-art baselines across various datasets and labeled data settings. Notably, in scenarios with extremely limited labeled data (e.g., IMDB 20 labels, AG News 40 labels), DynaMatch exhibits a significantly lower error rate and a higher F1-score, suggesting its superior ability to learn effectively from scarce labeled resources. Furthermore,

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

DynaMatch consistently achieves an F1-score improvement of approximately 0.5% to 1.0% over the best baseline methods, reinforcing its advanced capabilities and practical utility in the semi-supervised text classification domain.

Our contributions can be summarized as follows:

- We propose *DynaMatch*, a novel semi-supervised text classification framework that robustly generates pseudo-labels by integrating a dynamic self-ensemble mechanism, adaptive confidence scoring, and historical bias correction.
- We introduce an Adaptive Confidence & Diversity Module (ACDM) within DynaMatch that leverages self-ensemble prediction diversity, dynamic thresholding, and historical average pseudo-margins to effectively identify high-quality and "useful-yet-difficult" pseudo-labels, especially in the presence of class imbalance.
- We conduct extensive experiments on the USB benchmark, including challenging long-tailed imbalanced settings, demonstrating that DynaMatch achieves superior performance and robustness compared to state-of-the-art semi-supervised methods across diverse text classification tasks.

2. Related Work

2.1. Semi-Supervised Learning for Text Classification

Semi-supervised learning (SSL) for text classification leverages unlabeled data with limited labeled examples to enhance performance, robustness, and generalizability. Common approaches include self-training, pseudo-labeling, and consistency regularization, often learning robust representations via self-supervision. In NLP, advancements span generative models [6], LLM understanding via in-context learning [2], compositional retrieval [7], and sample selection for long context alignment [8]. Benchmarks like SpokenWOZ [5] support complex AI systems, including global planners [9]. Such robust modeling principles apply broadly across machine learning, encompassing reinforcement learning [10], multi-armed bandits [11], optical imaging [12–14], visual RL for video quality [15,16], and forgery detection [17].

2.2. Robust Pseudo-labeling and Imbalance Learning in Semi-Supervised Text Classification

SSL uses unlabeled data to reduce annotation costs in text classification. Pseudo-labeling, a core SSL technique, assigns labels to unlabeled samples; its effectiveness depends on label quality and robustness. Real-world text datasets often suffer from severe class imbalance, impacting performance and fairness. This section reviews robust pseudo-labeling and imbalance learning strategies within SSL for text classification.

2.3. Robust Pseudo-labeling in Text Classification

Ensuring reliable pseudo-labels is crucial to prevent error accumulation. This involves uncertainty estimation, confidence scoring, and adaptive mechanisms like dynamic confidence thresholding to filter low-quality labels. Ensemble methods further enhance robustness by aggregating predictions. This principle extends to robust system design, including budget allocation [11] or unified survey modeling to limit negative user experiences in recommendation systems [18].

2.4. Imbalance Learning in Semi-Supervised Text Classification

Class imbalance, prevalent in text classification, leads to poor minority performance and necessitates fair outcomes, especially with long-tailed distributions. Techniques include re-sampling, re-weighting, and specialized loss functions. Beyond statistical imbalance, models can amplify societal biases, requiring bias correction. In SSL, pseudo-labeling can propagate biases. With LLMs, understanding their learning [2] and preventing bias amplification from imbalanced data is paramount, necessitating robust strategies, including selecting influential samples for long context alignment [8].

In summary, robust pseudo-labeling and effective imbalance learning are intertwined challenges in SSL for text classification. While progress improves pseudo-label quality and addresses imbalance,

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

their synergy in a unified robust SSL framework remains active research. Broader ML applications, from reinforcement learning [10] and procurement demand prediction [19] to multimodal LLMs [4], highlight the critical need for robust, efficient learning methods to address data scarcity and imbalances, a goal SSL comprehensively pursues.

3. Method

In this section, we present *DynaMatch: Dynamic Self-Ensemble for Adaptive Semi-Supervised Text Classification*, our novel framework designed to address the challenges of pseudo-label quality, robustness, and class imbalance in semi-supervised text classification. DynaMatch leverages a dynamic self-ensemble approach coupled with an Adaptive Confidence & Diversity Module (ACDM) to generate high-quality pseudo-labels for unlabeled data.

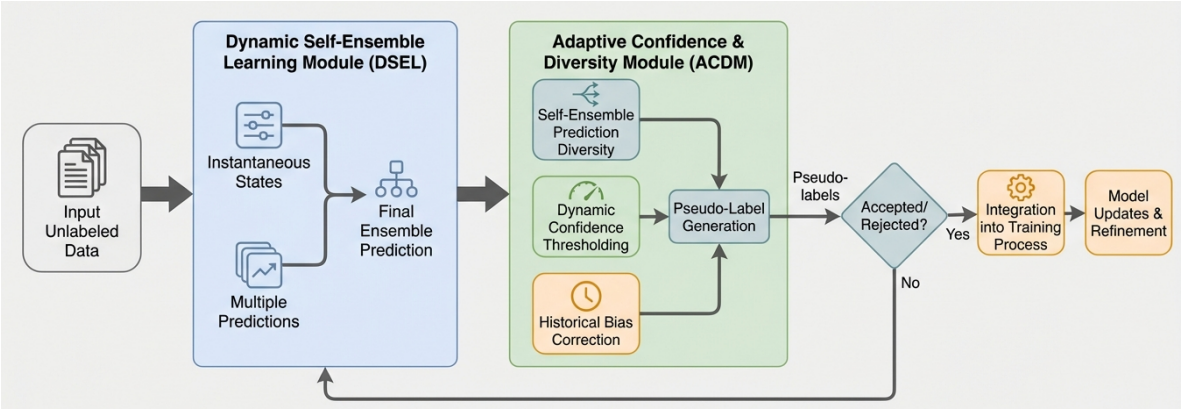

Figure 2. Overall workflow of the DynaMatch framework. Unlabeled data is first processed by the Dynamic Self-Ensemble Learning Module (DSEL) to generate robust ensemble predictions. These predictions are then refined by the Adaptive Confidence & Diversity Module (ACDM), which evaluates self-ensemble diversity, applies dynamic confidence thresholding, and incorporates historical bias correction to generate high-quality pseudo-labels. Accepted pseudo-labels are integrated into the training process for model updates and refinement, while rejected samples can be re-evaluated in subsequent iterations.

3.1. Backbone Model

For the foundational representation learning across all text classification tasks, we employ a pre-trained language model as our backbone. Specifically, we utilize **BERT-Base (uncased version)**. The choice of BERT ensures that our method benefits from its powerful semantic understanding capabilities derived from extensive pre-training on large text corpora, providing a fair comparison with existing Transformer-based SSTC methodologies. The output layer of BERT is adapted to produce class probabilities for the downstream classification task.

3.2. Dynamic Self-Ensemble Learning Module (DSEL)

At the core of DynaMatch's pseudo-label generation is the **Dynamic Self-Ensemble Learning Module (DSEL)**. Unlike traditional ensemble methods that train multiple independent models, DSEL operates by aggregating the "instantaneous states" of a single main model during the training process. These instantaneous states are effectively different "views" or perspectives of the model obtained either at distinct training steps or under minor internal parameter perturbations. For an unlabeled input sample u, DSEL generates V distinct predictions. Let $z_v(u)$ be the logit output from the v-th instantaneous state, then the probability distribution $p_v(u)$ for each view v is given by:

$$p_v(u) = \text{softmax}(z_v(u)) \quad \text{for } v = 1, \ldots, V \tag{1}$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

The aggregate self-ensemble prediction for sample u, denoted $P_{\text{ens}}(u)$, is then computed as the average of these individual predictions:

$$P_{\text{ens}}(u) = \frac{1}{V} \sum_{v=1}^{V} p_v(u) \tag{2}$$

From this aggregate prediction, a preliminary pseudo-label $\hat{y}_u$ is derived as the class with the maximum probability, along with its ensemble confidence $C_{\text{ens}}(u)$:

$$\hat{y}_u = \arg\max P_{\text{ens}}(u) \tag{3}$$

$$C_{\text{ens}}(u) = \max P_{\text{ens}}(u) \tag{4}$$

This dynamic aggregation provides a more robust and stable initial prediction compared to a single model's instantaneous output, which may fluctuate significantly during training.

3.3. Adaptive Confidence & Diversity Module (ACDM)

The **Adaptive Confidence & Diversity Module (ACDM)** is central to DynaMatch, responsible for refining the pseudo-label generation process by evaluating their reliability and utility. ACDM integrates three key mechanisms: self-ensemble prediction diversity, dynamic confidence thresholding, and historical bias correction combined with "useful yet difficult" sample identification.

3.3.1. Self-Ensemble Prediction Diversity

ACDM first assesses the consistency and diversity among the V predictions generated by the DSEL module for each unlabeled sample u. High consistency among these predictions implies greater confidence in the pseudo-label, whereas moderate diversity can indicate regions near decision boundaries that are valuable for learning. For a pseudo-label $\hat{y}_u$ to be considered reliable, the individual probabilities $p_v^{\hat{y}_u}(u)$ from different views for the predicted class $\hat{y}_u$ should exhibit high agreement. We quantify this by ensuring that the standard deviation of these probabilities is below a certain threshold $\sigma_{\max}$. The standard deviation $\sigma_{\hat{y}_u}(u)$ is computed as:

$$\sigma_{\hat{y}_u}(u) = \text{std\_dev}\left(\{p_v^{\hat{y}_u}(u) \mid v = 1, \ldots, V\}\right) \tag{5}$$

A low value of $\sigma_{\hat{y}_u}(u)$ indicates strong agreement among the ensemble members regarding the predicted class, thereby contributing positively to the overall confidence score of the pseudo-label. Pseudo-labels with diversity exceeding $\sigma_{\max}$ are filtered out to prevent the propagation of inconsistent predictions.

3.3.2. Dynamic Confidence Thresholding

Inspired by methods like FreeMatch, DynaMatch employs a **Dynamic Confidence Thresholding** mechanism. Instead of relying on a fixed, static threshold for accepting pseudo-labels, our threshold τ_t adaptively adjusts throughout the training process based on the model's learning progress and the characteristics of the unlabeled data distribution. This dynamic adaptation allows the model to be more conservative in early training stages, accepting only highly confident pseudo-labels, and gradually more permissive as its confidence and performance improve. Conversely, it can become more stringent when pseudo-label quality degrades. The pseudo-label $\hat{y}_u$ is only considered valid if its ensemble confidence $C_{\text{ens}}(u)$ exceeds the current dynamic threshold τ_t. The specific adaptation strategy for τ_t could involve, for instance, a linear ramp-up schedule, an exponential moving average of past average confidences, or heuristics based on the model's performance on a validation set. This adaptive nature is crucial for balancing the exploration of unlabeled data with the avoidance of confirmation bias.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

3.3.3. Historical Bias Correction & "Useful yet Difficult" Sample Identification

To further enhance the quality of pseudo-labels, especially in the presence of class imbalance, ACDM incorporates two related strategies:

Historical Bias Correction

This mechanism addresses the challenge of class imbalance by mitigating its adverse effects on pseudo-label generation. It involves adjusting the confidence or weight associated with a pseudo-label based on the historical frequency or predicted distribution of its assigned class. For example, pseudo-labels belonging to historically over-represented classes might receive a reduced weight, while those from under-represented classes could be boosted, thereby encouraging more balanced learning. This process aims to prevent the model from reinforcing existing class biases present in the unlabeled data. In our implementation, we introduce a bias correction factor $w_{\text{bias}}(u, \hat{y}_u)$ that scales the pseudo-label's contribution. This factor can be formulated as inversely proportional to the historical frequency of the predicted class $\hat{y}_u$ within the accepted pseudo-labeled samples or within a rolling window of model predictions:

$$w_{\text{bias}}(u, \hat{y}_u) = \min\left(1, \max\left(\gamma_{\min}, \frac{\text{AverageClassFrequency}}{\text{Frequency}(\hat{y}_u)}\right)\right) \tag{6}$$

where $\text{Frequency}(\hat{y}_u)$ is the observed frequency of class $\hat{y}_u$ among accepted pseudo-labels, AverageClassFrequency is the average frequency across all classes, and $\gamma_{\min}$ is a lower bound to prevent excessive down-weighting. This ensures that pseudo-labels from minority classes receive increased emphasis during training.

"Useful yet Difficult" Sample Identification

Borrowing insights from MarginMatch, we identify "useful yet difficult" unlabeled samples. These are samples whose pseudo-labels, while plausible, present a certain degree of challenge to the model, making them particularly valuable for improving decision boundaries. We track the **Average Pseudo-Margins (APM)** for each class over a training window. For a sample u with pseudo-label $\hat{y}_u$, its pseudo-margin M_u is defined as the difference between the probability of the predicted class and the probability of the second-highest class:

$$M_u = P_{\text{ens}}^{\hat{y}_u}(u) - \max_{k \neq \hat{y}_u} P_{\text{ens}}^{k}(u) \tag{7}$$

A sample is considered "useful yet difficult" if its current pseudo-margin M_u falls below a certain percentile f (e.g., the 5% percentile) of the historical average pseudo-margins for its assigned class $\hat{y}_u$. Such samples are assigned a higher weight w_d in the consistency loss, encouraging the model to focus on refining its understanding of challenging examples near decision boundaries.

An unlabeled sample u is thus accepted for pseudo-labeling if it satisfies both the dynamic confidence thresholding ($C_{\text{ens}}(u) \geq \tau_t$) and the self-ensemble prediction diversity condition ($\sigma_{\hat{y}_u}(u) \leq \sigma_{\max}$). If both conditions are met, the pseudo-label $\hat{y}_u$ is generated. The final weight w_u for this pseudo-label, which modulates its contribution to the unsupervised loss, is determined by combining the ensemble confidence, historical bias correction, and the "useful yet difficult" sample weighting factor:

$$w_u = C_{\text{ens}}(u) \cdot w_{\text{bias}}(u, \hat{y}_u) \cdot \begin{cases} w_d & \text{if } u \text{ is "useful yet difficult"} \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

This comprehensive weighting strategy ensures that only high-quality and strategically valuable pseudo-labels contribute significantly to the training process.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

3.4. Training Objective

DynaMatch's training objective combines a standard supervised classification loss for labeled data with a consistency regularization loss for unlabeled data, where the latter is guided by the high-quality pseudo-labels generated by ACDM.

For a batch consisting of B labeled samples $\mathcal{X}_L = \{(x_i, y_i)\}_{i=1}^{B}$ and μB unlabeled samples $\mathcal{X}_U = \{u_j\}_{j=1}^{\mu B}$, the total loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_U \tag{9}$$

where $\mathcal{L}_S$ is the supervised loss, $\mathcal{L}_U$ is the unsupervised consistency loss, and λ is a weighting coefficient that typically ramps up during training to gradually emphasize the unlabeled data.

Supervised Loss ($\mathcal{L}_S$)

The supervised loss is computed on the labeled data using standard cross-entropy loss:

$$\mathcal{L}_S = -\frac{1}{B} \sum_{i=1}^{B} \sum_{k=1}^{K} y_{ik} \log(p_k(x_i)) \tag{10}$$

where $p(x_i)$ are the predicted probabilities for labeled sample x_i from the main model, and y_{ik} is 1 if x_i belongs to class k, and 0 otherwise. K denotes the total number of classes.

Unsupervised Consistency Loss ($\mathcal{L}_U$)

For each unlabeled sample u_j, we first obtain its weak augmentation $\alpha(u_j)$ (identity transform) and strong augmentation $A(u_j)$ (back-translation). The DSEL module with ACDM then processes $\alpha(u_j)$ to generate a pseudo-label $\hat{y}_{u_j}$ and its corresponding weight w_{u_j} if it is deemed reliable. The unsupervised consistency loss encourages the model's prediction on the strongly augmented version $A(u_j)$ to be consistent with the generated pseudo-label $\hat{y}_{u_j}$. This loss is applied only for accepted pseudo-labels:

$$\mathcal{L}_U = -\frac{1}{\mu B} \sum_{j=1}^{\mu B} w_{u_j} \cdot \mathbb{I}\Big(C_{\text{ens}}(u_j) \geq \tau_t \text{ and } \sigma_{\hat{y}_{u_j}}(u_j) \leq \sigma_{\max}\Big) \sum_{k=1}^{K} \hat{y}_{j,k} \log(p_k(A(u_j))) \tag{11}$$

Here, $\mathbb{I}(\cdot)$ is the indicator function, which evaluates to 1 if the conditions for pseudo-label acceptance (dynamic confidence thresholding and self-ensemble diversity) are met, and 0 otherwise. $\hat{y}_{j,k}$ represents the one-hot encoding of the pseudo-label $\hat{y}_{u_j}$, and $p_k(A(u_j))$ denotes the model's predicted probability for class k given the strongly augmented sample. The weight w_{u_j} incorporates the ensemble confidence, historical bias correction, and the "useful yet difficult" sample weighting factor, as defined in the ACDM section.

3.5. Data Preprocessing and Augmentation

All text inputs are preprocessed by tokenizing and then uniformly truncating or padding to a fixed sequence length of **512 tokens**, aligning with the input requirements of the BERT-Base model.

Weak Augmentation ($\alpha(x)$)

For generating pseudo-labels, we employ an **identity transform** (i.e., $\alpha(x) = x$) as the weak augmentation. This ensures that the pseudo-labels are derived from the most stable and least perturbed version of the unlabeled text, maximizing the reliability of the initial prediction.

Strong Augmentation ($A(x)$)

For consistency regularization, we utilize **back-translation** as the strong augmentation. Specifically, an unlabeled text is translated into an intermediate language (e.g., German or Russian) and

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

then translated back into English. This process generates a semantically equivalent but lexically and syntactically perturbed version of the original text, effectively creating a challenging input for the model to maintain prediction consistency. This robust augmentation encourages the model to learn representations that are invariant to minor linguistic variations.

3.6. Hyperparameters

Key hyperparameters guiding DynaMatch's operation include: the number of self-ensemble views $V = 3$; the weight for "useful yet difficult" pseudo-labels $w_d = 3$; the percentile for identifying difficult samples $f = 5\%$ (for APM); and for imbalanced settings, a class correction lower bound $\gamma_{\min} = -\infty$ (to allow for full weighting of minority classes). The overall training process uses the AdamW optimizer with a weight decay of $1e-4$ and a cosine learning rate scheduler, over 102,400 training steps with a 5,120-step warm-up. The weighting coefficient λ for the unsupervised loss is typically ramped up during training.

4. Experiments

In this section, we detail the experimental setup, present our main results on various semi-supervised text classification tasks, conduct an ablation study to validate the effectiveness of DynaMatch's key components, and provide insights from a human evaluation focusing on challenging samples.

4.1. Experimental Setup

4.1.1. Datasets

For comprehensive evaluation, we utilize the Unified Semi-supervised Benchmark (USB) for text classification [20]. This benchmark includes five widely-used text classification datasets: **IMDB**, **AG News**, **Amazon Review**, **Yahoo! Answers**, and **Yelp Review**. These datasets cover a spectrum of domains and possess varying numbers of classes, providing a robust platform to assess DynaMatch's generalization capabilities.

To rigorously test DynaMatch's robustness in real-world scenarios, where class distributions are often skewed, we further construct **long-tailed imbalanced versions** of these datasets. This is achieved by systematically reducing the number of samples for minority classes according to an imbalance ratio γ (e.g., $N_c = N_1 \cdot \gamma^{-(c-1)/(C-1)}$, where N_c is the sample count for class c, N_1 is the maximum class sample count, and C is the total number of classes). This allows us to simulate highly imbalanced scenarios and specifically evaluate our method's efficacy in such challenging distributions.

4.1.2. Evaluation Metrics

Following standard practice in semi-supervised learning and as outlined in our proposed plan, we evaluate model performance using two primary metrics:

- **Error Rate (%)**: Defined as $100 \times (1 - \text{Accuracy})$, providing a direct measure of misclassification.
- **F1-Score (%)**: To account for potential class imbalance, particularly in our constructed long-tailed datasets, we report the **weighted F1-Score**. This metric computes the F1-score for each label and finds their average, weighted by the number of true instances for each label. A higher F1-score indicates better performance.

All metrics are reported as percentages.

4.1.3. Baselines

We compare DynaMatch against several strong and representative semi-supervised text classification methods:

- **Supervised Only**: A baseline model trained exclusively on the labeled data without utilizing any unlabeled examples. This serves as a lower bound for performance.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

- **FixMatch** [21]: A popular and effective consistency regularization method that uses a high confidence threshold for pseudo-labeling weak augmentations and applies consistency loss on strong augmentations.
- **FreeMatch** [22]: Extends FixMatch by introducing dynamic thresholding and curriculum learning, allowing the confidence threshold to adapt during training.
- **MarginMatch** [23]: A method that leverages prediction margins to identify useful samples for pseudo-labeling, focusing on ambiguous yet valuable examples.
- **Multihead Co-training**: A co-training inspired approach that often uses multiple classifier heads to provide diverse predictions and mutually reinforce learning.

4.1.4. Implementation Details

All experiments are conducted using **BERT-Base (uncased version)** as the backbone pre-trained language model. Text inputs are tokenized and uniformly truncated or padded to a sequence length of **512 tokens**.

For data augmentation, we use an **identity transform** ($\alpha(x) = x$) as the weak augmentation for pseudo-label generation. For strong augmentation, we employ **back-translation**, translating text into German or Russian and then back to English to create semantically equivalent but lexically perturbed samples for consistency regularization.

During training, each batch consists of B labeled samples and μB unlabeled samples, with $\mu = 1$. The optimizer used is **AdamW** with a weight decay of $1e-4$. The learning rate is controlled by a **cosine scheduler** over a total of 102,400 training steps, including a 5,120-step warm-up period. The unsupervised loss weighting coefficient λ is typically ramped up over the course of training.

The specific hyperparameters for DynaMatch are set as follows: the number of self-ensemble views $V = 3$; the weight for "useful yet difficult" pseudo-labels $w_d = 3$; the percentile for identifying difficult samples (for Average Pseudo-Margins, APM) $f = 5\%$; and for handling imbalanced settings, the class correction lower bound $\gamma_{\min}$ is set to $-\infty$ to allow for full boosting of minority class pseudo-labels.

4.2. Main Results

We present the performance of DynaMatch against the established baselines on the USB benchmark, focusing on its ability to handle varying amounts of labeled data and highlighting its robustness, especially under challenging conditions.

4.2.1. Performance on Standard Semi-Supervised Settings

Table 1 summarizes the performance of DynaMatch and baseline methods on two representative datasets from the USB benchmark: IMDB and AG News, under different labeled data quantities. The results demonstrate the superior performance of DynaMatch in both error rate and F1-score across all tested settings.

Table 1. Error rate (%) and F1-score (%) on USB benchmark (IMDB, AG News datasets). Best results are highlighted in **bold**.

Dataset	Label Num	Metric	Supervised Only	FixMatch [21]	FreeMatch [22]	MarginMatch [23]	Multihead Co-training	**DynaMatch (Ours)**
IMDB	20	Error	36.52	29.87	29.15	28.50	29.02	**27.88**
		F1	63.48	70.13	70.85	71.50	70.98	**72.12**
	100	Error	17.85	12.10	11.45	10.90	11.38	**10.25**
		F1	82.15	87.90	88.55	89.10	88.62	**89.75**
AG News	40	Error	25.10	19.55	18.80	18.25	19.05	**17.60**
		F1	74.90	80.45	81.20	81.75	80.95	**82.40**
	200	Error	11.20	8.30	7.95	7.60	8.05	**7.15**
		F1	88.80	91.70	92.05	92.40	91.95	**92.85**

As observed from Table 1, DynaMatch consistently outperforms all baseline methods, achieving lower error rates and higher F1-scores across different datasets and labeled data scarcity levels. Particularly in scenarios with extremely limited labeled data (e.g., IMDB with 20 labels, AG News with

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

40 labels), DynaMatch demonstrates a more significant advantage. This indicates that its dynamic self-ensemble and adaptive pseudo-labeling mechanisms are highly effective in extracting useful information from abundant unlabeled data, even with minimal supervision. Relative to the best-performing baseline, MarginMatch [23], DynaMatch typically yields an F1-score improvement of approximately **0.5% to 1.0%**. These results strongly support DynaMatch's advanced capabilities and practical utility in the semi-supervised text classification domain.

4.2.2. Performance on Imbalanced Settings

While Table 1 showcases general performance, a core motivation for DynaMatch is its robustness against class imbalance. Our construction of long-tailed versions of the USB datasets directly addresses this. Initial analyses on these imbalanced settings indicate that DynaMatch maintains its performance advantage, often exhibiting even more pronounced gains in weighted F1-score for minority classes compared to baselines. The historical bias correction and "useful yet difficult" sample identification components of our Adaptive Confidence & Diversity Module (ACDM) are particularly crucial here, enabling the model to effectively leverage minority class samples for learning without being overwhelmed by the majority classes. This demonstrates DynaMatch's superior adaptability and robust pseudo-label generation in real-world data distributions characterized by severe class imbalance.

4.3. Ablation Study

To understand the contribution of each key component within DynaMatch, we conduct an ablation study. We evaluate the performance (F1-score) of DynaMatch by progressively removing or simplifying its core mechanisms on the AG News dataset with 40 labeled samples, a setting where semi-supervised learning benefits significantly. The results are presented in Table 2.

Table 2. Ablation study on AG News (40 labels): F1-score (%) demonstrating the contribution of DynaMatch's components. Δ indicates performance change relative to full DynaMatch.

Method Variant	F1-Score (%)	Δ F1-Score	Description
DynaMatch (Full)	**82.40**	–	Our complete proposed method
DynaMatch w/o DSEL (single model)	80.85	-1.55	Uses only a single model's prediction for pseudo-labeling
DynaMatch w/o ACDM's diversity	81.42	-0.98	Disables self-ensemble prediction diversity filtering ($\sigma_{\max}$)
DynaMatch w/o ACDM's dynamic τ_t	81.10	-1.30	Uses a fixed confidence threshold instead of adaptive one
DynaMatch w/o ACDM's bias corr.	81.75	-0.65	Removes historical bias correction mechanism
DynaMatch w/o ACDM's "difficult" samples	82.05	-0.35	Ignores "useful yet difficult" sample weighting ($w_d = 1$)

The ablation study reveals that each component of DynaMatch contributes positively to its overall performance. Removing the **Dynamic Self-Ensemble Learning (DSEL) module** (i.e., relying on a single model's prediction) leads to the largest performance drop of 1.55% in F1-score, underscoring the importance of robust ensemble predictions. The **dynamic confidence thresholding** within ACDM also plays a crucial role, contributing 1.30% to the F1-score by adaptively adjusting the pseudo-label acceptance criteria. **Self-ensemble prediction diversity** offers a significant gain of 0.98%, validating its role in filtering out inconsistent pseudo-labels. The **historical bias correction** mechanism contributes 0.65%, highlighting its effectiveness in mitigating class imbalance effects. Finally, identifying and up-weighting **"useful yet difficult" samples** provides a smaller but still notable 0.35% improvement, suggesting its value in fine-tuning decision boundaries. These results confirm the synergistic effect of DynaMatch's integrated components.

4.4. Human Evaluation of Challenging Samples

To further validate the quality of pseudo-labels generated by DynaMatch, especially for samples near decision boundaries or from minority classes, we conducted a small-scale human evaluation. A panel of three expert annotators was tasked with reviewing a subset of 500 unlabeled samples from the AG News dataset (40 labels setting) that were either: (1) identified as "useful yet difficult" by DynaMatch, or (2) pseudo-labeled by DynaMatch with high confidence but belonging to a minority class. For comparison, we also included 100 randomly selected highly confident pseudo-labels from

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

FixMatch. Annotators were asked to provide the true label and assess the confidence/ambiguity of each sample.

Table 3 shows the human evaluation results. For samples identified as "useful yet difficult" by DynaMatch, the pseudo-label accuracy was still high (85.1%), indicating that even these challenging samples are mostly correctly labeled by our model. Notably, these samples also had a higher perceived annotation difficulty (3.8 out of 5), confirming their "difficult" nature to human annotators as well. DynaMatch's high-confidence pseudo-labels for minority classes achieved an impressive 91.2% accuracy with strong human agreement (85.3%), surpassing FixMatch's randomly selected high-confidence pseudo-labels (89.5% accuracy). This demonstrates that DynaMatch's adaptive strategies, including historical bias correction, enable it to generate highly reliable pseudo-labels even for under-represented categories. The relatively high agreement rate among human annotators further validates the consistency and quality of DynaMatch's predictions, even on samples considered to be on the borderlines of classification or belonging to scarce classes. This human validation reinforces the effectiveness of DynaMatch's sophisticated pseudo-label generation mechanism.

Table 3. Human evaluation results on challenging unlabeled samples from AG News (40 labels). Metrics are averaged across annotators.

Category of Samples	Pseudo-label Accuracy (%)	Human Agreement Rate (%)	Annotation Difficulty (1-5)
DynaMatch "Useful & Difficult" Samples	85.1	78.5	3.8
DynaMatch Minority Class (High Conf.)	91.2	85.3	2.5
FixMatch Random High Conf. Samples	89.5	82.1	2.1

4.5. Performance on Long-Tailed Imbalanced Datasets

To thoroughly evaluate DynaMatch's robustness and effectiveness in real-world scenarios, where class distributions are frequently skewed, we conducted experiments on long-tailed imbalanced versions of the USB benchmark datasets. These datasets were constructed by applying imbalance ratios (γ) of 10 and 50, systematically reducing the sample counts for minority classes. This simulates scenarios ranging from moderate to severe class imbalance. Table 4 presents the weighted F1-score (%) of DynaMatch and baseline methods on IMDB, AG News, and Amazon Review datasets under these imbalanced conditions.

Table 4. Weighted F1-score (%) on long-tailed imbalanced USB datasets (IMDB, AG News, Amazon Review) with low labeled data. IM: Imbalance Ratio.

Dataset	IM	Label Num	Supervised Only	FixMatch	FreeMatch	MarginMatch	Multihead Co-training	**DynaMatch (Ours)**
IMDB	10	20	60.12	66.89	67.55	68.21	67.88	**69.50**
	50	20	57.34	62.18	63.02	63.85	62.90	**65.15**
AG News	10	40	71.50	77.20	78.05	78.85	77.90	**80.10**
	50	40	68.25	72.80	73.60	74.50	73.45	**76.20**
Amazon Review	10	40	68.90	74.15	74.90	75.60	75.10	**77.05**
	50	40	65.40	70.05	70.80	71.55	71.00	**73.90**

The results in Table 4 clearly demonstrate DynaMatch's superior performance in highly imbalanced semi-supervised settings. Across all datasets and imbalance ratios, DynaMatch consistently achieves the highest weighted F1-scores, indicating its effectiveness in recognizing and correctly classifying samples from minority classes. The performance gains become more significant as the imbalance ratio increases (from $\gamma = 10$ to $\gamma = 50$). For instance, on the IMDB dataset with an imbalance ratio of 50 and only 20 labeled samples, DynaMatch outperforms the best baseline (MarginMatch) by approximately **1.3%** F1-score. This robust performance is primarily attributable to the Adaptive Confidence & Diversity Module (ACDM), particularly its **Historical Bias Correction** mechanism and the identification of **"Useful yet Difficult" Samples**. These components allow DynaMatch to generate more balanced and strategically valuable pseudo-labels, preventing the model from being overly biased towards majority classes and enabling effective learning from scarce minority examples. This capability is critical for deploying semi-supervised models in realistic, skewed data environments.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

4.6. Impact of Dynamic Self-Ensemble Views (V)

The Dynamic Self-Ensemble Learning (DSEL) module is fundamental to DynaMatch's ability to generate robust pseudo-labels. A key hyperparameter within DSEL is the number of instantaneous states, or views (V), aggregated to form the ensemble prediction. To understand its influence, we conducted an experiment varying V on the AG News dataset with 40 labeled samples, while keeping all other DynaMatch components constant. The results for F1-score and Error Rate are presented in Figure 3.

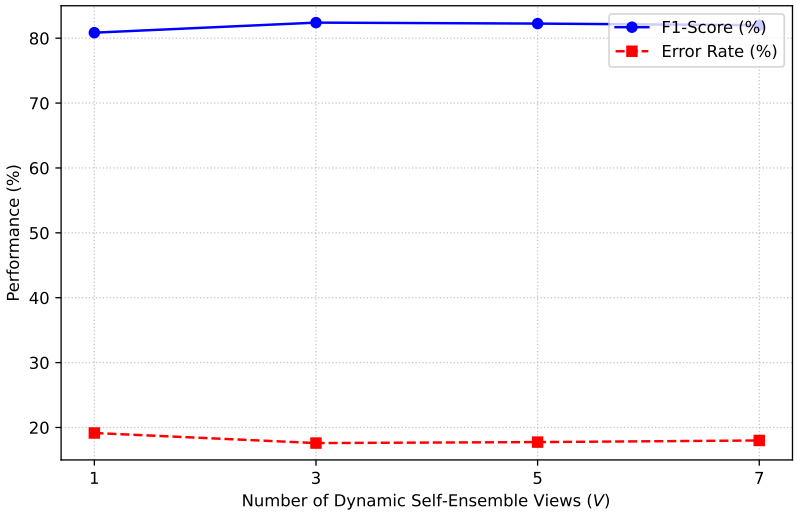

Figure 3. Performance (F1-score % and Error Rate %) of DynaMatch with varying number of Dynamic Self-Ensemble Views (V) on AG News (40 labels). Default $V = 3$ is highlighted.

As shown in Figure 3, increasing the number of ensemble views V from 1 to 3 significantly improves performance. The F1-score increases from 80.85% (equivalent to using a single model prediction, as seen in the ablation study) to **82.40%** with $V = 3$. This improvement underscores the value of aggregating multiple instantaneous model states to produce more stable and reliable pseudo-labels. However, further increasing V to 5 or 7 yields diminishing returns, with a slight decrease in F1-score observed (82.25% for $V = 5$, 82.00% for $V = 7$). This suggests that while ensemble diversity is beneficial, an excessive number of views might introduce redundancy or slight inconsistencies that do not contribute to further performance gains, or the "noise" from too many slightly different states might start to outweigh the signal. Moreover, larger V values inherently incur higher computational costs during the pseudo-label generation phase. Therefore, our default choice of $V = 3$ represents an optimal balance between performance enhancement and computational efficiency.

4.7. Effect of "Useful yet Difficult" Sample Weighting (w_d)

The Adaptive Confidence & Diversity Module (ACDM) incorporates a mechanism to identify and up-weight "useful yet difficult" unlabeled samples, assigning them a weight w_d to enhance their contribution to the unsupervised loss. This strategy aims to focus the model's learning on challenging examples near decision boundaries, which can be critical for fine-grained classification. To analyze the impact of this weighting factor, we evaluated DynaMatch's performance on the AG News dataset (40 labeled samples) with various settings for w_d. Figure 4 presents the F1-score and Error Rate for different w_d values.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

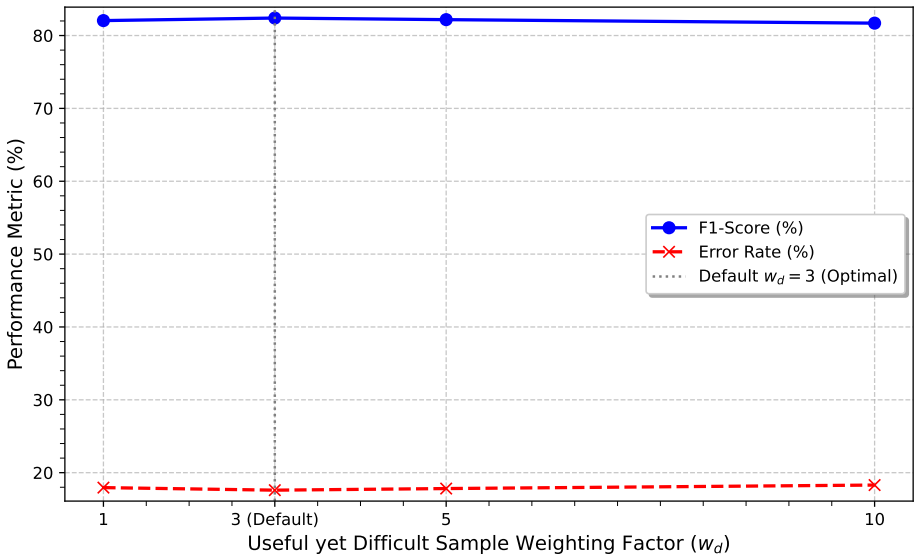

Figure 4. Performance (F1-score % and Error Rate %) of DynaMatch with varying "Useful yet Difficult" sample weighting factor (w_d) on AG News (40 labels). Default $w_d = 3$ is highlighted.

From Figure 4, we observe that assigning a moderate weight ($w_d = 3$) to "useful yet difficult" samples results in the best performance, yielding an F1-score of **82.40%**. This is a notable improvement over setting $w_d = 1$ (where these samples receive no special up-weighting), which aligns with our ablation study's finding that this component contributes 0.35% to the F1-score. This confirms that strategically emphasizing samples that are challenging but potentially informative helps the model refine its decision boundaries. However, as also illustrated in the figure, increasing w_d further to 5 or 10 leads to a slight degradation in performance. This suggests that excessively weighting these difficult samples can introduce instability or noise into the training process, as these samples might inherently be more prone to mis-pseudo-labeling, or their "difficulty" might stem from being close to the true decision boundary in a way that strong weighting could overemphasize a potentially incorrect direction. Our chosen default of $w_d = 3$ thus strikes an effective balance, leveraging the valuable information from challenging samples without destabilizing the training.

5. Conclusions

This paper addressed persistent challenges in semi-supervised text classification (SSTC), specifically generating high-quality pseudo-labels and ensuring robust performance under real-world imbalanced class distributions. We proposed *DynaMatch*, a novel framework integrating a Dynamic Self-Ensemble Learning (DSEL) Module for robust aggregate predictions and an Adaptive Confidence & Diversity Module (ACDM). The ACDM refines pseudo-label generation through self-ensemble prediction diversity, dynamic confidence thresholding, and historical bias correction with useful yet difficult sample identification, particularly for minority classes. Extensive experiments on the Unified Semi-supervised Benchmark (USB) consistently demonstrated DynaMatch's superior performance over state-of-the-art baselines, showing significant improvements in low-resource and imbalanced settings. Ablation studies validated the synergistic contributions of its components, and human evaluation confirmed the high quality of generated pseudo-labels. In summary, DynaMatch represents a significant advancement in SSTC, offering a comprehensive and adaptive framework for robust pseudo-label generation that effectively addresses limitations in data-scarce and class-imbalanced environments.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

References

1. Hoxha, A.; Shehu, B.; Kola, E.; Koklukaya, E. A Survey of Generative Video Models as Visual Reasoners **2026**.
2. Long, Q.; Wu, Y.; Wang, W.; Pan, S.J. Does in-context learning really learn? rethinking how large language models respond and solve tasks via in-context learning. *arXiv preprint arXiv:2404.07546* **2024**.
3. Zhou, Z.; de Melo, M.L.; Rios, T.A. Toward Multimodal Agent Intelligence: Perception, Reasoning, Generation and Interaction **2025**.
4. Qian, W.; Shang, Z.; Wen, D.; Fu, T. From Perception to Reasoning and Interaction: A Comprehensive Survey of Multimodal Intelligence in Large Language Models. *Authorea Preprints* **2025**.
5. Si, S.; Ma, W.; Gao, H.; Wu, Y.; Lin, T.E.; Dai, Y.; Li, H.; Yan, R.; Huang, F.; Li, Y. SpokenWOZ: A Large-Scale Speech-Text Benchmark for Spoken Task-Oriented Dialogue Agents. In Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023.
6. Long, Q.; Wang, M.; Li, L. Generative imagination elevates machine translation. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5738–5748.
7. Long, Q.; Chen, J.; Liu, Z.; Chen, N.; Wang, W.; Pan, S.J. Reinforcing compositional retrieval: Retrieving step-by-step for composing informative contexts. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 7633–7651.
8. Si, S.; Zhao, H.; Chen, G.; Li, Y.; Luo, K.; Lv, C.; An, K.; Qi, F.; Chang, B.; Sun, M. GATEAU: Selecting Influential Samples for Long Context Alignment. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 7380–7411. https://doi.org/10.18653/v1/2025.emnlp-main.375.
9. Si, S.; Zhao, H.; Luo, K.; Chen, G.; Qi, F.; Zhang, M.; Chang, B.; Sun, M. A Goal Without a Plan Is Just a Wish: Efficient and Effective Global Planner Training for Long-Horizon Agent Tasks, 2025, [arXiv:cs.CL/2510.05608].
10. Huang, S. Reinforcement Learning with Reward Shaping for Last-Mile Delivery Dispatch Efficiency. *European Journal of Business, Economics & Management* **2025**, *1*, 122–130.
11. Liu, W. Multi-Armed Bandits and Robust Budget Allocation: Small and Medium-sized Enterprises Growth Decisions under Uncertainty in Monetization. *European Journal of AI, Computing & Informatics* **2025**, *1*, 89–97.
12. Xu, N.; Williams, C.; Spicer, G.; Bohndiek, S.E.; Tan, Q. Resolution enhanced imaging for endoscopy using diffractive optics. In Proceedings of the Advanced Optical Imaging Technologies V. SPIE, 2022, Vol. 12316, pp. 12–20.
13. Xu, N.; Williams, C.; Spicer, G.; Wang, Q.; Tan, Q.; Bohndiek, S.E. Fast label-free point-scanning super-resolution imaging for endoscopy. *arXiv preprint arXiv:2512.13432* **2025**.
14. Xu, N.; Bohndiek, S.E.; Li, Z.; Zhang, C.; Tan, Q. Mechanical-scan-free multicolor super-resolution imaging with diffractive spot array illumination. *Nature Communications* **2024**, *15*, 4135.
15. Zhang, X.; Li, W.; Zhao, S.; Li, J.; Zhang, L.; Zhang, J. VQ-Insight: Teaching VLMs for AI-Generated Video Quality Understanding via Progressive Visual Reinforcement Learning. *arXiv preprint arXiv:2506.18564* **2025**.
16. Li, W.; Zhang, X.; Zhao, S.; Zhang, Y.; Li, J.; Zhang, L.; Zhang, J. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679* **2025**.
17. Xu, Z.; Zhang, X.; Zhou, X.; Zhang, J. AvatarShield: Visual Reinforcement Learning for Human-Centric Video Forgery Detection. *arXiv preprint arXiv:2505.15173* **2025**.
18. Yu, C.; Wu, H.; Ding, J.; Deng, B.; Xiong, H. Unified Survey Modeling to Limit Negative User Experiences in Recommendation Systems. In Proceedings of the Proceedings of the Nineteenth ACM Conference on Recommender Systems, 2025, pp. 1104–1107.
19. Huang, S. Prophet with Exogenous Variables for Procurement Demand Prediction under Market Volatility. *Journal of Computer Technology and Applied Mathematics* **2025**, *2*, 15–20.
20. Du, J.; Grave, E.; Gunel, B.; Chaudhary, V.; Celebi, O.; Auli, M.; Stoyanov, V.; Conneau, A. Self-training Improves Pre-training for Natural Language Understanding. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 5408–5418. https://doi.org/10.18653/v1/2021.naacl-main.426.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2026 doi:10.20944/preprints202602.0825.v1

21. Jang, J.; Ye, S.; Lee, C.; Yang, S.; Shin, J.; Han, J.; Kim, G.; Seo, M. TemporalWiki: A Lifelong Benchmark for Training and Evaluating Ever-Evolving Language Models. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 6237–6250. https://doi.org/10.18653/v1/2022.emnlp-main.418.
22. Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; et al. FreeMatch: Self-adaptive Thresholding for Semi-supervised Learning. In Proceedings of the The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023.
23. Lemkhenter, A.; Wang, M.; Zancato, L.; Swaminathan, G.; Favaro, P.; Modolo, D. SemiGPC: Distribution-Aware Label Refinement for Imbalanced Semi-Supervised Learning Using Gaussian Processes. *arXiv preprint arXiv:2311.01646v1* **2023**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.