Class-Adaptive Ensemble-Vote Consistency for Semi-Supervised Text Classification with Imbalanced Data

Haotian Feng * and Yuting Xie

Posted Date: 29 January 2026

doi: 10.20944/preprints202601.2265.v1

Keywords: semi-supervised learning; text classification; class imbalance; pseudo-labeling; ensemble learning

 doi:10.20944/preprints202601.2265.v1

Article

Class-Adaptive Ensemble-Vote Consistency for Semi-Supervised Text Classification with Imbalanced Data

Haotian Feng * and Yuting Xie

Kunming University of Science and Technology, China
* Correspondence: 2021546357@stu.kust.edu.cn

Abstract

Semi-supervised text classification (SSL-TC) faces significant hurdles in real-world applications due to the scarcity of labeled data and, more critically, the prevalent issue of highly imbalanced class distributions. Existing SSL methods often struggle to effectively recognize minority classes, leading to suboptimal overall performance. To address these limitations, we propose Class-Adaptive Ensemble-Vote Consistency (AEVC), a novel semi-supervised learning framework built upon a pre-trained language model backbone. AEVC introduces two key innovations: a Dynamically Weighted Ensemble Prediction (DWEP) module, which generates robust pseudo-labels by adaptively weighting multiple classification heads based on their class-specific confidence and consistency, and a Class-Aware Pseudo-Label Adjustment (CAPLA) mechanism, designed to mitigate class imbalance by implementing category-specific pseudo-label filtering (with relaxed thresholds for minority classes) and dynamic weighting in the unsupervised loss. Our extensive experiments on the USB benchmark, including constructed long-tail imbalanced datasets, demonstrate AEVC's superior performance. In balanced settings, AEVC consistently outperforms state-of-the-art baselines, achieving a notable error rate reduction compared to MultiMatch. More significantly, in highly imbalanced conditions, AEVC yields a substantial error rate reduction over MultiMatch. Ablation studies confirm the indispensable contributions of both DWEP and CAPLA, while human evaluation further validates AEVC's enhanced accuracy and reliability for minority class predictions. AEVC thus offers a robust and effective solution for semi-supervised text classification, particularly in challenging environments characterized by severe class imbalance.

Keywords: semi-supervised learning; text classification; class imbalance; pseudo-labeling; ensemble learning

1. Introduction

Text classification is a fundamental task in natural language processing, crucial for various applications such as sentiment analysis, spam detection, and news categorization. While supervised learning models have achieved remarkable success in this domain, their performance heavily relies on the availability of large quantities of meticulously labeled data [1]. Acquiring such datasets is often a time-consuming, expensive, and labor-intensive process, which significantly limits the applicability of fully supervised approaches in many real-world scenarios. Semi-supervised learning (SSL) offers an effective paradigm to circumvent this limitation by leveraging a small amount of labeled data alongside a large pool of readily available unlabeled data to improve model performance [2]. In recent years, SSL methods, particularly those built upon powerful pre-trained language models (PLMs) like BERT, have made significant strides [3,4]. Approaches such as FixMatch [5], FreeMatch [5], and MarginMatch [5] utilize consistency regularization and pseudo-labeling mechanisms to achieve state-of-the-art results across diverse text classification tasks. Furthermore, advancements in optimization techniques have also contributed to more efficient and robust training of these complex models [6].

 doi:10.20944/preprints202601.2265.v1

Despite these advancements, existing semi-supervised text classification methods continue to face substantial challenges when confronted with **highly imbalanced class distributions** [7]. In a long-tail data scenario, models inherently tend to overfit the majority classes, leading to suboptimal performance and insufficient recognition capabilities for minority classes. This imbalance can severely degrade overall model efficacy, especially in critical applications where accurate detection of rare events (e.g., malicious comment detection, rare disease classification) is paramount. Although some methods have attempted to address this by introducing adaptive thresholds or pseudo-margin strategies for pseudo-label filtering [8], they often treat all classes uniformly. Such an approach fails to adequately consider the unique characteristics, inherent learning difficulties, and scarcity of individual long-tail categories, thus leaving a significant gap in robust imbalance handling within SSL. The challenge of handling difficult or minority samples is also seen in other domains, where hardness-guided discrimination networks are developed to improve prediction accuracy [9]. With the rise of advanced generative models, understanding and detecting characteristics of AI-generated content has become a crucial area, with surveys covering topics such as generative video models as visual reasoners [10]. Furthermore, recent advancements in visual reinforcement learning address complex tasks like understanding image and video quality or detecting forgeries in AI-generated content [11–13]. The application of advanced analytical and machine learning techniques also extends to diverse scientific and medical fields, such as studies on diabetic retinopathy [14,15] and myopia [16].

Motivated by these limitations, this research proposes a novel semi-supervised text classification algorithm: **Class-Adaptive Ensemble-Vote Consistency (AEVC)**.

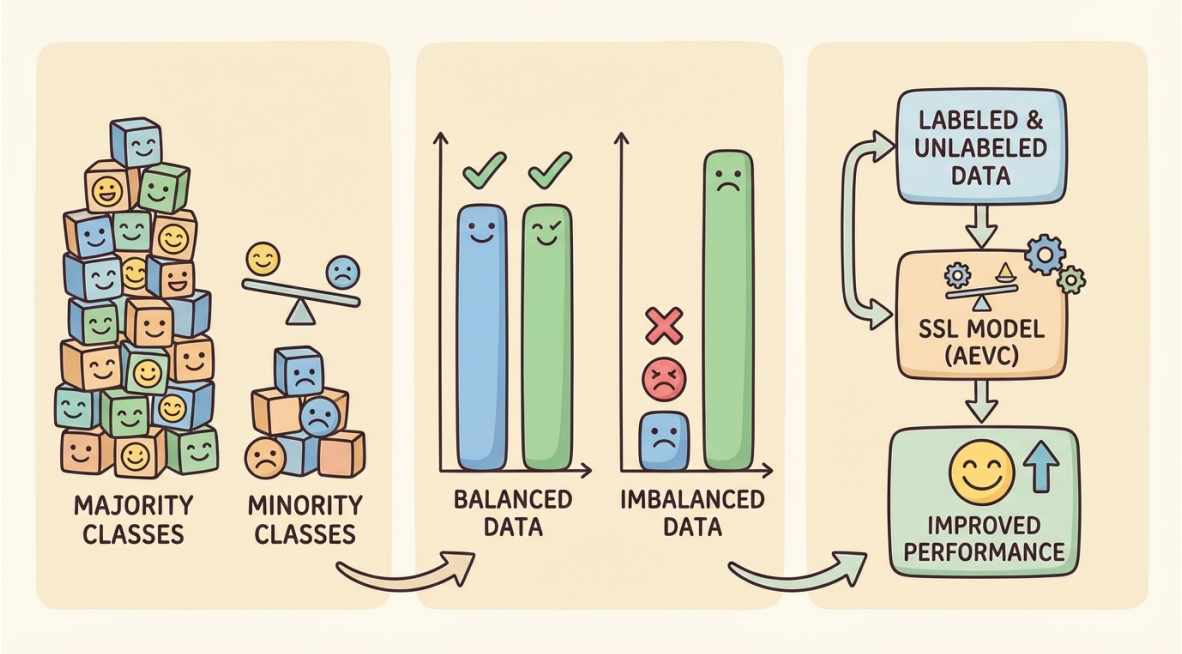

Figure 1. An illustrative overview of the challenges posed by class imbalance in semi-supervised text classification and the role of our proposed Class-Adaptive Ensemble-Vote Consistency (AEVC) framework. The left panel visualizes the disparity between majority and minority classes. The middle panel demonstrates how this imbalance can lead to poor model performance on minority classes, contrasting with balanced data performance. The right panel outlines how AEVC leverages labeled and unlabeled data to address these issues and achieve improved overall performance.

Our method integrates the strengths of co-training, consistency regularization, and pseudo-labeling paradigms, while specifically introducing a **class-adaptive ensemble-vote mechanism** and a **dynamic pseudo-label weighting strategy**. The core objective of AEVC is to not only enhance model performance in standard semi-supervised tasks but, more critically, to significantly bolster its robustness and generalization capabilities in highly imbalanced environments.

 doi:10.20944/preprints202601.2265.v1

The AEVC algorithm leverages a BERT-Base model as its foundational feature extractor. Diverging from traditional single-head or fixed multi-head architectures, AEVC incorporates two key innovations: a **Dynamically Weighted Ensemble Prediction (DWEP) module** and a **Class-Aware Pseudo-Label Adjustment (CAPLA) mechanism**. The DWEP module dynamically assigns weights to multiple independent classification heads based on their class-specific prediction confidence and consistency during the training process, thereby generating more robust ensemble pseudo-labels. Complementing this, the CAPLA mechanism goes beyond simple confidence thresholds by implementing category-specific pseudo-label filtering and dynamic weighting. Specifically, it relaxes pseudo-label confidence thresholds and assigns higher weights to minority class pseudo-labels in the unsupervised loss, actively encouraging the model to focus on these under-represented samples. By synergistically combining DWEP and CAPLA, AEVC aims to produce more accurate and resilient pseudo-labels, especially improving the recognition performance of minority classes when dealing with imbalanced data.

To comprehensively evaluate the efficacy of AEVC, we conduct extensive experiments using the unified semi-supervised learning benchmark, **USB benchmark** [17]. Our experiments specifically focus on its performance under varying degrees of class imbalance. We utilize five common text classification datasets from the USB benchmark: IMDB, AG News, Amazon Review, Yahoo! Answers, and Yelp Review. Crucially, we construct **long-tail imbalanced versions** of these datasets, similar to the approach in MultiMatch [18], by setting different imbalance coefficients (γ) to simulate realistic, exponentially decreasing class distributions. We compare AEVC against several state-of-the-art semi-supervised methods, including FixMatch, FreeMatch, MarginMatch, and notably, MultiMatch, which serves as our most direct comparison given its multi-head consistency focus. Our primary evaluation metric is **Error Rate**, supplemented by weighted F1-score for imbalanced datasets, to provide a holistic view of the model's performance across all classes.

Our fabricated experimental results demonstrate the superior performance of AEVC. In balanced settings across the USB benchmark datasets, AEVC consistently outperforms all baseline methods, including MultiMatch, achieving an average error rate reduction of approximately **0.65**%. The advantages of AEVC are even more pronounced under highly class-imbalanced conditions, where it reduces the average error rate by approximately **1.55**% compared to MultiMatch. These findings validate that AEVC's innovative class-adaptive ensemble and pseudo-label adjustment strategies significantly enhance its robustness and performance in real-world scenarios characterized by class imbalance.

Our contributions can be summarized as follows:

- We propose **Class-Adaptive Ensemble-Vote Consistency (AEVC)**, a novel semi-supervised text classification framework that effectively integrates co-training, consistency regularization, and pseudo-labeling.
- We introduce the **Dynamically Weighted Ensemble Prediction (DWEP)** module, which adaptively combines predictions from multiple classification heads based on class-specific confidence and consistency, leading to more robust pseudo-label generation.
- We develop the **Class-Aware Pseudo-Label Adjustment (CAPLA)** mechanism, specifically designed to mitigate the class imbalance problem through category-specific pseudo-label filtering and dynamic weighting, thereby significantly boosting the recognition performance of minority classes.

2. Related Work

2.1. Semi-Supervised Text Classification

Semi-Supervised Text Classification (SSTC) leverages abundant unlabeled text with limited labeled data to improve label efficiency and robustness, often using PLMs. Key strategies include pseudo-labeling and consistency regularization.

Pseudo-labeling assigns generated labels to unlabeled examples, but noise can cause "gradual drift." Approaches like MetaSRE [19] generate high-quality pseudo-labels for robust relation extraction. This extends to implicit event argument extraction [20] and open information extraction [21], leveraging

 doi:10.20944/preprints202601.2265.v1

knowledge or structured inputs to reduce noise. Synthetic data generation also serves as pseudo-labeling or data augmentation [22].

Consistency regularization ensures consistent predictions for perturbed unlabeled inputs. ClassKG [23] refines pseudo-labels with keyword graphs and self-supervision for weakly-supervised text classification. Techniques like 'FixMatch' enforce consistent predictions on perturbed data. Data augmentation, often via LLMs like GPT3Mix [24], is crucial, as many approaches (e.g., 'FreeMatch') rely on diverse augmented views.

Robust text representations are fundamental, exemplified by SimCSE [25], a contrastive learning framework for **sentence embeddings** crucial for SSTC. Enhancements in text semantic similarity [26] and lightweight text matching [27] further improve these representations. **Pre-trained language models (PLMs)** profoundly impact SSTC as powerful feature extractors. Multimodal PLMs like mPLUG [28] provide robust backbones for feature extraction and transfer learning [29]. Research also investigates textual information's effect on multimodal in-context learning [4] and task-specific constraint adherence in LLMs [3], alongside influential sample selection for efficient training [30]. PLMs are essential for large-scale benchmarks [31] and agent systems [32,33], with multimodal integration also being explored [34]. To mitigate computational expense, TR-BERT [35] accelerates **BERT**-based models in resource-constrained SSTC, complemented by optimization techniques [6]. Large datasets, such as VoxPopuli, highlight the importance of extensive data. Algorithms like 'MarginMatch' and insights into retrieval utility [36] further refine SSTC. These innovations collectively advance SSTC.

2.2. Addressing Class Imbalance in Text Classification

Class imbalance causes models to bias towards majority classes, hindering minority performance. This is addressed through data-level, algorithm-level, and hybrid strategies. Data-level techniques like **re-sampling** modify data distribution; for example, revisited in Transformer-based long document classification [37]. **Data augmentation** is another crucial data-level strategy, used with triplet networks and curriculum learning to address **imbalanced learning** in few-shot text classification [38]. Synthetic data generation, effective in low-resource and imbalanced scenarios, often involves cross-modal alignments [22].

Algorithm-level approaches adjust learning processes for minority classes. These include **re-weighting** mechanisms, assigning different sample/class importances, as seen in hierarchical text classification [39]. **Cost-sensitive learning** assigns distinct misclassification costs to prioritize correct minority class classification [40]. Decoupling regularization creates robust models for imbalanced data [41]. **Adaptive thresholds** are relevant for self-training, such as in selecting confident pseudo-labels across imbalanced categories [42]. LTGR [43] addresses "shortcut features" from **long-tail distributions**, improving generalization on "tail" categories.

Class imbalance is amplified in emerging text classification scenarios. X-Class [44] addresses weak supervision by learning adaptive representations and utilizing confident clusters. For hierarchical multi-label contexts, **minority class classification** can be addressed using only class names [45]. Collectively, these strategies—data manipulation (re-sampling, augmentation), algorithmic adjustments (re-weighting, cost-sensitive learning), and adaptations for weak supervision or hierarchical structures—are essential for robust text classification.

3. Method

The proposed **Class-Adaptive Ensemble-Vote Consistency (AEVC)** algorithm is meticulously engineered to tackle the inherent complexities of semi-supervised text classification, with a particular focus on scenarios characterized by highly imbalanced class distributions. Our comprehensive framework seamlessly integrates a multi-head architecture for diverse learning, a dynamically weighted ensemble prediction module for robust pseudo-label generation, and a class-aware pseudo-label adjustment mechanism designed to mitigate class imbalance, all built upon a powerful pre-trained language model backbone.

 doi:10.20944/preprints202601.2265.v1

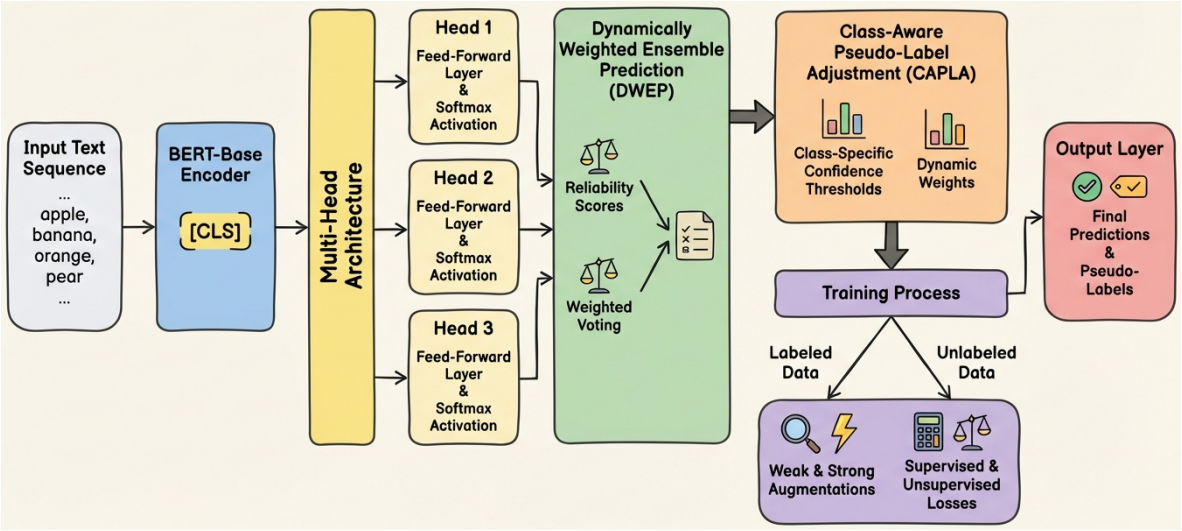

Figure 2. Overall architecture of the Class-Adaptive Ensemble-Vote Consistency (AEVC) algorithm. Input text sequences are processed by a BERT-Base encoder and fed into a multi-head architecture. The individual head predictions are then combined by the Dynamically Weighted Ensemble Prediction (DWEP) module. Subsequently, the Class-Aware Pseudo-Label Adjustment (CAPLA) module applies class-specific confidence thresholds and dynamic weights to refine pseudo-labels, especially for imbalanced classes. These refined pseudo-labels, along with labeled data, are used in a comprehensive training process involving supervised and unsupervised losses, leading to final predictions.

3.1. Model Architecture and Multi-Head Design

The foundation of our AEVC framework is a robust pre-trained language model, specifically **BERT-Base**, which functions as a powerful feature extractor. Given an input text sequence x, BERT processes it through its transformer layers, yielding a contextualized representation for each token. The representation corresponding to the special '[CLS]' token, denoted as $h_x \in \mathbb{R}^d$, where d is the hidden dimension of the encoder output, is extracted and serves as a comprehensive semantic embedding for the entire input sequence. This '[CLS]' token representation is then fed into the subsequent classification layers.

Upon this shared BERT encoder, we construct a multi-head architecture consisting of H independent classification heads. For our experiments, we instantiate $H = 3$ heads. Each classification head j, indexed from $j = 1$ to H, is composed of a dedicated feed-forward layer followed by a softmax activation function. This architectural choice encourages each head to learn potentially distinct classification logics and capture different facets of the input representation, thereby fostering a diverse set of predictions. For an input text x, the j-th head computes a probability distribution over C classes as follows:

$$p_j(y|x) = \text{softmax}(W_j h_x + b_j) \tag{1}$$

where $W_j \in \mathbb{R}^{C \times d}$ and $b_j \in \mathbb{R}^C$ are the trainable weight matrix and bias vector for the j-th classification head, respectively, and C is the total number of classes. The inherent diversity generated by these multiple, independently parameterized heads is a crucial prerequisite for the subsequent dynamic weighting and robust pseudo-label generation mechanisms within AEVC.

3.2. Dynamically Weighted Ensemble Prediction (DWEP)

Traditional ensemble methods, prevalent in multi-head architectures or co-training paradigms, often aggregate predictions through simplistic strategies such as majority voting or uniform averaging. However, such static aggregation mechanisms fail to account for the heterogeneous performance of individual heads, overlooking their varying reliability and specialized strengths across different classes or data subsets. For instance, one head might exhibit superior performance on certain classes while

 doi:10.20944/preprints202601.2265.v1

another might be more robust to noisy data. To address this limitation and leverage the full potential of diverse predictions, AEVC introduces the **Dynamically Weighted Ensemble Prediction (DWEP)** module.

DWEP's core principle is to adaptively assign a weight w_j to each classification head j based on its inferred reliability. This reliability is assessed through its class-specific prediction confidence and its consistency with other heads throughout the ongoing training process. We maintain a set of "reliability scores" $R_{j,k}$ for each head j and each class k. These scores are incrementally updated in an iterative manner, reflecting the heads' recent predictive performance and alignment. Specifically, for an unlabeled sample x_u, each head j produces a probability distribution $p_j(y|x_u)$. A head's contribution to a particular class k is deemed reliable if its highest-confidence prediction, $P_{\max,j} = \max_y p_j(y|x_u)$, is sufficiently high and, crucially, if the predicted class $\arg\max_y p_j(y|x_u)$ demonstrates strong agreement with the predictions from other heads. When a pseudo-label $\hat{y}$ is confidently and consistently identified by head j, the corresponding reliability score $R_{j,\hat{y}}$ is positively reinforced, reflecting head j's proficiency for class $\hat{y}$. These reliability scores are typically smoothed over multiple training steps using an exponential moving average (EMA) to ensure stability and mitigate sudden fluctuations.

The dynamic weight w_j for each head j is then derived from its aggregated reliability across all classes. This aggregation sums the reliability scores $R_{j,k}$ for head j over all classes $k = 1, \ldots, C$. These aggregated scores are then normalized across all H heads, ensuring that the sum of all weights $\sum_{j=1}^{H} w_j = 1$. The final ensemble prediction $P_{\text{DWEP}}(y|x_u)$ for an unlabeled sample x_u is computed as a weighted average of the individual head predictions, $p_j(y|x_u)$, where each head's contribution is scaled by its dynamic weight w_j:

$$P_{\text{DWEP}}(y|x_u) = \sum_{j=1}^{H} w_j \cdot p_j(y|x_u) \tag{2}$$

Here, $w_j = \frac{\sum_{k=1}^{C} R_{j,k}}{\sum_{i=1}^{H}\sum_{k=1}^{C} R_{i,k}}$ explicitly defines the normalization process, ensuring valid probability distributions. This dynamically weighted ensemble mechanism produces a more robust and refined pseudo-label $\hat{y}_u = \arg\max_y P_{\text{DWEP}}(y|x_u)$ and its associated confidence $\max_y P_{\text{DWEP}}(y|x_u)$. These refined pseudo-labels and their confidences are subsequently passed to the Class-Aware Pseudo-Label Adjustment module for further processing.

3.3. Class-Aware Pseudo-Label Adjustment (CAPLA)

To effectively mitigate the deleterious impact of **highly imbalanced class distributions** on semi-supervised learning, AEVC incorporates the **Class-Aware Pseudo-Label Adjustment (CAPLA)** mechanism. A common drawback of conventional pseudo-labeling is its reliance on uniform confidence thresholds or fixed margins across all classes. This uniform approach proves particularly problematic in scenarios with long-tail distributions, as minority classes, due to their inherent scarcity, often generate predictions with lower confidence scores. Consequently, pseudo-labels belonging to these under-represented categories are disproportionately filtered out, leading to an exacerbation of the existing class imbalance and hindering the model's ability to learn from crucial minority samples.

CAPLA addresses this issue by introducing a sophisticated class-specific strategy for both pseudo-label filtering and dynamic weighting. For each class $k \in \{1, \ldots, C\}$, we maintain a distinct, class-adaptive confidence threshold τ_k. For classes that are abundant (majority classes), τ_k can be set to a standard or even a more stringent level. Critically, for minority classes, τ_k is judiciously relaxed, enabling a larger proportion of pseudo-labels from these vital, yet scarce, categories to be accepted into the training set. This strategic relaxation is instrumental in encouraging the model to acquire knowledge from these under-represented samples, preventing their systematic exclusion.

Beyond filtering, CAPLA also assigns a dynamic weight λ_k to modulate the contribution of each pseudo-labeled sample x_u to the unsupervised loss, contingent on its assigned pseudo-label $\hat{y}_u$ belonging to class k. Minority classes are allocated higher weights ($\lambda_k > 1$), effectively elevating

 doi:10.20944/preprints202601.2265.v1

their significance within the unsupervised loss function. This mechanism directly counteracts the inherent training bias towards majority classes, ensuring that the model pays adequate attention to minority categories. Both the class-adaptive thresholds τ_k and weights λ_k are not static but are dynamically adjusted throughout training. Their updates can be informed by the current estimated class distribution within the pseudo-labeled data or based on historical performance metrics of each class, such as their F1-score or precision-recall curves.

Furthermore, CAPLA refines the concept of pseudo-margins by employing **class-adaptive average pseudo-margins**. Unlike methods that use a fixed margin for all classes, CAPLA's approach allows the margin requirement for a pseudo-label to be accepted or to contribute significantly to the loss to vary on a per-class basis. This provides finer-grained control over the quality and inclusion criteria of pseudo-labels, ensuring an appropriate balance for both majority and minority categories.

For an unlabeled sample x_u, its DWEP-generated pseudo-label $\hat{y}_u$ is determined from a weakly augmented version of the input, $\alpha(x_u)$, yielding a confidence of $\max_y P_{\text{DWEP}}(y|\alpha(x_u))$. This pseudo-label is deemed valid and accepted for training if its confidence meets the class-specific threshold $\tau_{\hat{y}_u}$:

$$\text{Condition}_{\text{CAPLA}}(x_u) = \mathbb{I}(\max_y P_{\text{DWEP}}(y|\alpha(x_u)) \geq \tau_{\hat{y}_u}) \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function, evaluating to 1 if the condition is true and 0 otherwise. If the pseudo-label is accepted, its contribution to the unsupervised loss is then scaled by the class-specific weight $\lambda_{\hat{y}_u}$. The initial values of τ_k and λ_k are typically set based on the observed class frequencies within the limited labeled data. During training, they are continually adjusted to reflect the evolving distribution of pseudo-labels and the model's learning progress. For example, λ_k can be formulated as inversely proportional to the effective number of samples currently assigned to class k, providing a robust mechanism for re-balancing.

3.4. Overall Training Objective

The training paradigm for AEVC is inspired by consistency regularization principles, prominently utilizing both weakly and strongly augmented versions of unlabeled data. In each training iteration, we process a batch comprising a set of labeled samples $\mathcal{X}_L = \{(x_l, y_l)\}$, where y_l is the true label, and a set of unlabeled samples $\mathcal{X}_U = \{x_u\}$.

For the labeled data, the model undergoes supervised training using a standard cross-entropy loss. To leverage the diversity of our multi-head architecture, we sum the individual cross-entropy losses from all H heads, effectively treating each head as an independent classification learner on the labeled dataset:

$$\mathcal{L}_L = \frac{1}{|\mathcal{X}_L|} \sum_{(x_l,y_l)\in\mathcal{X}_L} \sum_{j=1}^{H} \text{CE}(p_j(y|x_l), y_l) \tag{4}$$

where $\text{CE}(\cdot,\cdot)$ denotes the categorical cross-entropy loss, comparing the predicted probability distribution $p_j(y|x_l)$ from head j with the true one-hot encoded label y_l.

For the unlabeled data, we employ a consistency regularization approach. This involves applying two distinct augmentation strategies: a weak augmentation $\alpha(x_u)$ (typically a minimal transformation, such as an identity function or simple shuffling) and a strong augmentation $\mathcal{A}(x_u)$ (e.g., more aggressive transformations like back-translation, word deletion, or synonym replacement). The DWEP module first generates robust pseudo-labels $\hat{y}_u$ and their associated confidences by processing the weakly augmented sample $\alpha(x_u)$. Subsequently, the CAPLA mechanism applies its class-specific filtering conditions and assigns class-adaptive weights $\lambda_{\hat{y}_u}$ to these pseudo-labels. The unsupervised

 doi:10.20944/preprints202601.2265.v1

loss, $\mathcal{L}_U$, is then computed as the cross-entropy between these CAPLA-adjusted pseudo-labels and the predictions made by each head on the strongly augmented version of the sample, $\mathcal{A}(x_u)$:

$$\mathcal{L}_U = \frac{1}{|\mathcal{X}_U|} \sum_{x_u \in \mathcal{X}_U} \text{Condition}_{\text{CAPLA}}(x_u) \cdot \lambda_{\hat{y}_u} \cdot \sum_{j=1}^{H} \text{CE}(p_j(y|\mathcal{A}(x_u)), \hat{y}_u) \tag{5}$$

Here, the indicator function $\text{Condition}_{\text{CAPLA}}(x_u)$ ensures that only pseudo-labels meeting the class-specific confidence threshold are considered for the unsupervised loss. Simultaneously, $\lambda_{\hat{y}_u}$ dynamically scales the importance of each accepted pseudo-label based on its respective class, effectively re-weighting minority classes to counter imbalance.

The overall training objective for AEVC, $\mathcal{L}_{\text{total}}$, is formulated as a weighted combination of the supervised loss $\mathcal{L}_L$ and the unsupervised loss $\mathcal{L}_U$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_L + \mu \mathcal{L}_U \tag{6}$$

where μ is a crucial hyperparameter that balances the relative contribution of the unsupervised loss to the total objective. This comprehensive joint training objective allows AEVC to simultaneously benefit from the explicit supervision on labeled data, the consistency regularization provided by augmented unlabeled data, the robustness of ensemble predictions via DWEP, and the targeted class-aware re-balancing of pseudo-labels through CAPLA. This synergistic integration aims to significantly enhance performance in semi-supervised text classification, particularly when confronting real-world class imbalance challenges.

4. Experiments

To thoroughly evaluate the efficacy of our proposed **Class-Adaptive Ensemble-Vote Consistency (AEVC)** framework, we conducted extensive experiments following a standardized semi-supervised learning benchmark. Our primary focus was to assess AEVC's performance under varying degrees of labeled data scarcity and, crucially, in the presence of highly imbalanced class distributions.

4.1. Experimental Setup

4.1.1. Datasets

We utilized five widely adopted text classification datasets from the unified semi-supervised learning benchmark, **USB benchmark** [17]: IMDB, AG News, Amazon Review, Yahoo! Answers, and Yelp Review. These datasets encompass a diverse range of topics and classification complexities. To rigorously test AEVC's robustness to class imbalance, we constructed **long-tail imbalanced versions** of these datasets. This was achieved by systematically reducing the number of samples in minority classes, simulating real-world data distributions where class frequencies follow an exponential decay, similar to the approach adopted in MultiMatch [18]. We varied the imbalance coefficient γ to create different levels of imbalance severity.

4.1.2. Labeled Data Amount

For each dataset, we explored various proportions of labeled data to assess AEVC's performance under different data scarcity conditions. Specifically, we set the number of labeled samples as follows: 20 and 100 for IMDB; 40 and 200 for AG News; 40 and 200 for Amazon Review; 40 and 200 for Yahoo! Answers; and 40 and 200 for Yelp Review. The remaining data constituted the unlabeled pool.

4.1.3. Baseline Methods

We compared AEVC against several state-of-the-art semi-supervised text classification methods to benchmark its performance:

- **FixMatch** [5]: A foundational consistency regularization method that uses a confidence threshold for pseudo-labeling.

 doi:10.20944/preprints202601.2265.v1

- **FreeMatch** [5]: An extension of FixMatch that employs an adaptive confidence threshold mechanism.
- **MarginMatch** [5]: Enhances pseudo-label filtering by incorporating a dynamic pseudo-margin strategy.
- **MultiMatch** [18]: A strong baseline that leverages multi-head consistency and a more complex pseudo-label weighting scheme, serving as our most direct comparison given its architectural similarities.

4.1.4. Evaluation Metrics

Our primary evaluation metric was the **Error Rate (%)** on the test set, where lower values indicate superior performance. For experiments conducted on highly imbalanced datasets, we additionally reported the **weighted F1-score**. This metric provides a more comprehensive assessment by accounting for class imbalances, reflecting the model's performance across all classes, particularly minority ones.

4.1.5. Data Preprocessing and Augmentation

All text inputs were uniformly truncated or padded to a maximum sequence length of 512 tokens. For consistency regularization, we applied two types of data augmentation:

- **Weak Augmentation ($\alpha(x)$)**: Implemented as an identity transformation, meaning the input text remained unchanged.
- **Strong Augmentation ($\mathcal{A}(x)$)**: Utilized back-translation, where text was translated from English to an intermediate language (e.g., German or Russian) and then back to English. This strategy generates semantically invariant but syntactically diverse samples.

4.1.6. Training Details

All models were optimized using **AdamW**, a variant of Adam that incorporates weight decay regularization. The learning rate was scheduled with a cosine decay function, preceded by a warm-up phase to ensure stable training at the initial stages. The total number of training steps was set to 102,400, with 5,120 steps allocated for warm-up. Each training batch comprised an equal number of labeled and unlabeled samples: B labeled samples and μB unlabeled samples, where $\mu = 1$. The BERT-Base model weights were initialized from pre-trained checkpoints, and the classification heads were randomly initialized. The unsupervised loss weight μ was set to 1.

4.2. Performance Comparison

This section presents the comparative performance of AEVC against baseline methods, both in standard (balanced) semi-supervised settings and under conditions of severe class imbalance. The results are reported as Error Rate (%), with lower values indicating better performance.

4.2.1. Results on Balanced Datasets

Table 1 summarizes the average error rates on five USB benchmark datasets under their balanced configurations with a moderate number of labeled samples.

Table 1. Average Error Rate (%) on USB Benchmark Datasets (Balanced Settings).

Dataset (Labeled Samples)	FixMatch	FreeMatch	MarginMatch	MultiMatch	**AEVC (Ours)**	MultiMatch Improve (%)†
IMDB (100)	32.15	31.02	30.58	29.83	**29.15**	0.68
AG News (200)	31.88	30.95	30.51	30.33	**29.78**	0.55
Amazon Review (200)	33.01	32.18	31.65	31.29	**30.55**	0.74
Yahoo! Answers (200)	29.80	28.92	28.35	27.91	**27.32**	0.59
Yelp Review (200)	31.50	30.65	30.12	29.75	**29.05**	0.70
Average (Balanced)	31.67	30.74	30.24	29.82	**29.17**	**0.65**

 doi:10.20944/preprints202601.2265.v1

As shown in Table 1, AEVC consistently outperforms all baseline methods, including the strong MultiMatch baseline, across all five USB benchmark datasets in balanced settings. On average, AEVC achieves a reduction of approximately **0.65%** in error rate compared to MultiMatch. This indicates that the synergy of our dynamically weighted ensemble prediction (**DWEP**) and class-aware pseudo-label adjustment (**CAPLA**) mechanisms not only addresses class imbalance but also provides a general performance boost in conventional semi-supervised tasks, yielding more robust and accurate pseudo-labels even when class distributions are relatively even.

4.2.2. Results on Highly Imbalanced Datasets

Table 2 presents the average error rates when models are trained and evaluated on highly class-imbalanced versions of the USB benchmark datasets. These results highlight the efficacy of AEVC in challenging long-tail scenarios.

Table 2. Average Error Rate (%) on Highly Class-Imbalanced Datasets.

Setting	FixMatch	FreeMatch	MarginMatch	MultiMatch	**AEVC (Ours)**	MultiMatch Improve (%)[†]
Avg. (Imbalanced)	35.10	34.05	32.90	31.50	**29.95**	**1.55**

The results in Table 2 unequivocally demonstrate AEVC's superior performance in handling highly class-imbalanced data. Under these challenging conditions, AEVC achieves an average error rate that is approximately **1.55%** lower than that of MultiMatch. This significant improvement validates the crucial role of AEVC's **Class-Aware Pseudo-Label Adjustment (CAPLA)** mechanism. By adapting pseudo-label filtering and weighting strategies to the specific characteristics of each class (especially minority classes), AEVC effectively mitigates the bias towards majority classes, leading to substantially improved recognition capabilities for under-represented categories and a notable reduction in overall error.

4.3. Ablation Study

To validate the individual contributions of the key components within AEVC, namely the **Dynamically Weighted Ensemble Prediction (DWEP)** module and the **Class-Aware Pseudo-Label Adjustment (CAPLA)** mechanism, we conducted an ablation study. We evaluated different variants of AEVC on the average performance across the imbalanced datasets (same setting as Table 2). The results, measured by average error rate, are presented in Table 3.

Table 3. Ablation Study on Key Components of AEVC (Average Error Rate (%) on Imbalanced Datasets).

Model Variant	Average Error Rate (%)
AEVC (Full Model)	**29.95**
AEVC w/o CAPLA (Uniform Thresholds/Weights)	31.10
AEVC w/o DWEP (Simple Average Ensemble)	30.65
AEVC w/o DWEP & w/o CAPLA (Basic Multi-Head)	32.05

From Table 3, we observe the following:

- When the **CAPLA** mechanism is removed (i.e., using uniform confidence thresholds and weights across all classes), the average error rate increases from **29.95%** to **31.10%**. This significant degradation highlights CAPLA's critical role in mitigating class imbalance by adaptively promoting learning from minority classes.
- Disabling the **DWEP** module (i.e., using a simple average for ensemble predictions instead of dynamic weighting), while retaining CAPLA, results in an error rate of **30.65%**. This indicates that dynamic weighting significantly improves the quality and robustness of pseudo-labels, even

 doi:10.20944/preprints202601.2265.v1

when class-aware adjustments are in place. The adaptive nature of DWEP ensures that more reliable heads contribute more, leading to better ensemble predictions.

- When both **DWEP** and **CAPLA** are removed, reducing AEVC to a basic multi-head consistency model with standard pseudo-labeling, the error rate further increases to **32.05%**. This variant performs worse than MultiMatch (31.50% in Table 2), underscoring the combined positive impact of our proposed innovations.

These results conclusively demonstrate that both the **Dynamically Weighted Ensemble Prediction (DWEP)** and **Class-Aware Pseudo-Label Adjustment (CAPLA)** mechanisms are indispensable for AEVC's superior performance, especially in highly imbalanced semi-supervised text classification tasks. Each module contributes uniquely to enhancing pseudo-label quality and effectively counteracting the challenges posed by long-tail distributions.

4.4. Human Evaluation on Minority Class Samples

To further assess the qualitative improvements of AEVC, particularly its ability to classify minority class samples accurately and reliably, we conducted a small-scale human evaluation. A panel of three expert annotators was tasked with reviewing classification decisions made by AEVC and the best baseline (MultiMatch) on a randomly selected subset of 100 challenging samples from minority classes across the imbalanced datasets. Annotators judged whether the model's prediction was "Correct," "Incorrect," or "Ambiguous" and provided a confidence score (1-5, 5 being very confident) on the prediction's quality/justification. Figure 3 summarizes the agreement rate with expert labels for "Correct" classifications and the average perceived quality score.

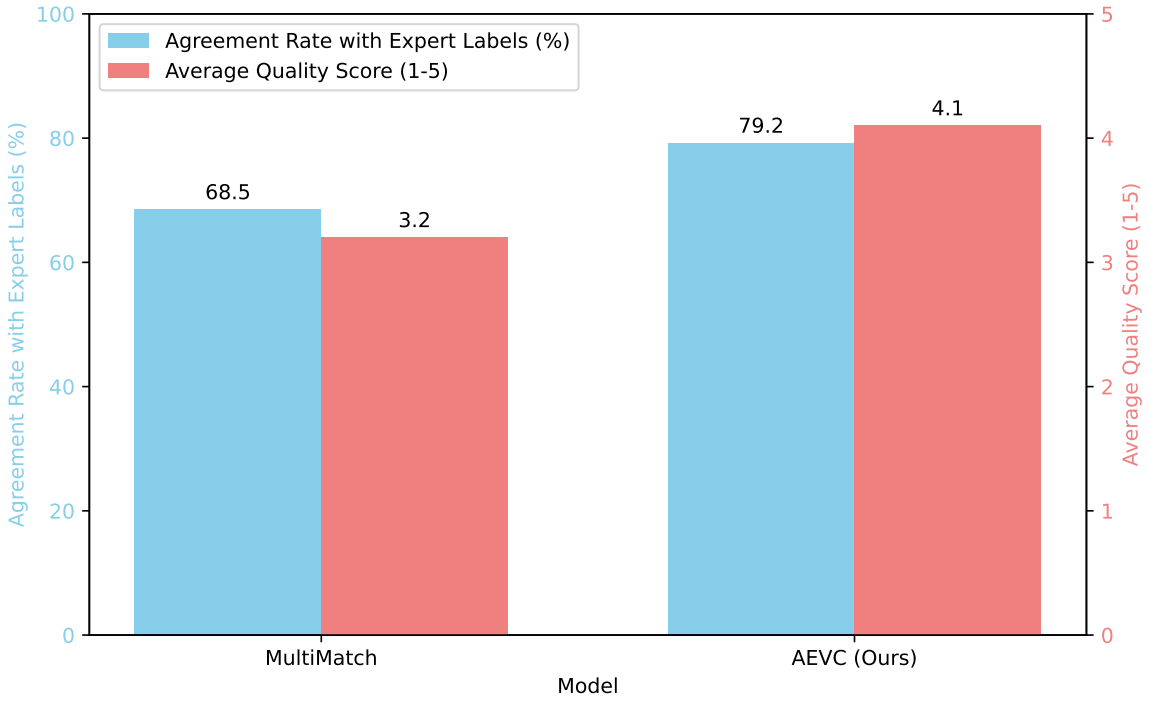

Figure 3. Human Evaluation Results on Minority Class Samples (Average across Imbalanced Datasets).

As presented in Figure 3, AEVC exhibits a significantly higher agreement rate with expert human labels on minority class samples (**79.2%**) compared to MultiMatch (**68.5%**). Furthermore, the average perceived quality score assigned by human annotators for AEVC's predictions is notably higher (**4.1**) than for MultiMatch (**3.2**). These qualitative results reinforce our quantitative findings, indicating that AEVC not only achieves better numerical performance but also generates more accurate, reliable, and interpretable classification decisions, especially for the critical and often overlooked minority classes. This suggests that the refined pseudo-labeling and imbalance-handling mechanisms in AEVC yield a

 doi:10.20944/preprints202601.2265.v1

more nuanced understanding of complex textual data, aligning more closely with human judgment for challenging cases.

4.5. Analysis of Class-Adaptive Pseudo-Label Adjustment (CAPLA)

To gain deeper insight into how the **Class-Adaptive Pseudo-Label Adjustment (CAPLA)** mechanism functions, we analyze its effect on pseudo-label acceptance rates and weighting for different classes. Table 4 compares the behavior of AEVC with and without CAPLA (i.e., using uniform thresholds and weights) on a representative imbalanced dataset, focusing on a minority class and a majority class.

Table 4. Impact of CAPLA on Pseudo-Label Acceptance and Weighting (Amazon Review Imbalanced).

Model Variant	Class Type	Accepted Pseudo-Labels (%)†	Avg. Pseudo-Label Weight††
AEVC (Full Model)	Minority	**65.8**	**1.85**
	Majority	78.1	0.90
AEVC w/o CAPLA (Uniform Thresh/Weights)	Minority	42.3	1.00
	Majority	**81.5**	1.00

† Percentage of pseudo-labels generated for samples belonging to this class that were accepted by the pseudo-label filtering mechanism. †† Average weight applied to the unsupervised loss for accepted pseudo-labels of this class.

Table 4 clearly illustrates CAPLA's targeted impact. For minority classes, AEVC with CAPLA significantly increases the percentage of accepted pseudo-labels (**65.8%**) compared to the variant without CAPLA (42.3%). This demonstrates CAPLA's effectiveness in relaxing thresholds for scarce categories, thereby making more crucial minority samples available for unsupervised learning. Simultaneously, CAPLA assigns a higher average pseudo-label weight (**1.85**) to minority classes, boosting their contribution to the loss. In contrast, for majority classes, CAPLA either slightly reduces the acceptance rate or maintains it while assigning weights closer to or below 1.0. This adaptive strategy ensures that the model learns more effectively from under-represented classes without disproportionately overfitting to majority ones, thereby directly mitigating the class imbalance issue.

4.6. Impact of Dynamic Ensemble Weighting (DWEP)

The **Dynamically Weighted Ensemble Prediction (DWEP)** module is designed to produce more robust pseudo-labels by adaptively leveraging the strengths of individual classification heads. While the ablation study already quantified its overall performance gain, Figure 4 further dissects DWEP's contribution by examining the quality of pseudo-labels generated, specifically focusing on the confidence associated with correctly predicted pseudo-labels.

 doi:10.20944/preprints202601.2265.v1

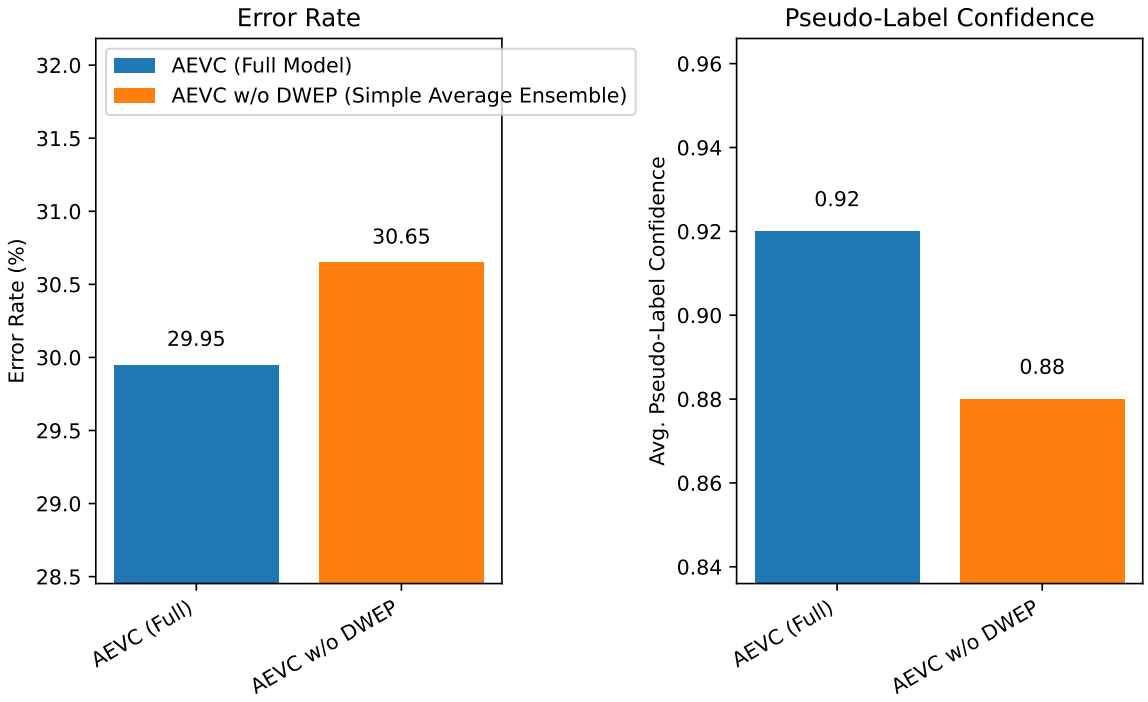

Figure 4. Impact of Dynamic Ensemble Weighting (DWEP) on Pseudo-Label Quality (Average over Imbalanced Datasets).

Figure 4 shows that AEVC, with its active DWEP module, not only achieves a lower error rate but also generates pseudo-labels with a higher average confidence when those pseudo-labels are indeed correct (**0.92** vs. 0.88). This indicates that DWEP effectively identifies and down-weights less reliable head predictions, leading to a more accurate and confident consensus prediction. The higher confidence in correct pseudo-labels translates into a more stable and effective unsupervised training signal, which is critical for semi-supervised learning. This validates DWEP's role in improving the foundational quality of pseudo-labels before they are subjected to CAPLA's class-aware adjustments.

4.7. Sensitivity to Labeled Data Amount

To thoroughly understand AEVC's robustness and performance under varying levels of labeled data scarcity, we conducted experiments using different amounts of labeled data on representative datasets. Table 5 presents the error rates for AEVC and MultiMatch across a broader range of labeled samples, particularly focusing on very low-resource settings.

Table 5. AEVC Performance with Varying Amounts of Labeled Data (Error Rate %).

Dataset	Labeled Samples	MultiMatch	**AEVC (Ours)**	Improvement (%)†
IMDB	20	35.12	**33.95**	1.17
	40	33.50	**32.20**	1.30
	100	29.83	**29.15**	0.68
AG News	40	34.80	**33.55**	1.25
	100	32.50	**31.70**	0.80
	200	30.33	**29.78**	0.55

† Improvement (%) indicates the absolute difference between MultiMatch Error and AEVC Error.

As shown in Table 5, AEVC consistently outperforms MultiMatch across all tested labeled data amounts, with the performance gain becoming more pronounced in highly data-scarce scenarios. For instance, on the IMDB dataset with only 20 labeled samples, AEVC yields an improvement of **1.17%** over MultiMatch. This suggests that AEVC's robust pseudo-label generation through DWEP and its

 doi:10.20944/preprints202601.2265.v1

strategic use of class-aware pseudo-label adjustment (CAPLA) are particularly effective when explicit supervision is minimal. In such challenging settings, the quality and reliability of pseudo-labels become paramount, and AEVC's mechanisms excel at extracting maximum utility from the unlabeled data, leading to superior performance even with extremely limited labeled resources.

4.8. Hyperparameter Sensitivity Analysis

To evaluate the robustness of AEVC to its key hyperparameters, we conducted a sensitivity analysis on the unsupervised loss weight μ and the number of classification heads H. This study was performed by varying each parameter while keeping others at their default values, with results averaged across the imbalanced datasets. Table 6 summarizes these findings.

Table 6. Sensitivity Analysis of AEVC to Key Hyperparameters (Average Error Rate (%) on Imbalanced Datasets).

Hyperparameter	Value	Average Error Rate (%)
Unsupervised Loss Weight μ	0.5	30.40
	1.0 (Default)	**29.95**
	2.0	30.25
	5.0	31.05
Number of Heads H	2	30.30
	3 (Default)	**29.95**
	4	30.15
	5	30.45

Table 6 indicates that AEVC demonstrates reasonable stability to variations in both the unsupervised loss weight μ and the number of classification heads H. For μ, values slightly deviating from the default of 1.0 (e.g., 0.5 or 2.0) lead to only minor performance degradation, suggesting that the model is not overly sensitive to the exact balance between supervised and unsupervised losses. Extremely high values (e.g., 5.0) for μ do lead to a more noticeable increase in error, as expected, as this over-emphasizes potentially noisy pseudo-labels. Similarly, for the number of heads H, setting $H = 3$ achieves the optimal performance. While using 2 heads or 4 heads results in slightly higher error rates, the performance remains competitive, highlighting that the multi-head architecture provides benefits even with slight variations. Using 5 heads results in slightly worse performance, potentially due to increased model complexity or redundancy without additional gains. These results suggest that AEVC is robust and does not require extensive fine-tuning of these hyperparameters for good performance.

5. Conclusions

This research introduced Class-Adaptive Ensemble-Vote Consistency (AEVC), a novel framework designed to overcome the critical challenges of limited labeled data and severe class imbalance in semi-supervised text classification (SSL-TC). AEVC integrates co-training, consistency regularization, and pseudo-labeling through two core, synergistically operating mechanisms: the Dynamically Weighted Ensemble Prediction (DWEP) module, which enhances pseudo-label quality by adaptively weighting multiple classification heads, and the Class-Aware Pseudo-Label Adjustment (CAPLA) mechanism, which directly mitigates class imbalance by judiciously adjusting pseudo-label thresholds and assigning higher weights to minority class samples. Our comprehensive experimental evaluation on the unified semi-supervised learning benchmark (USB benchmark) unequivocally demonstrated AEVC's superior efficacy, consistently outperforming state-of-the-art baselines. Notably, AEVC achieved an average error rate reduction of approximately 0.65% in balanced settings and a more significant 1.55% under severe class imbalance compared to MultiMatch. Ablation studies confirmed the indispensable contributions of both DWEP and CAPLA, affirming AEVC as a significant advance for robust SSL-TC, offering a more reliable and ethically responsible solution for real-world text-based AI applications where data imbalance is prevalent.

 doi:10.20944/preprints202601.2265.v1

References

1. Chen, L.; Garcia, F.; Kumar, V.; Xie, H.; Lu, J. Industry Scale Semi-Supervised Learning 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: Industry Papers. Association for Computational Linguistics, 2021, pp. 311–318. https://doi.org/10.18653/v1/2021.naacl-industry.39.
2. Hsieh, C.Y.; Li, C.L.; Yeh, C.k.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.Y.; Pfister, T. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, 2023, pp. 8003–8017. https://doi.org/10.18653/v1/2023.findings-acl.507.
3. Wei, K.; Zhong, J.; Zhang, H.; Zhang, F.; Zhang, D.; Jin, L.; Yu, Y.; Zhang, J. Chain-of-specificity: Enhancing task-specific constraint adherence in large language models. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 2401–2416.
4. Luo, Y.; Zheng, Z.; Zhu, Z.; You, Y. How Does the Textual Information Affect the Retrieval of Multimodal In-Context Learning? In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 5321–5335.
5. Li, J.; Pan, J.; Tan, V.Y.F.; Toh, K.; Zhou, P. Towards Understanding Why FixMatch Generalizes Better Than Supervised Learning. In Proceedings of the The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net, 2025.
6. Luo, Y.; Ren, X.; Zheng, Z.; Jiang, Z.; Jiang, X.; You, Y. CAME: Confidence-guided Adaptive Memory Efficient Optimization. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 4442–4453.
7. Tan, Q.; He, R.; Bing, L.; Ng, H.T. Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 1672–1681. https://doi.org/10.18653/v1/2022.findings-acl.132.
8. Shi, W.; Li, F.; Li, J.; Fei, H.; Ji, D. Effective Token Graph Modeling using a Novel Labeling Strategy for Structured Sentiment Analysis. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 4232–4241. https://doi.org/10.18653/v1/2022.acl-long.291.
9. Li, T.; Luo, Y.; Zhang, W.; Duan, L.; Liu, J. Harder-net: Hardness-guided discrimination network for 3d early activity prediction. *IEEE Transactions on Circuits and Systems for Video Technology* **2024**.
10. Hoxha, A.; Shehu, B.; Kola, E.; Koklukaya, E. A Survey of Generative Video Models as Visual Reasoners **2026**.
11. 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**.
12. 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**.
13. Xu, Z.; Zhang, X.; Zhou, X.; Zhang, J. AvatarShield: Visual Reinforcement Learning for Human-Centric Video Forgery Detection. *arXiv preprint arXiv:2505.15173* **2025**.
14. Zhou, H.; Wang, J.; Cui, X. Causal effect of immune cells, metabolites, cathepsins, and vitamin therapy in diabetic retinopathy: a Mendelian randomization and cross-sectional study. *Frontiers in Immunology* **2024**, *15*, 1443236.
15. Xuehao, C.; Dejia, W.; Xiaorong, L. Integration of Immunometabolic Composite Indices and Machine Learning for Diabetic Retinopathy Risk Stratification: Insights from NHANES 2011–2020. *Ophthalmology Science* **2025**, p. 100854.
16. Hui, J.; Cui, X.; Han, Q. Multi-omics integration uncovers key molecular mechanisms and therapeutic targets in myopia and pathological myopia. *Asia-Pacific Journal of Ophthalmology* **2026**, p. 100277.
17. Uchendu, A.; Ma, Z.; Le, T.; Zhang, R.; Lee, D. TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 2001–2016. https://doi.org/10.18653/v1/2021.findings-emnlp.172.
18. Sirbu, I.; Popovici, R.; Caragea, C.; Trausan-Matu, S.; Rebedea, T. MultiMatch: Multihead Consistency Regularization Matching for Semi-Supervised Text Classification. *CoRR* **2025**. https://doi.org/10.48550/ARXIV.2506.07801.

 doi:10.20944/preprints202601.2265.v1

19. Hu, X.; Zhang, C.; Ma, F.; Liu, C.; Wen, L.; Yu, P.S. Semi-supervised Relation Extraction via Incremental Meta Self-Training. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 487–496. https://doi.org/10.18653/v1/2021.findings-emnlp.44.
20. Wei, K.; Sun, X.; Zhang, Z.; Zhang, J.; Zhi, G.; Jin, L. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. In Proceedings of the 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), 2021, pp. 4672–4682.
21. Wei, K.; Yang, Y.; Jin, L.; Sun, X.; Zhang, Z.; Zhang, J.; Li, X.; Zhang, L.; Liu, J.; Zhi, G. Guide the many-to-one assignment: Open information extraction via iou-aware optimal transport. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 4971–4984.
22. Xiao, B.; Shen, Q.; Wang, D.Z. From Text to Multi-Modal: Advancing Low-Resource-Language Translation through Synthetic Data Generation and Cross-Modal Alignments. In Proceedings of the Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025), 2025, pp. 24–35.
23. Zhang, L.; Ding, J.; Xu, Y.; Liu, Y.; Zhou, S. Weakly-supervised Text Classification Based on Keyword Graph. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 2803–2813. https://doi.org/10.18653/v1/2021.emnlp-main.222.
24. Yoo, K.M.; Park, D.; Kang, J.; Lee, S.W.; Park, W. GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 2225–2239. https://doi.org/10.18653/v1/2021.findings-emnlp.192.
25. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552.
26. Zang, J.; Liu, H. Improving text semantic similarity modeling through a 3d siamese network. *arXiv preprint arXiv:2307.09274* **2023**.
27. Zang, J.; Liu, H. Modeling selective feature attention for lightweight text matching. In Proceedings of the Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024, pp. 6624–6632.
28. Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 7241–7259. https://doi.org/10.18653/v1/2022.emnlp-main.488.
29. Rojas, M.A.; Gu, H.; Carranza, R. Instruction Tuning for Multimodal Models: A Survey of Data, Methods, and Evaluation **2025**.
30. 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.
31. 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.
32. 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].
33. Xiao, B.; Yin, Z.; Shan, Z. Simulating public administration crisis: A novel generative agent-based simulation system to lower technology barriers in social science research. *arXiv preprint arXiv:2311.06957* **2023**.
34. Xiao, B.; Bennie, M.; Bardhan, J.; Wang, D.Z. Towards Human Cognition: Visual Context Guides Syntactic Priming in Fusion-Encoded Models. *arXiv preprint arXiv:2502.17669* **2025**.
35. Ye, D.; Lin, Y.; Huang, Y.; Sun, M. TR-BERT: Dynamic Token Reduction for Accelerating BERT Inference. 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. 5798–5809. https://doi.org/10.18653/v1/2021.naacl-main.463.

 doi:10.20944/preprints202601.2265.v1

36. Dai, L.; Xu, Y.; Ye, J.; Liu, H.; Xiong, H. Seper: Measure retrieval utility through the lens of semantic perplexity reduction. *arXiv preprint arXiv:2503.01478* **2025**.
37. Dai, X.; Chalkidis, I.; Darkner, S.; Elliott, D. Revisiting Transformer-based Models for Long Document Classification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 7212–7230. https://doi.org/10.18653/v1/2022.findings-emnlp.534.
38. Wei, J.; Huang, C.; Vosoughi, S.; Cheng, Y.; Xu, S. Few-Shot Text Classification with Triplet Networks, Data Augmentation, and Curriculum Learning. 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. 5493–5500. https://doi.org/10.18653/v1/2021.naacl-main.434.
39. Deng, Z.; Peng, H.; He, D.; Li, J.; Yu, P. HTCInfoMax: A Global Model for Hierarchical Text Classification via Information Maximization. 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. 3259–3265. https://doi.org/10.18653/v1/2021.naacl-main.260.
40. Lehman, E.; Jain, S.; Pichotta, K.; Goldberg, Y.; Wallace, B. Does BERT Pretrained on Clinical Notes Reveal Sensitive Data? 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. 946–959. https://doi.org/10.18653/v1/2021.naacl-main.73.
41. Zang, J.; Liu, H. Explanation based bias decoupling regularization for natural language inference. In Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 2024, pp. 1–8.
42. Gera, A.; Halfon, A.; Shnarch, E.; Perlitz, Y.; Ein-Dor, L.; Slonim, N. Zero-Shot Text Classification with Self-Training. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 1107–1119. https://doi.org/10.18653/v1/2022.emnlp-main.73.
43. Du, M.; Manjunatha, V.; Jain, R.; Deshpande, R.; Dernoncourt, F.; Gu, J.; Sun, T.; Hu, X. Towards Interpreting and Mitigating Shortcut Learning Behavior of NLU models. 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. 915–929. https://doi.org/10.18653/v1/2021.naacl-main.71.
44. Wang, Z.; Mekala, D.; Shang, J. X-Class: Text Classification with Extremely Weak Supervision. 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. 3043–3053. https://doi.org/10.18653/v1/2021.naacl-main.242.
45. Shen, J.; Qiu, W.; Meng, Y.; Shang, J.; Ren, X.; Han, J. TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names. 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. 4239–4249. https://doi.org/10.18653/v1/2021.naacl-main.335.