多重匹配：用于半监督文本分类的多头一致性正则化匹配

Iustin Sirbu[1], Robert-Adrian Popovici[1],
Cornelia Caragea[2], Stefan Trausan-Matu[1], Traian Rebedea[1,3]

[1]National University of Science and Technology POLITEHNICA Bucharest,
[2]University of Illinois Chicago,
[3]NVIDIA

Correspondence: iustin.sirbu@upb.ro

Abstract

我们介绍了一种多重匹配算法，即一种新的半监督学习（SSL）算法，它结合了协同训练、一致性正则化和伪标签的范式。多重匹配算法的核心是一个三重伪标签加权模块，专为三个关键目的而设计：根据头部一致性和模型置信度选择和过滤伪标签，并根据感知的分类难度对其进行加权。该新模块增强并统一了三种现有技术——来自多头协同训练的头部一致性、来自自由匹配的自适应阈值，以及来自边缘匹配的平均伪边距——从而形成一种整体方法，提高了 SSL 环境中的鲁棒性和性能。基准数据集的实验结果突出显示了多重匹配算法的优异表现，在来自 5 个自然语言处理数据集的 10 种设置中，有 9 种达到了最先进的结果，并在 19 种方法中根据 Friedman 测试排名第一。此外，多重匹配在高度不平衡的设置中表现出异常强的鲁棒性，性能超过第二优方案 3.26 %——数据不平衡是许多文本分类任务的关键因素。我们将公开我们的代码。

1 介绍

深度学习模型通常需要大量的标记数据来在各种任务中实现高性能 (Vaswani, 2017) 。此外，数据标注的过程既昂贵又耗时，经常在现实应用中构成显著的瓶颈。为减轻这一挑战，半监督学习方法（SSL）作为一种有前途的替代方案出现。这些方法通过利用少量标记示例和大量未标记语料，大大减少了人工标注所需的时间和资源，同时仍能保持竞争性的模型性能 (Berthelot et al., 2019b) 。

近年来，半监督学习主要被两种范式所主导。第一种是结合一致性正则化的伪标签，它通过对同一样本的不同增强保持预测不变性，利用对弱增强数据的高置信度预测为强增强的对应样本生成伪标签。第二种是协同训练，它通过并行训练两个网络来减轻误差积累，每个网络为另一个提供伪标签。

这两种范式都取得了显著的进展。在一致性正则化的伪标签中，研究主要集中在改进伪标签过滤，其中 FreeMatch (Wang et al., 2023) 引入了按类别自适应阈值，MarginMatch (Sosea and Caragea, 2023) 则加入了过去周期的预测来提高可靠性。在互培训范式中，Multihead Co-training (Chen et al., 2021) 利用多头架构减少了对多个网络的需求，而 JointMatch (Zou and Caragea, 2023) 强调了模型产生不同预测的样本的重要性。然而，有限的研究探索了这些范式的整合，以充分利用它们的互补优势。

为了解决这个差距，我们引入了 MultiMatch，这是一种新颖的算法，包含了一个三重伪标签加权模块（PLWM）。该模块结合了多头协同训练中的投票机制和一种自适应的按类别过滤策略，该策略利用了历史和当前预测。具体来说，它结合了 Margin Match 引入的平均伪边距（APM），根据历史预测来细化伪标签过滤，同时也利用了 Free Match 的自适应过滤机制，该机制基于模型的当前预测动态调整阈值。此外，我们定义了一种“无用”、“有用 & 难”和“有用 & 易”示例的分类法，并基于此提出了一种加权机制。

我们在统一的 SSL 基准——USB (Wang et al., 2022) 上的五个文本分类数据集上评估了 MultiMatch，与 18 个基线进行比较。为了模拟在现实世界中文本分类任务中常遇到的类别不平衡挑战，我们还构建了这五个数据集的高度不平衡版本。我们的算法始终优于所有基线，在平衡和不平衡设置下，根据 Friedman 检验和所有数据集的平均误差均排名第一。

总结来说，我们的主要贡献是：1）我们引入了 MultiMatch，这是一种新颖的 SSL 算法，可以增强和结合不同 SSL 范式的优势；2）我们在五个广泛用于文本分类的数据集上获得了最先进的性能；3）我们展示了 MultiMatch 在非平衡类分布中的鲁棒性，在高度不平衡的环境中取得了最先进的结果；4）我们提供了全面的消融研究和分析。

在引入 FixMatch (Sohn et al., 2020) 之后，半监督学习中广泛采用的范式是将一致性正则化和伪标签相结合。一致性正则化 (Sajjadi et al., 2016) 基于这样的原则：模型的预测应在输

入样本上施加合理扰动时保持不变。伪标签 (McLachlan, 1975) 利用模型在未标记数据上高置信度的预测，将其视为训练过程中纳入的伪标签。

早期的方法，如 FixMatch (Sohn et al., 2020) 和 UDA (Xie et al., 2020a)，利用预先定义的模型置信度阈值来确定何时可以将预测转换为伪标签。然而，这种方法有两个显著的局限性。首先，它没有考虑不同类别之间难度的差异，导致模型为较容易的类别生成更多的伪标签，从而偏向这些类别 (Zhang et al., 2022; Wang et al., 2023)。其次，模型的置信度在训练的初始阶段往往较低，并随着训练的进展而提高。因此，这种方法开始时只生成少量伪标签实例，然后逐渐为几乎所有样本生成高置信度的伪标签，这可能导致错误累积和泛化能力低下 (Arazo et al., 2020)。

这些限制在后续的工作中通过各种方式得到了改进。例如，FlexMatch (Zhang et al., 2022) 和 FreeMatch (Wang et al., 2023) 提出了根据不同类别的估计学习状态自适应地调整类别阈值。MarginMatch (Sosea and Caragea, 2023) 引入了平均伪边际（APM）评分，该评分跟踪每个样本的历史预测，并提供了更可靠的置信估计。其他工作采用协同训练 (Blum and Mitchell, 1998) 框架。Qiao et al. (2018) 旨在通过使用两个同时训练的不同网络来缓解错误积累，每个网络为另一个生成伪标签。JointMatch (Zou and Caragea, 2023) 结合了类似于 FreeMatch 的自适应局部阈值调整，并引入了用于不一致示例的加权机制，结果表明这些示例提供了更多的信息。Multihead Co-training (Chen et al., 2021) 提出了一种多头架构，以缓解必须训练多个分类器的限制，其中每个头部使用其他头部的一致性作为伪标签。

MultiMatch 建议利用协同训练和一致性正则化的优点。虽然 JointMatch 已经在这方面迈出了初步的步伐，但结合多头结构与基于历史预测的自适应过滤机制这一点仍未被探索。

半监督文本分类。 在文本分类领域受到广泛关注，因为它能够通过各种方法利用大规模未标记数据和有限标记示例。例如，MixText (Chen et al., 2020) 通过在隐藏空间内融合标记和未标记示例以减轻在有限数据上的过拟合。UDA (Xie et al., 2020a) 通过使用强数据增强和一致性正则化来确保对扰动输入预测的稳定性。AUM-ST (Sosea and Caragea, 2022) 通过使用边缘下的面积 (Pleiss et al., 2020) 筛选不可靠的伪标签来优化自训练，而 SAT (Chen et al., 2022) 通过根据输入相似性重新排序强和弱增强来加强 FixMatch (Sohn et al., 2020)。

2 多重匹配

符号 设 $\mathcal{X} = \{(x_b, y_b); b \in \{1, 2, ..., B\}\}$ 是一个大小为 B 的有标签样本批次，而 $\mathcal{U} = \{u_b; b \in \{1, 2, ..., \mu B\}\}$ 是一个无标签样本批次，其中 μ 表示批次中的有标签数据与无标签数据的比例。设 $\alpha(\cdot)$ 和 $\mathcal{A}(\cdot)$ 分别为弱增强函数和强增强函数。给定输入的模型预测分布表示为 $q = p_m(y|\cdot)$，而 $\hat{q} = argmax(q)$ 是独热编码的伪标签。$\mathcal{H}(\cdot, \cdot)$ 代表交叉熵函数。

2.1 背景

FixMatch (Sohn et al., 2020) 引入了 SSL 中最近最受欢迎的方向，该方向由结合一致性正则化 (Sajjadi et al., 2016) 和伪标签法 (McLachlan, 1975) 组成。

具体来说，对于一个未标记的样本 u_b，FixMatch 使用弱和强增强函数分别创建两个增强版本 $\alpha(\cdot)$ 和 $\mathcal{A}(\cdot)$。模型通过这两个版本 $\alpha(u_b)$ 和 $\mathcal{A}(u_b)$ 获得预测结果 $q_b = p_m(y|\alpha(u_b))$ 和 $Q_b = p_m(y|\mathcal{A}(u_b))$。对弱增强版本的预测被转换为伪标签 $\hat{q}_b = \arg\max_y(q_b)$。FixMatch 使用预定义阈值 τ，以便只有通过过滤函数 $\mathbb{1}_\tau(q_b) = (\max(q_b) > \tau)$ 的高置信度伪标签才会用于无监督损失计算：

$$l_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}_\tau(q_b) H(\hat{q}_b, Q_b) \quad (1)$$

FlexMatch (Zhang et al., 2022) 和 FreeMatch (Wang et al., 2023) 表明，使用固定置信度阈值 τ 是次优的，因为它没有考虑到不同类别的难度变化以及模型随时间增加的置信度。他们通过引入类别特定阈值来解决这些问题，根据每个类别的学习状态自适应地缩放 $\tau(c)$。具体来说，FreeMatch 计算了一个自适应的全局阈值 τ_t，它反映了模型在整个未标记数据集上的置信度，以及一个自适应的局部阈值 $\tilde{p}_t(c)$，反映了模型对特定类别的置信度。然后，将这两个阈值结合起来以获得最终的自适应类别阈值 $\tau_t(c)$，其为归一化的局部阈值与全局阈值的乘积。$\tau_t(c)$ 的计算在附录 C 中有正式描述。这些阈值用于替换 FixMatch 中的过滤函数：

$$\mathbb{1}_{Free}^{(t)} = \mathbb{1}(\max(q_b) > \tau_t(\hat{q}_b)) \quad (2)$$

MarginMatch (Sosea and Caragea, 2023) 通过展示模型的当前置信度不一定是伪标签质量的良好估计来建立在先前的方法之上。相反，他们提出历史信念也应该被考虑在内。为此，他们引入了伪边界（PM）的概念，它是针对每个样本相对于迭代中的伪标签 c 计算的，计算方

式是分配的伪标签 c 对应的 logit z_c 与最大的其他 logit 之间的差：

$$PM_c^{(t)}(u\_b) = z_c^{(t)} - \max_{i \neq c} z_i^{(t)} \quad (3)$$

然后，使用指数移动平均法计算平均伪边距（APM），其中使用平滑参数 λ_m 以偏向后期迭代：

$$APM_c^{(t)}(u\_b) = PM_c^{(t)}(u\_b) \cdot \frac{\lambda_m}{1+t} + APM_c^{(t-1)}(u\_b) \cdot (1 - \frac{\lambda_m}{1+t}) \quad (4)$$

最终，当前迭代的 APM 置信度阈值 $\gamma^{(t)}$ 通过一个新引入的虚拟错误样本类计算得出，过滤函数变为 $\mathbb{1}_{APM}^{(t)} = \mathbb{1}(APM_{\hat{q}_b}^{(t)}(u_b) > \gamma^{(t)})$ 。

2.2 多重匹配：多头一致性正则化匹配

MultiMatch 是我们提出的半监督学习算法，它基于并扩展了多个 SSL 框架的元素。三个主要的灵感来源是多头协同训练 (Chen et al., 2021) 、Margin Match (Sosea and Caragea, 2023) 和 Free Match (Wang et al., 2023) ，它们分别作为协同训练和通过伪标签进行一致性正则化的代表方法。

使用 MultiMatch 训练模型的流程如图 1 所示。每一批未标记的样本分别通过弱增强函数 $\alpha(\cdot)$ 和强增强函数 $\mathcal{A}(\cdot)$ 进行增强。这两个增强版本通过模型进行预测。请注意，与多头协同训练类似，该模型是多头的，这使得协同训练功能得以实现，而共享的骨干网络保留了单一模型的推理时间。在我们的算法介绍中，我们使用了三个头部，既为了简化也因为 Chen et al. (2021) 将其识别为最佳选择。

由于获得了多个预测结果，因此伪标签的选择比传统伪标签方法要复杂一些，除非舍弃有分歧的样本——我们认为这种策略并不是最优的。为了解决这个问题，我们引入了一种新的伪标签加权模块（PLWM），它有三个主要功能：选择伪标签、过滤掉不可靠的伪标签以及根据样本的估计难度进行加权。这些功能是通过我们三方面的方法整体实现的，该方法集成并优化了头结点的一致性、自适应阈值及 APM 得分。

MultiMatch 的伪代码演示可以在算法 1 中查看。有关该算法的更多详细信息将在下文中提供。

2.3 伪标签加权模块（PLWM）

在协同训练框架中，训练样本通常分为两类：模型在预测上达成一致的样本和存在分歧的样本。然而，不同的方法对它们的处理方式不同。一方面，一些方法认为一致样本无关紧要，因为它们不传递新的信息，因此从损失计算中去除它们是为了防止模型的收敛。这些方法只使用分歧样本进行训练。另一方面，多头方法认为一致样本提供了高置信度的伪标签，因此使用这些样本并从损失计算中去除分歧样本。最后，JointMatch 尝试通过仅将一致性作为一种加权机制来平衡这两方面，而过滤本身通过一个独特的基于置信度的启发式方法来完成。

我们引入了一种利用头部一致性的新机制。通过将其与 APM 分数结合，我们将生成的伪标签分为三类：无用的、有用 & 困难的和有用 & 简单的。

如果用于生成伪标签的模型的头部达成一致（$\hat{q}_b^i = \hat{q}_b^j$），并且两者在 APM 分数方面都具有高度信心，则认为该样本是有用的 & 简单的，因此它被纳入损失计算，权重为 1 使用商定的伪标签 。如果其中一个头具有高度信心而另一个头具有低信心，就 APM 分数而言，我们认为该样本是有用的 & 困难的，因为一个头对预测不自信。如果两个头都具有低信心，或者如果两个头都具有高度信心但 对标签有异议，我们认为该样本在当前迭代中无用，并为其分配权重 0 。为了定义 APM 分数方面的高度信心，我们引入了一种新的方法来计算自适应 APM 阈值，详细内容见第 2.4 节。最后，我们根据每个头的当前信心采用额外的自适应过滤。头部 h 和未标记批次 u_b 的权重 W_{Multi}^h 的计算在公式 5 中正式描述。

$$W_{Multi}^h = [1 \cdot (\mathbb{1}_{Multi}^i \wedge \mathbb{1}_{Multi}^j \wedge \mathbb{1}_{Agree}^h) + w_d \cdot (\mathbb{1}_{Multi}^i \oplus \mathbb{1}_{Multi}^j)] \cdot \mathbb{1}_{FreeMulti}^h \quad (5)$$

在中，$\mathbb{1}_{Multi}^k = \mathbb{1}(APM_{\hat{q}_b^k}^{(t)}(u_b) > \gamma_{k,c}^{(t-1)})$ 和 $k \in \{i, j\}$ 指示头部 k 的高置信度，$\mathbb{1}_{Agree}^h = [\hat{q}\_b^i = \hat{q}\_b^j], \{i, j\} = \{1, 2, 3\} \setminus \{h\}$ 表示其余头部 i 和 j 的一致性，$\mathbb{1}_{FreeMulti}^h = \mathbb{1}_{Free}^i \vee \mathbb{1}_{Free}^j$ 是基于当前置信度的自适应阈值滤波，适用于多个头部。

2.4 自适应 APM 阈值

使用平均伪边缘 (Sosea and Caragea, 2023) 在解决依赖当前模型置信度评估伪标签质量时出现的常见问题上展示了巨大的潜力。然而，MarginMatch 使用的阈值法仍然有其局限性，因为自适应阈值 $\gamma^{(t)}$ 对所有类别都是相同的。我们通过使用考虑到每个类别不同学习难度的分类阈值来解决这个问题。

进一步，我们利用多头架构，不引入额外的虚拟类别，而是使用头部达成一致的示例子集作为阈值估计所用的子集。

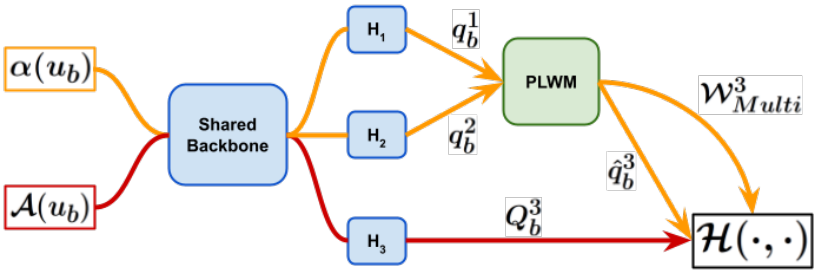

Figure 1: MultiMatch 架构 有三个头 。橙色和红色线条分别显示了弱增强和强增强样本的路径。绿色的 PLWM 模块接收来自头 H_1 和 H_2 的弱增强样本的 q_b^1 和 q_b^2 预测，然后生成伪标签 $\hat{q}_b^3$ 和对应的权重 $\mathcal{W}_{Multi}^3$ 。它们与头 H_3 的强增强样本的预测 Q_b^3 一起用于采用交叉熵函数计算无监督损失 $\mathcal{L}_u^3$ 。尽管图中显示的是头 H_3 的无标签损失计算，类似的计算也用于 H_1 和 H_2 。

Algorithm 1 多重匹配

Require: Labeled batch $\mathcal{X} = \{(x_b, y_b); b \in (1, 2, ..., B)\}$, unlabeled batch $\mathcal{U} = \{u_b; b = (1, 2, ..., \mu B)\}$, where B is the batch size of labeled data, μ is the ratio of unlabeled to labeled data, unsupervised loss weight w_u , disagreement weight δ , FreeMatch EMA decay λ_f , MarginMatch EMA decay λ_m , weak and strong augmentations $\alpha(\cdot)$, $\mathcal{A}(\cdot)$.

for h in 1 to 3 **do** ▷ For each classification head.
  $i, j = \{1, 2, 3\} \setminus \{h\}$ ▷ The other 2 heads are used for generating pseudo-labels.
  $\mathcal{L}_s^h = \frac{1}{B} \sum_{b=1}^{B} \mathcal{H}(p_b, p_m^h(y|\alpha(x_b)))$ ▷ Compute the supervised loss for head h using the labeled data.
  Compute $\mathbb{1}_{Free}^{(t)}(i)$ and $\mathbb{1}_{Free}^{(t)}(j)$ using 方程 2 for the remaining heads i and j .
  Compute $APM_{k,c}^{(t)}(u_b)$ for all classes $c \in \{1, ..., C\}$, heads $k \in \{i, j\}$ and all samples $u_b \in \mathcal{U}$ using Eq. 3 and 4.
  Select $\hat{q}_b^h$ and compute W_{Multi}^h using Eq. 5. ▷ Use PLWM to select, filter, and weight pseudo-labels.
  $\mathcal{L}_{u,t}^h = \frac{1}{\mu B} \sum_{b=1}^{\mu B} W_{Multi}^h \cdot H(\hat{q}_b^h, Q_b^h)$ ▷ Compute the unsupervised loss for head h using the generated pseudo-labels and weights.
end for
Compute $\gamma_{k,c}^{(t)}$, $k \in \{1, 2, 3\}$, using Eq. 6. ▷ Update our self-adaptive APM thresholds.
$\mathcal{L}_s = \sum_{k \in \{1,2,3\}} \mathcal{L}_s^k$ and $\mathcal{L}_u = \sum_{k \in \{1,2,3\}} \mathcal{L}_u^k$
$\mathcal{L} = \mathcal{L}_s + w_u \cdot \mathcal{L}_u$ ▷ Compute the total loss.

通过从一个假设被正确伪标记的子集开始，并选择 $f = 5^{th}$ 百分位数，我们获得了一个倾向于正确伪标记样本的阈值估计，最小化通过阈值的错误伪标签的数量。另一方面，MarginMatch 从一个错误分类的子集开始，并使用 95^{th} 百分位值，因此可能选择一个允许更多错误伪标签通过的阈值，导致模型错误累积。最后，按照最小化噪声伪标签的原则，我们为计算的阈值设定了一个 0 的下界。

方程 6 显示了在迭代 t 时，使用其他头的共识来选择头 h 和类别 c 的阈值 $\gamma_{h,c}^{(t)}$ 的公式：

$$\gamma_{h,c}^{(t)} = max(0, \text{perc}_f(\cup_{k \neq h} \{APM_{k,c}^{(t)}(u_b) \mid \mathbb{1}_{Agree}^{h,c}\})) \quad (6)$$

，其中 $\mathbb{1}_{Agree}^{h,c} = [\hat{q}\_b^i = \hat{q}\_b^j = c], \{i, j\} = \{1, 2, 3\} \setminus \{h\}$ 标记其余两个头是否对伪标签 c 达成一致。

3 实验

我们评估了我们的 MultiMatch 算法在各种 SSL 基准数据集上的性能。具体来说，我们使用了 USB 基准 (Wang et al., 2022) ，该基准为视觉、语言和音频任务中的半监督分类提供了标准化的评估框架。我们在 USB 中包含的自然语言处理（NLP）数据集上进行了不同数量标记样本的实验：IMDB (Maas et al., 2011) 、AG News (Zhang et al., 2015) 、Amazon Review (McAuley and Leskovec, 2013) 、Yahoo! Answer (Chang et al., 2008) 和 Yelp Review (Asghar, 2016) 。为了与 USB 中已经包含的 16 个 SSL 算法进行简单而公平的比较，我们保持其原始实验设置不作修改。这确保了评估的一致性，包括使用 BERT-Base (Devlin, 2018) 非区分大小写的版本作为共享骨干网络，保持训练-验证-测试的数据划分，每类的标记和未标记样本数，以及所有的训练超参数。我们在附录 E 中提供了数据集统计信息和划分信息，并在附

Dataset	IMDB		AG News		Amazon Review		Yahoo! Answer		Yelp Review		Mean	Fried.	Final
# Label	20	100	40	200	250	1000	500	2000	250	1000	error	rank	rank
Supervised (Full)	5.69±0.15	5.72±0.13	5.78±0.07	5.73±0.11	36.4±0.05	36.4±0.05	24.87±0.04	24.84±0.04	32.04±0.03	32.04±0.03	20.95	-	-
Supervised (Small)	20.31±2.79	14.02±1.22	15.06±1.08	14.25±0.97	52.31±1.28	47.53±0.69	37.43±0.29	33.26±0.1	51.22±0.98	46.71±0.37	33.21	-	-
AdaMatch	8.09±0.99	7.11±0.2	11.73±0.17	11.22±0.95	46.72±0.72	42.27±0.25	32.75±0.35	30.44±0.31	45.4±0.96	40.16±0.49	27.59	5.70	3
FixMatch	7.72±0.33	7.33±0.13	30.17±1.87	11.71±1.95	47.61±0.83	43.05±0.54	33.03±0.49	30.51±0.53	46.52±0.94	40.65±0.46	29.83	8.70	10
FlexMatch	7.82±0.77	7.41±0.38	16.38±3.94	12.08±0.73	45.73±1.6	42.25±0.33	35.61±1.08	31.13±0.18	43.35±0.69	40.51±0.34	28.23	8.20	8
Dash	8.34±0.86	7.55±0.35	31.67±13.19	13.76±1.67	47.1±0.74	43.09±0.6	35.26±0.33	31.19±0.29	45.24±2.02	40.14±0.79	30.33	10.30	11
CRMatch	8.96±0.88	7.16±0.09	12.28±1.43	11.08±1.24	45.49±0.98	43.07±0.5	32.51±0.4	29.98±0.07	45.71±0.63	40.62±0.28	27.69	6.40	5
CoMatch	7.44±0.3	7.72±1.14	11.95±0.76	10.75±0.35	48.76±0.9	43.36±0.21	33.48±0.51	30.25±0.35	45.4±1.12	40.27±0.51	27.94	7.10	7
SimMatch	7.93±0.55	7.08±0.33	14.26±1.51	12.45±1.37	45.91±0.95	42.21±0.3	33.06±0.2	30.16±0.21	46.12±0.48	40.26±0.62	27.94	6.60	6
FreeMatch	8.94±0.21	7.95±0.45	12.98±0.58	11.73±0.63	46.41±0.6	42.64±0.06	32.77±0.26	30.32±0.18	47.95±1.45	40.37±1.0	28.21	8.40	9
SoftMatch	7.76±0.58	7.97±0.72	11.9±0.27	11.72±1.58	45.29±0.95	42.21±0.2	33.07±0.31	30.44±0.62	44.09±0.5	39.76±0.13	27.42	5.70	3
Multihead Co-training	8.70±0.88	7.46±0.68	22.72±6.09	13.48±1.51	46.22±0.70	43.07±0.79	35.17±1.58	30.81±0.28	46.46±1.28	40.79±0.49	29.49	10.50	12
MarginMatch	7.19±0.39	6.99±0.19	10.65 ±0.19	11.03±0.99	44.81±1.23	42.14±0.67	32.08±0.70	29.55±0.15	42.93±1.48	39.13±0.34	26.65	2.00	2
MultiMatch	6.89 ±0.07	6.98 ±0.27	11.14±0.96	10.59 ±0.66	44.43 ±0.98	42.09 ±0.28	30.90 ±0.70	29.39 ±0.39	42.16 ±0.79	39.08 ±0.55	26.37	1.10	1

Table 1: 在 IMDB、AG 新闻、亚马逊评论、雅虎！问答和 Yelp 评论数据集上测试错误率，使用两个不同大小的标记集配置。每个配置的最佳结果用 蓝色 标出。完整的表格在附录 G 中展示。

录 D 中提供了实验设置的更多细节。

对于在 USB 中实现的监督基线和 SSL 基线，我们报告来自 USB GitHub 页面的结果[1]，通过使用更大的批量大小，这些结果比 Wang et al. (2022) 最初报告的有改进。对于其他 2 个基线（多头联合训练和 MarginMatch）以及 MultiMatch，我们将实现添加到 USB 框架中，并报告每个实验的平均测试错误（使用相同的种子）进行 3 次运行的结果。除了跨任务的平均错误率外，我们还计算每个算法的 Friedman 排名，即在所有任务中的平均排名 $\text{rank}_F = \frac{1}{m}\sum_{i=1}^{m}\text{rank}_i$，其中 $m = 10$ 是评估设置的数量，rank_i 是算法在第 i 个设置中的排名。

尽管在自然语言处理中的半监督学习研究普遍使用平衡的设置进行评估，许多现实世界文本分类数据集的类别分布是不平衡的。此外，虽然众所周知在不平衡数据上训练的分类器对多数类别有偏差，这一问题对于半监督学习算法来说更加严重，因为使用偏向性的伪标签随着时间的推移会增加偏差。

为了评估 MultiMatch 在数据不平衡情况下的鲁棒性，我们借鉴了 ABC 在计算机视觉中的 (Lee et al., 2021) 设置，创建了来自 USB 的文本数据集的不平衡版本。由于我们专注于高不平衡场景，我们考虑长尾不平衡，其中数据点数量从最大类到最小类呈指数级减少。形式上，类 c 中的样本数量为 $N_c = N_1 \cdot \gamma^{-\frac{c-1}{C-1}}$，其中 C 是类的数量，$c \in \{1, ..., C\}$，并且 $\gamma = \frac{N_1}{N_C}$ 是最大类与最小类之间的不平衡因子。对于每个数据集，我们考虑两种高度不平衡的设置：首先，我们考虑在有标签和无标签数据集中相似的长尾不平衡。在这种情况下，我们设置不平衡因子为 $\gamma = 100$，最大类的大小为 $N_1 = 1000$，最小类为 $N_C = 10$，并且无标签集合是有标签集合的 10 倍大。对于第二种设置，我们保持有标签数据集的不平衡不变，但我们使用一个无标签不平衡因子为 $\gamma = -100$，这表示一种长尾分布，其中在有标签数据集中最具代表性的类在无标签集合中是最不具代表性的。这代表了一种极端情况，即无标签语料库是从有标签数据集中单独收集的，且类分布可能完全不同。

由于 SSL 算法在不平衡情况下的表现与在平衡情况下明显不同，并且在不平衡环境中算法之间的性能差距更大，我们旨在整合一种类不平衡半监督学习（CISSL）方法来评估缩小这一差距的潜力。鉴于其多功能性——能够与现有的 SSL 算法无缝集成——我们通过在不平衡情况下用 ABC (Lee et al., 2021) 增强每个测试的 SSL 算法，进行了额外的实验。有关 CISSL 和 ABC 的更多细节，请参见附录 A 。

4 结果

表格 1 中展示了在 USB 基准中的文本分类数据集上，MultiMatch 和其他 SSL 方法的比较。我们的结果显示，MultiMatch 一直优于所有基线，在 10 个设置中取得了 9 次最优表现，并且根据 Friedman 测试在 19 个 SSL 算法中排名第一。虽然与第二佳方法相比，总体平均错误率也有绝对值 0.28 % 的降低，但我们注意到，当标注的样本数量较少时，提升更加显著。具体来说，在资源较少的设置中，相较于提供更多标注数据的情况下，MultiMatch 相比 MarginMatch 的平均改进为 0.43 %，而后者为 0.14 %。

通过分析为我们的方法提供灵感的基线，我们观察到 MarginMatch 一致地优于除 MultiMatch 之外的所有方法，这强化了利用历史证据进行伪标签过滤的重要性。相比之下，多头协同训练和 FreeMatch 分别排名为 12^{th} 和

[1] https://github.com/microsoft/Semi-supervised-learning

Dataset	IMDB		AG News		Amazon Review		Yahoo! Answer		Yelp Review		Mean	Fried.	Final
Imbalance	100	-100	100	-100	100	-100	100	-100	100	-100	error	rank	rank
FixMatch	49.95±0.06	49.92±0.14	31.95±5.30	36.09±6.50	64.69±0.92	61.93±4.41	51.64±0.75	52.78±1.45	65.57±3.19	57.43±6.22	52.20	8.40	10
+ ABC	49.69±0.43	45.90±7.07	38.08±8.80	21.21±5.72	62.02±1.03	56.27±1.16	61.15±3.12	53.79±9.31	62.63±0.89	58.37±6.97	50.91	6.30	7
FreeMatch	49.95±0.03	42.75±12.56	26.55±4.87	24.58 ±3.71	62.26±0.48	59.22±1.94	44.93±1.00	40.96 ±0.43	63.92±3.98	59.13±1.08	47.43	6.00	6
+ ABC	49.97±0.04	30.62±18.18	23.76±5.45	24.01±1.66	58.13±0.93	54.86 ±0.34	49.60±7.68	44.04±4.18	61.58±2.69	54.02±3.44	45.06	3.90	4
Multihead Co-training	49.91±0.12	50.00±0.00	30.88±2.81	34.43±4.51	62.90±2.24	58.05 ±4.53	52.05±1.28	51.39±1.51	64.32±3.06	58.62±3.34	51.25	7.90	9
+ ABC	49.76±0.24	50.00±0.00	25.29±1.41	26.18±1.20	57.99 ±3.54	57.07±1.56	46.92±2.81	49.00±6.97	63.48±3.90	55.38±1.50	48.11	5.10	5
MarginMatch	49.60±0.62	49.99±0.01	26.46±0.99	33.85±2.93	64.74±0.39	66.45±0.81	50.59±0.86	53.96±0.81	63.33±0.58	63.70±2.41	52.27	7.80	8
+ ABC	45.04 ±4.14	42.55±12.85	23.74 ±2.77	17.88 ±0.75	59.98±1.45	55.71±3.13	48.48±9.07	52.75±2.71	61.43±3.15	55.42±0.33	46.30	3.20	2
MultiMatch	49.17 ±0.88	26.05 ±6.28	21.11 ±0.49	25.36±2.60	61.03 ±1.64	59.66±2.58	41.01 ±0.79	41.46±4.34	60.14 ±3.41	56.71 ±1.87	44.17	3.30	3
+ ABC	48.81±1.98	17.21 ±3.78	29.36±11.90	19.65±4.39	62.13±1.46	57.12±0.88	41.05 ±1.34	42.72 ±2.57	60.71 ±2.68	53.87 ±3.03	43.26	3.10	1

Table 2: 在高度不平衡的设置中测试错误率。最佳结果 没有 / 与 ABC 在 蓝色 / 紫色 中突出显示。不平衡 100：标记和未标记数据具有相似分布；不平衡 -100：未标记数据集呈现相反的长尾分布。

8^{th}，这表明投票机制和自适应阈值虽然有益，但单独使用时并不是最具影响力的技术。然而，当经过改进并融入我们独特的三重伪标签加权模块（PLWM）时，它们的综合效果在所有评估设置中均带来了显著的性能提升。

不平衡设置。 我们在来自 USB 的相同数据集的新创建的高度不平衡设置中评估 MultiMatch 及其融合技术的代表。具体来说，如表格 2 所示，我们将 MultiMatch 与 MarginMatch 进行比较，该方法代表平均伪边距方法；与 Multihead Co-training 进行比较，后者体现了联合训练中的投票机制；与 FreeMatch 进行比较，代表自适应类阈值技术；以及与 FixMatch 进行比较，后者作为伪标签的一致性正规化的基准。

MultiMatch 在所有基准中表现优异，在 10 个设定中排名第一，在所有任务中平均比第二好的方法 FreeMatch 高出 3.26 %。特别是，在不平衡因子为 100、未标记分布与标记分布相似的所有五个设置中，MultiMatch 达到了最先进的性能，并在未标记分布不同的情况下超过 FreeMatch 3.48 %。与平衡设置的主要竞争对手 MarginMatch 相比，MultiMatch 在不平衡设置中平均超出 8.1 %，当未标记分布与标记分布不同时，有 11.74 % 的显著提升。附录 B 中提供了详细的技术讨论，解释了 MultiMatch 的设计选择是如何对其显著的性能提升做出贡献的。

通过增强 SSL 算法的 ABC，提升了每种算法的整体性能。虽然 MarginMatch 受益最大，提升了 5.97 %，而 MultiMatch 受益最小，仅提升了 0.91 %，但 MultiMatch 仍然是表现最佳的方法，在五个设置中取得了最先进的结果，其错误率比排名第二的 FreeMatch+ABC 低了 1.8 %。值得注意的是，即使没有 ABC 增强，MultiMatch 在平均错误率方面仍然优于所有其他经过 ABC 增强的算法。这突显了我们的方法在应对不平衡类别分布方面的稳健性，而这是实际应用于文本分类数据集所必需的特性。

5 消融研究与分析

为了证明 MultiMatch 中每个组件的有效性，我们在系统地移除不同元素后评估其性能，如表格 3 所示。我们观察到，在移除每个组件后性能都有所下降，这表明 MultiMatch 中的所有组件在实现最佳性能中都起到了重要作用。

首先，为了评估有用 & 困难样本的重要性，我们去掉了权重 w_d，因此它们将和有用 & 易样本被赋予相同的权重。这导致 10 个设置中的 9 个性能下降，平均错误率增加了 0.3 %。同样，通过完全从训练数据中删除这些样本，我们平均没有注意到任何实质性的差异。其次，我们移除了基于模型当前置信度 $\mathbb{1}_{FreeMulti}$ 的自适应阈值的过滤。这导致我们的 PLWM 仅使用头部间的一致性和 APM 评分来过滤伪标签。这导致性能显著下降 1.65 %，突显了当前模型信念在我们的 PLWM 中的重要性。第三，我们从 MultiMatch 中移除了基于 APM 的过滤 $\mathbb{1}_{Multi}$，得到的 PLWM 模块仅使用至少一个头部对所选伪标签有高置信度的一致性样本。在这种情况下，平均错误率增加了 1.21 %，并且 Friedman 测试将其放在消融研究中最低的位置。最后，移除针对类阈值的低限 $\gamma_{k,c} > 0$ 的影响最小，仅导致 0.15 % 的错误率增加，但这仍为更严格的过滤机制在减少伪标签噪声和防止其逐步退化方面的重要性提供了额外的证明。

我们使用两种指标：掩码率和杂质来比较 MultiMatch、MarginMatch、多头协同训练和 FreeMatch 的伪标签质量。我们在 AG 新闻任务上展示了 200 个标签及其高度不平衡版本的结果，分别在图 **??** 和 2 中。掩码率被定义为在第 t 轮次训练中被排除的伪标签示例的比例，而杂质是指在轮次 t 参与训练但有错误标签的伪标签示例的比例。正如预期，我们的 PLWM 模块采用了更严格的过滤机制，导致比所有其他方法更高的掩码率。相比之下，仅依靠头部一致性被证明是无效的，因为多头协同训练的掩码率不到 1 %。同时，我们严格的过滤，结

Dataset	IMDB		AG News		Amazon Review		Yahoo! Answer		Yelp Review		Mean	Fried.	Final
# Label	20	100	40	200	250	1000	500	2000	250	1000	error	rank	rank
MultiMatch	6.89$_{\pm 0.07}$	6.98 $_{\pm 0.27}$	11.14$_{\pm 0.96}$	10.59$_{\pm 0.66}$	44.43$_{\pm 0.98}$	42.09$_{\pm 0.28}$	30.90 $_{\pm 0.70}$	29.39 $_{\pm 0.39}$	42.16 $_{\pm 0.79}$	39.08 $_{\pm 0.55}$	26.37	1.80	1
— w_d	7.08$_{\pm 0.32}$	7.01$_{\pm 0.31}$	10.93 $_{\pm 0.43}$	10.73$_{\pm 0.62}$	44.47$_{\pm 1.11}$	42.17$_{\pm 0.18}$	31.24$_{\pm 0.47}$	29.64$_{\pm 0.51}$	44.06$_{\pm 1.28}$	39.41$_{\pm 0.46}$	26.67	3.70	4
— $w_d = 0$	7.19$_{\pm 0.29}$	7.20$_{\pm 0.17}$	11.80$_{\pm 0.88}$	10.53$_{\pm 0.27}$	44.19$_{\pm 0.79}$	42.13$_{\pm 0.27}$	31.12$_{\pm 0.75}$	29.63$_{\pm 0.31}$	43.20$_{\pm 1.24}$	39.61$_{\pm 0.14}$	26.66	3.30	3
— $\mathbb{1}_{FreeMulti}$	7.23$_{\pm 0.18}$	7.02$_{\pm 0.24}$	18.41$_{\pm 5.86}$	11.28$_{\pm 0.86}$	44.58$_{\pm 0.96}$	42.08$_{\pm 0.29}$	36.15$_{\pm 1.36}$	30.35$_{\pm 0.70}$	43.94$_{\pm 1.68}$	39.16$_{\pm 0.36}$	28.02	4.60	5
— $\mathbb{1}_{Multi}$	6.88 $_{\pm 0.14}$	7.46$_{\pm 0.50}$	15.77$_{\pm 3.04}$	11.66$_{\pm 0.18}$	44.57$_{\pm 0.80}$	42.31$_{\pm 0.36}$	32.13$_{\pm 0.45}$	30.35$_{\pm 0.26}$	44.68$_{\pm 0.89}$	39.99$_{\pm 0.22}$	27.58	5.20	6
— $\gamma_{k,c} > 0$	7.09$_{\pm 0.24}$	6.98 $_{\pm 0.20}$	11.18$_{\pm 0.66}$	10.56 $_{\pm 0.63}$	44.11 $_{\pm 0.79}$	41.99 $_{\pm 0.18}$	31.14$_{\pm 0.61}$	29.59$_{\pm 0.27}$	43.37$_{\pm 0.65}$	39.17$_{\pm 0.31}$	26.52	2.40	2

Table 3: 通过从 MultiMatch 中移除一个组件进行消融研究。最佳结果在 蓝色 中突出显示。

合处理困难示例的针对性方法，有效地减少了两种设置中的杂质。相比之下，多头协同训练由于其弱过滤机制而包含许多错误的伪标签，而 FreeMatch 的自适应阈值未能在不平衡设置中防止错误积累，导致杂质随着训练的进行而上升。

6 结论

在这项工作中，我们介绍了 MultiMatch，这是一种新颖的半监督学习算法，将协同训练、一致性正则化和伪标签集成到一个统一的框架中。我们的方法有效地平衡了伪标签的选择、过滤和加权，提高了鲁棒性和性能。对 USB 基准的广泛实验表明，MultiMatch 在平衡和高度不平衡的设置中均持续优于 18 种基线方法，这对现实世界中的文本分类任务至关重要。此外，我们的消融研究突出了 MultiMatch 中每个组件的重要性，证明我们严格的三重过滤机制对于最小化错误的伪标签和提高性能至关重要。这些结果强调了混合 SSL 策略的有效性，并为开发更具适应性和鲁棒性的技术开辟了新方向。

7

限制

尽管 MultiMatch 在各种设置中表现出了强劲的性能，但它也存在某些限制。首先，与单头 SSL 方法相比，它依赖于多个分类头，增加了计算开销。在我们的案例中，参数数量大约增加了 1 %（从 110M 增加到 111M），但这可能会因主干网络的不同而有所变化。然而，与训练两个独立模型的传统协同训练方法相比，这种开销可以忽略不计。我们在附录 F 中扩展了效率方面的考虑。其次，将 MultiMatch 应用于专门的应用可能需要整合领域知识（例如，定制的增广策略）和额外的超参数调优。然而，这一限制是所有 SSL 方法共有的。在附录 H 中，我们采取初步步骤，通过在一个多模态的真实实验环境中应用它来评估 MultiMatch 的泛化能力。结果突显了 MultiMatch 在两个不同任务上表现出的特殊鲁棒性，即使没有任何超参数调优，这与其他基线方法形成对比。不过，对于高度敏感的应用，进一步分析仍然是明智的。

References

Eric Arazo, Diego Ortego, Paul Albert, Noel E O' Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. 2019a. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019b. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. 2021. Adamatch: A unified approach to semi-supervised learning and domain adaptation. *arXiv preprint arXiv:2106.04732*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Aaai*, volume 2, pages 830–835.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. 2023. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*.

Hui Chen, Wei Han, and Soujanya Poria. 2022. Sat: Improving semi-supervised text classification with simple instance-adaptive self-training. *arXiv preprint arXiv:2210.12653*.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *Preprint*, arXiv:2004.12239.

Mingcai Chen, Yuntao Du, Yi Zhang, Shuwei Qian, and Chongjun Wang. 2021. Semi-supervised learning with multi-head co-training. *Preprint*, arXiv:2107.04795.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yue Fan, Anna Kukleva, Dengxin Dai, and Bernt Schiele. 2023. Revisiting consistency regularization for semi-supervised learning. *International Journal of Computer Vision*, 131(3):626–643.

Qian Gui, Hong Zhou, Na Guo, and Baoning Niu. 2024. A survey of class-imbalanced semi-supervised learning. *Machine Learning*, 113(8):5057–5086.

Lan-Zhe Guo and Yu-Feng Li. 2022. Class-imbalanced semi-supervised learning with adaptive thresholding. In *International Conference on Machine Learning*, pages 8082–8094. PMLR.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta.

Hyuck Lee, Seungjae Shin, and Heeyoung Kim. 2021. Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. *Advances in Neural Information Processing Systems*, 34:7082–7094.

Junnan Li, Caiming Xiong, and Steven CH Hoi. 2021. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9475–9484.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Geoffrey J McLachlan. 1975. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Ferda Ofli, Firoj Alam, and Muhammad Imran. 2020. Analysis of social media data using multimodal deep learning for disaster response. In *17th International Conference on Information Systems for Crisis Response and Management*. ISCRAM.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. *Preprint*, arXiv:2001.10528.

Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. 2018. Deep co-training for semi-supervised image recognition. *Preprint*, arXiv:1803.05984.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. *Advances in neural information processing systems*, 28.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171.

Iustin Sirbu, Tiberiu Sosea, Cornelia Caragea, Doina Caragea, and Traian Rebedea. 2022. Multimodal semi-supervised learning for disaster tweet classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2711–2723, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*.

Tiberiu Sosea and Cornelia Caragea. 2022. Leveraging training dynamics and self-training for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4750–4762, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tiberiu Sosea and Cornelia Caragea. 2023. Marginmatch: Improving semi-supervised learning with pseudo-margins. *Preprint*, arXiv:2308.09037.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, Heli Qi, Zhen Wu, Yu-Feng Li, Satoshi Nakamura, Wei Ye, Marios Savvides, Bhiksha Raj, Takahiro Shinozaki, Bernt Schiele, Jindong Wang, Xing Xie, and Yue Zhang. 2022. Usb: A unified semi-supervised learning benchmark for classification. *Preprint*, arXiv:2208.07204.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. 2023. Freematch: Self-adaptive thresholding for semi-supervised learning. *Preprint*, arXiv:2205.07246.

Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. 2021. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10857–10866.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020a. Unsupervised data augmentation for consistency training. *Preprint*, arXiv:1904.12848.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020b. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. 2021. Dash: Semi-supervised learning with dynamic thresholding. In *International conference on machine learning*, pages 11525–11536. PMLR.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2022. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Preprint*, arXiv:2110.08263.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. 2022. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14471–14481.

Henry Peng Zou and Cornelia Caragea. 2023. Jointmatch: A unified approach for diverse and collaborative pseudo-labeling to semi-supervised text classification. *Preprint*, arXiv:2310.14583.

A 类不平衡的半监督学习

现有的半监督学习算法通常假设数据的类别分布是平衡的，但在实际场景中往往不是这样。此外，虽然分类器通常对多数类有偏向，这对于依赖伪标签的 SSL 算法来说是一个更大的问题。随着使用越来越偏向的预测来生成伪标签，偏差会进一步增加。为了解决这个问题，ABC (Lee et al., 2021) 提出了一个辅助的平衡分类器，可以附加到现有的 SSL 算法上以缓解类别不平衡。辅助分类器附加在模型的最后特征层上，因此它可以利用主干网络从整个数据集中学习的有用表示。然而，它通过以与其对应类别大小成反比的概率来屏蔽样本进行训练，从而实现使用平衡分类损失。请注意，ABC 损失只是基础 SSL 算法的补充损失。这使得该方法不同于简单的多数类下采样，因为基础 SSL 算法仍然完整地使用它。

由于 ABC 在使用 FixMatch (Sohn et al., 2020) 作为基础 SSL 算法的计算机视觉中类不平衡数据集上显示出的显著性能提升，以及其多功能性（因为它可以与任何基础 SSL 算法结合使用），我们也将其纳入我们的实验中。由于我们的 MultiMatch 方法在强烈不平衡的设置中比基线取得了特别令人印象深刻的结果，我们认为有必要评估将诸如 ABC 的平衡方法集成到训练过程中，是否能弥合我们的方法与其他 SSL 算法之间的差距。

为此，在高度不平衡的设置中测试的每个 SSL 算法，我们使用相同的算法与 ABC (Lee et al., 2021) 结合进行额外实验，并在表 2 中显示结果。

B 用于类别不平衡的多重匹配

根据 Gui et al. (2024)，CISSL 方法通过两种主要策略解决 SSL 算法中的偏差积累问题。第一种方法，已经在附录 A 中描述并由 ABC (Lee et al., 2021) 举例说明，通过调整分类器来模拟平衡的类别分布。在这种设置中，所有未标记的样本都有助于学习有用的表示，但分类损失的计算方式考虑了类别不平衡问题——通常通过基于类别频率加权或屏蔽示例。第二种更直接的策略旨在通过减少过度代表类别的影响来改进伪标签质量。例如，CReST (Wei et al., 2021) 修改了自训练范式，采用了一种重新采样技术，以优先选择少数类别的伪标签。同时，Adsh (Guo and Li, 2022) 引入了一种自适应阈值机制，强制对每个类别使用类似百分比的伪标签，减轻不平衡放大作用。最后，减少伪标签杂质——定义为每个周期的标记错误——可以帮助限制长期错误积累。

在这种情况下，MultiMatch 通过设计集成了

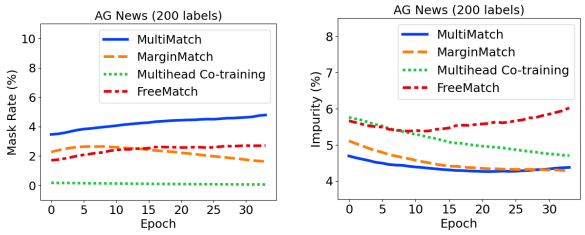

Figure 2: 在不平衡因子为 100 的 AG 新闻上，遮罩率（左）和杂质（右）。

几种 CISSL 技术，以解释其在高度不平衡环境中的卓越表现。首先，使用按类别划分的阈值允许算法区分各个类别的学习进展。从表格 2 中可以明显看出这种阈值策略的影响，通过分析 FreeMatch 的表现。虽然在平衡条件下它与 FixMatch 的性能相似，但在不平衡设置中却表现出显著的改进，在应用 ABC 增强之前排在第二位。

其次，MultiMatch 为每个类别使用基于百分位的阈值，类似于 Adsh，从而实现了跨类别的更公平的伪标签选择。尽管 MarginMatch 也采用了基于百分位的阈值，但由于缺乏逐类别的计算，因此未能达到相同的效果。

最后，MultiMatch 通过多种机制明确地促进降低伪标签的纯度。多重过滤器的使用增加了错误伪标签通过的难度，而用于阈值计算的子集通常会导致比 MarginMatch 高的阈值。如图 2 所示，MultiMatch 通过更高预测掩码率保持低伪标签纯度，从而减少错误积累。尽管 Sosea and Caragea (2023) 指出过度掩码可能会舍弃有信息的边缘例子，MultiMatch 通过给予通过所有过滤条件的有用且难度较大的样本更高的权重来补偿这一点。

C 扩展背景

FreeMatch (Wang et al., 2023) 表明使用固定的置信度阈值 τ 是次优的，因为它没有考虑到类之间难度的变化以及模型置信度随时间的增加。他们通过引入按类设置的阈值来解决这些问题，根据每个类的学习状态自适应地调整 $\tau(c)$ 的比例。具体来说，FreeMatch 计算出一个自适应的全局阈值 τ_t ，该阈值反映了模型在整个未标记数据集上的置信度：

$$\tau_t = \lambda\tau_{t-1} + (1-\lambda)\frac{1}{\mu B}\sum_{b=1}^{\mu B}\max(q_b) \quad (7)$$

和一个自适应局部阈值 $\tilde{p}_t(c)$ ，该阈值反映了模型对特定类别的置信度：

$$\tilde{p}_t(c) = \lambda\tilde{p}_{t-1}(c) + (1-\lambda)\frac{1}{\mu B}\sum_{b=1}^{\mu B}q_b(c) \quad (8)$$

与 $\tau_0 = \tilde{p}_0(c) = \frac{1}{C}$ 。

最后，这两个阈值结合在一起，以获得最终的自适应类别阈值 $\tau_t(c)$ ，作为归一化局部阈值和全局阈值的乘积。

$$\tau_t(c) = \frac{\tilde{p}_t(c)}{\max_{c'}(\tilde{p}_t(c'))} \cdot \tau_t \quad (9)$$

这些阈值用于替换 FixMatch 中的过滤功能：

$$\mathbb{1}_{Free}^{(t)} = \mathbb{1}(\max(q_b) > \tau_t(\hat{q}_b)) \quad (10)$$

D 实验设置

为了确保与 USB (Wang et al., 2022) 中已包含的算法进行直接而公平的比较，我们遵循他们的实验设置，并在所有实验中使用相同的超参数。更具体地说，我们在所有算法中使用了预训练的 BERT-Base uncased 版本，最大文本长度为 512 个标记。虽然这可能不是最优的，我们没有使用弱增强 ($\alpha(x) = x$)，以确保向后兼容。对于强增强，我们采用了使用 WMT19 的反向翻译 (Xie et al., 2020b) ，语言对为德-英和俄-英。对于其他 4 个数据集，增强已经在 USB 中提供，因此在所有实验中都是相同的，而对于 IMDB，我们必须重新生成翻译，因此由于翻译中使用的温度 0.9，可能会与 USB 的 16 个基准方法和我们实现的 3 种方法的框架之间存在差异。使用权重衰减为 1e-4 的 AdamW 优化器、总共 102,400 个训练步骤和 5,120 个预热步骤的余弦学习率调度器的优化在所有实验中均被保留。每个数据集的最佳学习率和层衰减率通过针对 FixMatch (Sohn et al., 2020) 方法的网格搜索确定，并且对所有方法均不做更改。相较于 USB (Wang et al., 2022) 中报告的设置，唯一的改变是使用了批次大小 8 而不是 4，因为这被证明可以提高性能 [2] 。基线的报告结果也使用了这种改进的批次大小。遵循 USB，我们还使用了无标签损失权重 $w_u = 1$ 以及批次中无标签和有标签样本之间的比例 $\mu = 1$ ，尽管已经表明更大的 μ 可以提高性能 (Sohn et al., 2020; Sirbu et al., 2022) 。

MultiMatch 超参数 MultiMatch 特定的超参数包括头数 H ，我们从一开始就将其设置为 3，与多头协同训练类似，使用的权重 w_d 用于有用 & 困难样本以及在计算自适应 APM 阈值时使用的百分位数 f 。对于 APM 百分位数 f ，我们尝试了 5 % 和 10 % 的值，选择了 5 % 作为最佳值。设置 f =10 % 确定了一个增加的掩码率，这排除了与无监督损失计算相关的

[2] https://github.com/microsoft/Semi-supervised-learning/blob/main/results/usb_nlp.csv

样本。对于有用 & 困难样本的权重 w_d，我们尝试了 3 和 5 的值，选择了 3 为最佳值。超参数选择是在 AG News 数据集上进行实验后完成的，其中包含 200 个标签，并选择了在 3 次运行中导致较小验证误差的值。由于 AG News 上的训练时间较快，因此选择了该方法，但这不排除不同的值可能在其他任务中是最佳的。然而，我们选择的值在所有设置中提供了稳定的性能。

我们在一个配备 8 个 Nvidia H100 GPU（80GB 显存）的系统上进行实验。每项实验在一个 GPU 上运行，多个 GPU 被用来同时运行多个实验。每次训练根据数据集不同需要 2 到 4 小时（AG News 为 2 小时，IMDB 和 Yahoo 为 3 小时，Amazon 和 Yelp 为 4 小时）。每次运行平均耗时 3 小时，每个设置 3 次运行，10 个设置（5 个数据集，每个数据集 2 种大小）和 3 个模型（MultiMatch, MarginMatch, Multi-head Co-training），主结果表的总训练时间约为 3x3x10x3 = 270 个 GPU 小时。对于不平衡设置，我们训练了 10 个模型（5 个模型带/不带 ABC），总共需要 3x3x10x10 = 900 个 GPU 小时。消融研究需要额外训练 5 个模型，总计 3x3x10x5 = 450 个 GPU 小时。加起来，总成本大约为 1,620 个 GPU 小时，或 67.5 个 GPU 天。

E 数据集

本文使用的文本分类数据集的统计数据如表 4 所示。

对于高度不平衡的设置，我们根据长尾分布为每个类别使用 1000 到 10 个标签，如 3 节所述。对于不平衡因子 100，我们为每个类别使用比有标签数据多 10 倍的无标签数据。对于不平衡因子 -100，无标签样本的总数仍然是有标签样本的 10 倍，但无标签集的长尾分布被反转。验证集和测试集与平衡设置相同。

F 效率考虑

尽管半监督学习（SSL）算法可以通过利用额外的无标签数据显著提高监督模型的性能，但由于需要处理更多的数据，它们通常会导致训练速度变慢。具体而言，对于一个监督模型，总的训练时间 $\mathcal{T}$ 与训练轮次的数量 N_e 和有标签样本的数量 N_L（即，$\mathcal{T} \approx N_L \times N_e$）成正比。然而，在 SSL 方法中，由于每个轮次处理 $(\mu + 1) \times N_L$ 个样本，总的训练时间变为 $\mathcal{T} \approx (\mu + 1) \times N_L \times N_e$。这适用于本工作使用的所有 SSL 算法。

此外，某些设计选择可能引入额外的计算开销。例如，传统的协同训练方法需要双倍的时间和内存，因为它同时训练两个模型。尽管多头协同训练通过在各个头之间共享一个主干网络来解决这个问题，但三个分类头的存在仍意味着通过两个额外的输出层进行额外计算。然而，分类头通常较小，对运行时间的影响因主干网络的不同而异。在我们使用 BERT 作为共享主干的实验中，与 FixMatch 相比，多头协同训练和 MultiMatch 的训练开销始终低于 1%。

在推理时，使用自监督学习训练的模型表现与其监督学习对手类似，因为通过模型的数据量相同。然而，多头架构的额外开销仍然存在。在我们的设置中，通过平均来自所有三个头的 logits 来进行预测，这导致相对于单头架构来说有一点推理开销（大约为 1%）。虽然这在实际部署中可能相关，但在训练后可以通过在推理时仅保留一个头来消除该开销。这在我们的任务中会导致大约 0.1% 的轻微性能下降，因为在最终训练阶段，三个头趋于高度一致。

G 结果

由于空间限制，我们从表格 1 中移除了 7 个最弱的基线。包含所有基线的完整表格版本在表格 6 中提供。结果摘要以及每个 SSL 算法的引用在表格 7 中提供。对于 共享 骨干网，我们在所有实验中使用了不区分大小写的 BERT (Devlin, 2018)。

MultiMatch 在 USB 基准测试上实现了显著的统计改进，这点通过 Friedman 检验及接下来的 Nemenyi 事后检验 (Demšar, 2006) 得以证实，这是一种用于在多个数据集上比较多个分类器的标准方法。Friedman 检验首先评估分类器在不同数据集上的排名是否存在显著差异。在发现统计显著差异后，再应用 Nemenyi 检验来识别哪些分类器对之间存在显著差异。我们的结果表明，MultiMatch 表现显著优于 USB 基准测试，突显了我们提出的方法在各种文本分类任务中的稳健性和泛化能力。

H 多模态设置

为了评估我们方法在真实世界多模态场景中的泛化能力，我们将 MultiMatch 应用于 Crisis-MMD 数据集的多模态版本 (Ofli et al., 2020)，并将其性能与几个具有代表性的基线进行比较。我们采用了由 Sirbu et al. (2022) 提出的实验设置，包括访问适合半监督学习的未标记数据。具体而言，我们使用 MMBT (Kiela et al., 2019) 作为监督骨干，并遵循其工作中的超参数配置。重要的是，我们在 NLP 实验中使用相同的超参数应用 MultiMatch，而无需任何特定任务的调优。

我们选择 CrisisMMD 有几个原因。首先，它提供了来自危机应对领域的真实世界数据。其

Dataset	Label Type	# Classes	# Labeled	# Unlabeled	# Validation	# Test
IMDB	Movie Review Sentiment	2	10 / 50	11,500	1,000	12,500
AG News	News Topic	4	10 / 50	25,000	2,500	1,900
Amazon Review	Product Review Sentiment	5	50 / 200	50,000	5,000	13,000
Yahoo! Answer	QA Topic	10	50 / 200	50,000	5,000	6,000
Yelp Review	Restaurant Review Sentiment	5	50 / 200	50,000	5,000	10,000

Table 4: 数据集统计和划分信息按每个类别的样本数量进行分类。

次，该数据集是多模态的——结合了文本和图像——这与之前使用的 NLP 数据集不同。第三，它支持具有不同特征的多种分类任务：例如，Informative 任务特征是一个几乎平衡的二元分类设置，而 Humanitarian 任务则涉及五个类别并且高度不平衡。关于数据集的更多详细信息可以在 (Ofli et al., 2020) 中找到。

如表格 5 所示，MultiMatch 显示出强大的泛化能力，在两个任务中始终优于所有其他 SSL 基准。相比之下，其他方法表现不一致。例如，FreeMatch 和 Multihead Co-training 在信息任务中表现良好，但在人道主义任务中陷入困难。以前在 NLP 数据集上表现出竞争力的 MarginMatch，在这种多模态环境中被几种替代方法超越。虽然对特定领域挑战的深入研究超出了这项工作的范围，但结果突出表明 MultiMatch 在多模态、特定领域和高度不平衡设置中的稳健性。

尽管探索更多的实际应用超出了本研究的范围，但本节的发现突出了 MultiMatch 作为通用 SSL 方法的强大潜力。它即使在特定领域和多模态环境中也能可靠地执行——无需进行任务特定的调整或超参数优化——支持其作为一种现成解决方案在各种低资源场景中的应用。

我们所有的代码将以 MIT 许可证共享，类似于 USB 框架 (Wang et al., 2022) 。我们旨在将 MultiMatch 作为开源贡献集成到 USB 中。

Task	Humanitarian				Informative			
	Acc	P	R	F1	Acc	P	R	F1
MMBT Supervised (Kiela et al., 2019)	86.71	87.20	86.75	86.74	89.44	90.07	90.06	89.87
FixMatch (Sohn et al., 2020)	88.55	88.87	88.59	88.51	89.96	89.91	90.00	89.91
FixMatch LS (Sirbu et al., 2022)	88.66	89.04	88.70	88.74	90.38	90.35	90.42	90.36
FreeMatch (Wang et al., 2023)	84.88	86.93	84.93	85.54	90.35	90.52	90.39	90.43
Multihead Co-training (Chen et al., 2021)	88.34	88.68	88.38	88.28	90.68	90.66	90.71	90.67
MarginMatch (Sosea and Caragea, 2023)	87.39	88.48	87.43	87.54	89.86	90.12	89.86	89.94
MultiMatch	89.18	89.58	89.18	89.14	91.36	91.37	91.36	91.37

Table 5: 关于多模态人道主义和信息性 CrisisMMD 任务的分类结果。每个指标的最佳结果在 blue 中突出显示。Acc 代表准确率，而 P、R 和 F1 分别是加权精确度、召回率和 F1。

Dataset	IMDB		AG News		Amazon Review		Yahoo! Answer		Yelp Review		Mean	Fried.	Final
# Label	20	100	40	200	250	1000	500	2000	250	1000	error	rank	rank
Supervised (Full)	5.69±0.15	5.72±0.13	5.78±0.07	5.73±0.11	36.4±0.05	36.4±0.05	24.87±0.04	24.84±0.04	32.04±0.03	32.04±0.03	20.95	-	-
Supervised (Small)	20.31±2.79	14.02±1.22	15.06±1.08	14.25±0.97	52.31±1.28	47.53±0.69	37.43±0.29	33.26±0.1	51.22±0.98	46.71±0.37	33.21	-	-
Pseudo-Labeling	45.45±4.43	19.67±1.01	19.49±3.07	14.69±1.88	53.45±1.9	47.0±0.79	37.7±0.65	32.72±0.31	54.51±0.82	47.33±0.2	37.20	15.30	16
Mean Teacher	20.06±2.51	13.97±1.49	15.17±1.21	13.93±0.65	52.14±0.52	47.66±0.84	37.09±0.18	33.43±0.28	50.6±0.62	47.21±0.31	33.13	14.10	14
Π -Model	49.99±0.01	44.75±3.99	60.7±19.09	12.58±0.57	77.22±1.5	53.17±2.56	44.91±1.32	32.45±0.45	75.73±4.01	59.82±0.61	51.13	16.60	17
VAT	25.93±2.58	11.61±1.79	14.7±1.19	11.71±0.84	49.83±0.46	46.54±0.31	34.87±0.41	31.5±0.35	52.97±1.41	45.3±0.32	32.50	11.90	13
MixMatch	26.12±6.13	15.47±0.65	13.5±1.51	11.75±0.6	59.54±0.67	61.69±3.32	35.75±0.71	33.62±0.14	53.98±0.59	51.7±0.68	36.31	14.30	15
ReMixMatch	50.0±0.0	50.0±0.0	75.0±0.0	75.0±0.0	80.0±0.0	80.0±0.0	90.0±0.0	90.0±0.0	80.0±0.0	80.0±0.0	75.00	18.90	19
AdaMatch	8.09±0.99	7.11±0.2	11.73±0.17	11.22±0.95	46.72±0.72	42.27±0.25	32.75±0.35	30.44±0.31	45.4±0.96	40.16±0.49	27.59	5.70	3
UDA	49.97±0.04	50.0±0.0	41.0±24.96	53.68±30.15	60.76±13.61	68.38±16.44	71.3±26.45	70.5±27.58	69.33±15.08	66.95±18.46	60.19	17.60	18
FixMatch	7.72±0.33	7.33±0.13	30.17±1.87	11.71±1.95	47.61±0.83	43.05±0.54	33.03±0.49	30.51±0.53	46.52±0.94	40.65±0.46	29.83	8.70	10
FlexMatch	7.82±0.77	7.41±0.38	16.38±3.94	12.08±0.73	45.73±1.6	42.25±0.33	35.61±1.08	31.13±0.18	43.35±0.69	40.51±0.34	28.23	8.20	8
Dash	8.34±0.86	7.55±0.35	31.67±13.19	13.76±1.67	47.1±0.74	43.09±0.6	35.26±0.33	31.19±0.29	45.24±2.02	40.14±0.79	30.33	10.30	11
CRMatch	8.96±0.88	7.16±0.09	12.28±1.43	11.08±1.24	45.49±0.98	43.07±0.5	32.51±0.4	29.98±0.07	45.71±0.63	40.62±0.28	27.69	6.40	5
CoMatch	7.44±0.3	7.72±1.14	11.95±0.76	10.75±0.35	48.76±0.9	43.36±0.21	33.48±0.51	30.25±0.35	45.4±1.12	40.27±0.51	27.94	7.10	7
SimMatch	7.93±0.55	7.08±0.33	14.26±1.51	12.45±1.37	45.91±0.95	42.21±0.3	33.06±0.2	30.16±0.21	46.12±0.48	40.26±0.62	27.94	6.60	6
FreeMatch	8.94±0.21	7.95±0.45	12.98±0.58	11.73±0.63	46.41±0.6	42.64±0.06	32.77±0.26	30.32±0.18	47.95±1.45	40.37±1.0	28.21	8.40	9
SoftMatch	7.76±0.58	7.97±0.72	11.9±0.27	11.72±1.58	45.29±0.95	42.21±0.2	33.07±0.31	30.44±0.62	44.09±0.5	39.76±0.13	27.42	5.70	3
Multihead Co-training	8.70±0.88	7.46±0.68	22.72±6.09	13.48±1.51	46.22±0.70	43.07±0.79	35.17±1.58	30.81±0.28	46.46±1.28	40.79±0.49	29.49	10.50	12
MarginMatch	7.19±0.39	6.99±0.19	10.65 ±0.19	11.03±0.99	44.81±1.23	42.14±0.67	32.08±0.70	29.55±0.15	42.93±1.48	39.13±0.34	26.65	2.00	2
MultiMatch	6.89 ±0.07	6.98 ±0.27	11.14±0.96	10.59 ±0.66	44.43 ±0.98	42.09 ±0.28	30.90 ±0.70	29.39 ±0.39	42.16 ±0.79	39.08 ±0.55	26.37	1.10	1

Table 6: 在 IMDB、AG News、Amazon Review、Yahoo! Answer 和 Yelp Review 数据集上，使用两种不同大小的标记集进行测试错误率。每种设置的最佳结果以 蓝色 突出显示。

Algorithm name	Citation	Mean error	Friedman rank	Final rank
BERT - Supervised (Full)	Devlin (2018)	20.95	-	-
BERT - Supervised (Small)	Devlin (2018)	33.21	-	-
Pseudo-Labeling	Lee et al. (2013)	37.20	15.30	16
Mean Teacher	Tarvainen and Valpola (2017)	33.13	14.10	14
Π -Model	Rasmus et al. (2015)	51.13	16.60	17
VAT	Miyato et al. (2018)	32.50	11.90	13
MixMatch	Berthelot et al. (2019b)	36.31	14.30	15
ReMixMatch	Berthelot et al. (2019a)	75.00	18.90	19
AdaMatch	Berthelot et al. (2021)	27.59	5.70	3
UDA	Xie et al. (2020a)	60.19	17.60	18
FixMatch	Sohn et al. (2020)	29.83	8.70	10
FlexMatch	Zhang et al. (2022)	28.23	8.20	8
Dash	Xu et al. (2021)	30.33	10.30	11
CRMatch	Fan et al. (2023)	27.69	6.40	5
CoMatch	Li et al. (2021)	27.94	7.10	7
SimMatch	Zheng et al. (2022)	27.94	6.60	6
FreeMatch	Wang et al. (2023)	28.21	8.40	9
SoftMatch	Chen et al. (2023)	27.42	5.70	3
Multihead Co-training	Chen et al. (2021)	29.49	10.50	12
MarginMatch	Sosea and Caragea (2023)	26.65	2.00	2
MultiMatch	Ours	26.37	1.10	1

Table 7: 我们实验中使用的所有方法作为强基准的引用。