MultiMatch: Multihead Consistency Regularization Matching for Semi-Supervised Text Classification

Iustin Sirbu[1,4], Robert-Adrian Popovici[1], Cornelia Caragea[2], Stefan Trausan-Matu[1], Traian Rebedea[1,3]
[1]National University of Science and Technology Politehnica of Bucharest, [2]University of Illinois Chicago, [3]NVIDIA, [4]Renius Technologies

Introduction & Motivation

Semi-Supervised Learning (SSL) for Text

SSL is critical for NLP tasks where **labeled data is scarce**, leveraging **large amounts of unlabeled text** to improve model generalization.

Consistency Regularization + Pseudo-labeling

- Uses high-confidence predictions from weakly augmented inputs as pseudo-labels for strongly augmented inputs.
  - **FreeMatch**: Introduces class-wise self-adaptive thresholding to mitigate class bias.
  - **MarginMatch**: Uses Average Pseudo-Margins (APM) to improve filtering by tracking predictions from previous epochs.

Co-training

- Two models are trained in parallel, each providing pseudo-labels for the other.
  - **Multihead Co-training**: Employs a multihead architecture, reducing the need for multiple networks.

Problem & Solution

- Advancements in each paradigm are independent and do not directly transfer to the other.
- **MultiMatch** introduces a **unified framework** that enhances and combines the strengths of both paradigms.

MultiMatch: Method & Architecture

For each head strongly augmented head input, the other heads are used to generate a pseudo-label based on the weakly augmented input. This is achieved through our novel **Pseudo-Label Weighting Module (PLWM)**, which combines **heads agreement with historical and current confidence** to effectively select, filter and weight pseudo-labels.

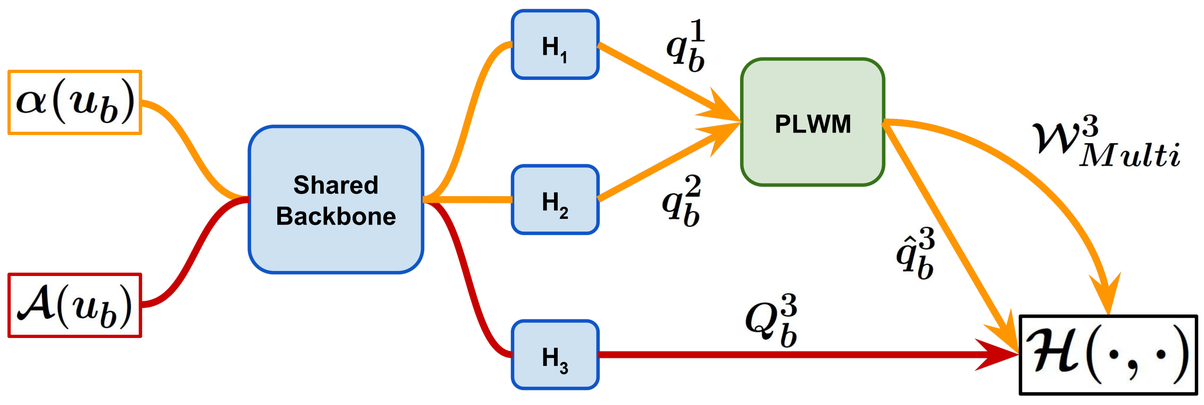

Figure 1. MultiMatch architecture with three heads.
Orange and red lines show the path of the weakly augmented and strongly augmented samples, respectively

PLWM Taxonomy & Weighting

The PLWM classifies unlabeled samples into three categories for loss computation:

- **Useful & Easy** (Weight 1): Both heads agree and both have high historical confidence (APM score).
- **Useful & Difficult** (Weight $w_d > 1$): Only one head has high confidence, indicating an informative but challenging example. The confident head gives the pseudo-label.
- **Not useful** (Weight 0): Both heads have low confidence, or they disagree while having high confidence.

Self-adaptive APM Thresholds

While employing the APM score from MarginMatch for historical stability, MultiMatch improves threshold computation for filtering confident pseudo-labels:

- **Self-Adaptive Thresholding**: Instead of a fixed threshold, we employ class-wise thresholds (similar to FreeMatch), making the filter dynamic and class-specific.
- **Improved Threshold Estimation**: We compute the threshold as the 5th percentile of the head agreement subset for each class.
  - **Benefits**: This avoids the need for the "virtual erroneous class" (used in MarginMatch) and sets a higher-quality threshold based on a confirmed, correctly-labeled subset.

Key Results

MultiMatch demonstrates superior generalization and resilience across SSL benchmarks, establishing new State-of-the-Art performance in both standard and challenging real-world settings.

High-Impact Findings:

- **Overall SOTA**: Ranks **First** according to the Friedman test among 21 competing methods.
- **Performance**: Achieved **SOTA results on 8 out of 10 setups** on 5 NLP datasets.
- **Robustness**: Outperforms baselines by **3.26%** in highly imbalanced settings, a critical advantage for real-world tasks.
- **Reduced bias**: Even without the class balancing method (ABC), MultiMatch outperforms all other **ABC-enhanced baselines** in terms of mean error.
- **Generalizability**: Outperforms baselines on the **multimodal CrisisMMD tasks** without any task-specific hyperparameter tuning.

Results on the USB benchmark

Dataset	IMDB		AG News		Amazon Rev.		Yahoo! Ans.		Yelp Review		Mean	Fried.	Final
# Label	20	100	40	200	250	1000	500	2000	250	1000	error	rank	rank
Supervised (Full)	5.69	5.72	5.78	5.73	36.40	36.40	24.87	24.84	32.04	32.04	20.95	-	-
Supervised (Small)	20.31	14.02	15.06	14.25	52.31	47.53	37.43	33.26	51.22	46.71	33.21	-	-
other 15 baselines													4-21
FixMatch	7.72	7.33	30.17	11.71	47.61	43.05	33.03	30.51	46.52	40.65	29.83	10.2	11
FreeMatch	8.94	7.95	12.98	11.73	46.41	42.64	32.77	30.32	47.95	40.37	28.21	9.7	10
Multihead Co-training	8.70	7.46	22.72	13.48	46.22	43.07	35.17	30.81	46.46	40.79	29.49	12.2	14
MarginMatch	7.19	6.99	**10.65**	11.03	44.81	42.14	32.08	29.55	42.93	39.13	26.65	2.3	2
CGMatch	7.07	**6.79**	11.95	11.29	44.77	42.61	32.15	29.85	44.34	40.14	27.10	3.9	3
MultiMatch	**6.89**	6.98	11.14	**10.59**	**44.43**	**42.09**	**30.90**	**29.39**	**42.16**	**39.08**	**26.37**	**1.2**	**1**

Table 1. Test error rates on IMDB, AG News, Amazon Review, Yahoo! Answer, and Yelp Review datasets using two setups with different sizes for the labeled set. The best result for each setup is highlighted in blue.

Results in highly imbalanced settings

Dataset	IMDB		AG News		Amazon Rev.		Yahoo! Ans.		Yelp Review		Mean	Fried.	Final
Imbalance	100	-100	100	-100	100	-100	100	-100	100	-100	error	rank	rank
FixMatch	49.95	49.92	31.95	36.09	64.69	61.93	51.64	52.78	65.57	57.43	52.20	8.4	10
+ ABC	49.69	45.90	38.08	21.21	62.02	56.27	61.15	53.79	62.63	58.37	50.91	6.3	7
FreeMatch	49.95	42.75	26.55	**24.58**	62.26	59.22	44.93	**40.96**	63.92	59.13	47.43	6.0	6
+ ABC	49.97	30.62	23.76	24.01	58.13	**54.86**	49.60	44.04	61.58	54.02	45.06	3.9	4
Multihead Co-training	49.91	50.00	30.88	34.43	62.90	**58.05**	52.05	51.39	64.32	58.62	51.25	7.9	9
+ ABC	49.76	50.00	25.29	26.18	**57.99**	57.07	46.92	49.00	63.48	55.38	48.11	5.1	5
MarginMatch	49.60	49.99	26.46	33.85	64.74	66.45	50.59	53.96	63.33	63.70	52.27	7.8	8
+ ABC	**45.04**	42.55	**23.74**	**17.88**	59.98	55.71	48.48	52.75	61.43	55.42	46.30	3.2	2
MultiMatch	**49.17**	**26.05**	**21.11**	25.36	**61.03**	59.66	**41.01**	41.46	**60.14**	**56.71**	**44.17**	**3.3**	**3**
+ ABC	48.81	**17.21**	29.36	19.65	62.13	57.12	**41.05**	**42.72**	**60.71**	**53.87**	**43.26**	**3.1**	**1**

Table 2. Test error rates in imbalanced setups. The best results without/with ABC are highlighted in blue/purple. Imbalance 100: similar distributions for labeled and unlabeled data; Imbalance -100: reversed long-tail distribution for the unlabeled set.

Results in multimodal settings

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 3. Classification results on the multimodal Humanitarian and Informative CrisisMMD tasks. The best result for each metric is highlighted in blue.

Pseudo-Label Quality Analysis

We investigate the evolution of pseudo-labels quality during training.

Metrics:

- **Mask Rate**: the proportion of pseudo-labels filtered out from training.
- **Impurity**: the proportion of incorrect pseudo-labels that participate in training.

Findings:

- MultiMatch employs a stricter filtering mechanism, resulting in a **higher mask rate**.
- MultiMatch **minimizes impurity**.
- MultiMatch **mitigates error accumulation** in the imbalanced setup.

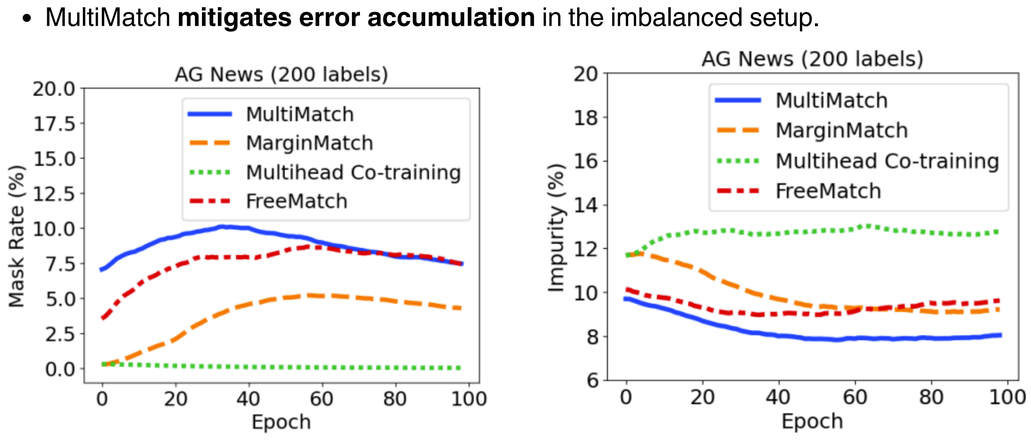

Figure 2. Mask rate and impurity on AG News with 200 labels.

Conclusion

MultiMatch establishes a unified, hybrid SSL approach that achieves superior performance and exceptional robustness in text classification tasks.

- **Novel framework**: Introduced MultiMatch, a novel algorithm integrating co-training, consistency regularization and pseudo-labeling into a unified framework.
- **SOTA performance**: Outperformed 20 baselines on the USB benchmark, ranking 1st overall.
- **Robustness**: Achieved SOTA in highly imbalanced and real-world settings.

Acknowledgements

- The project “Romanian Hub for Artificial Intelligence - HRIA”, Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906.
- The NSF IIS award 2107518.
- A UIC Discovery Partners Institute (DPI) award.