Task-Informed Anti-Curriculum by Masking Improves Downstream Performance on Text

Andrei Jarca, Florinel Alin Croitoru and **Radu Tudor Ionescu***
University of Bucharest
Bucharest, Romania

Abstract

Masked language modeling has become a widely adopted unsupervised technique to pre-train large language models (LLMs). However, the process of selecting tokens for masking is random, and the percentage of masked tokens is typically fixed for the entire training process. In this paper, we propose to adjust the masking ratio and to decide which tokens to mask based on a novel task-informed anti-curriculum learning scheme. First, we harness task-specific knowledge about useful and harmful tokens in order to determine which tokens to mask. Second, we propose a cyclic decaying masking ratio, which corresponds to an anti-curriculum schedule (from hard to easy). We exemplify our novel task-informed anti-curriculum by masking (TIACBM) approach across three diverse downstream tasks: sentiment analysis, text classification by topic, and authorship attribution. Our findings suggest that TIACBM enhances the ability of the model to focus on key task-relevant features, contributing to statistically significant performance gains across tasks. We release our code at https://github.com/JarcaAndrei/TIACBM.

1 Introduction

Nowadays, masked language modeling (MLM) (Devlin et al., 2019) is one of the most popular frameworks used to pre-train language models, as it enables the use of vast amounts of unlabeled data. However, the process of selecting tokens for masking is generally based on random selection, while the percentage of masked tokens is typically fixed for the entire training process (Wettig et al., 2023). To the best of our knowledge, there are only two studies attempting to dynamically adapt the masking ratio (Ankner et al., 2024; Yang et al., 2023). These studies concur that the optimal schedule is to use a decaying masking ratio during training. Interestingly, we find that this observation is deeply connected to the curriculum learning paradigm. Curriculum learning is a training strategy formulated by Bengio et al. (2009), where neural models learn the data in a systematic manner, starting with easy samples and gradually adding more difficult samples as the learning progresses. Intuitively, using a higher masking ratio makes the learning task more difficult. Hence, employing a decaying masking ratio corresponds to an anti-curriculum strategy (Liu et al., 2022; Soviany et al., 2022).

In this paper, we further develop and explore anti-curriculum learning based on MLM to fine-tune pre-trained models on downstream tasks. We propose a novel task-informed anti-curriculum by masking (TIACBM) scheme, which employs a cyclic decaying masking ratio and relies on task-specific knowledge to decide which tokens to mask. Our most important contribution is to harness task-specific knowledge about useful and harmful tokens in order to select tokens for masking. For example, in sentiment analysis, the masking probability of a token is determined based on its polarity scores recorded in SentiWordNet 3.0 (Baccianella et al., 2010). In text categorization by topic and authorship attribution, masking a token is conditioned by its part-of-speech tag. While content words receive higher masking probabilities for text categorization by topic, function words and punctuation tokens are more likely to be masked in authorship attribution (Kestemont, 2014).

We conduct fine-tuning experiments on four data sets: SST-2 (Socher et al., 2013), 20 Newsgroups (Lang, 1995), Reuters-21578 (Lewis, 1987), and PAN 2019 Cross-Domain Authorship Attribution (Kestemont et al., 2019). Our experiments cover a diverse set of downstream tasks, including sentiment analysis, text classification by topic, and authorship attribution. We compare with several baselines, including conventional fine-tuning and standard MLM, and state-of-the-art training strategies based on anti-curriculum learning (Ankner

*Corresponding author: raducu.ionescu@gmail.com.

Findings of the Association for Computational Linguistics: ACL 2025, pages 2192–2201
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

et al., 2024) as well as curriculum learning (Poesina et al., 2024). The results show that our strategy outperforms its competitors across all data sets. Moreover, TIACBM brings statistically significant performance gains, showing that harnessing task knowledge to mask tokens during fine-tuning is beneficial for multiple downstream tasks.

In summary, our contribution is threefold:

- We propose to leverage task-specific knowledge to determine the probability distribution used to mask input tokens in MLM.
- We introduce a cyclic decaying masking ratio that boosts performance over existing anti-curriculum learning strategies for MLM.
- We apply a dynamic MLM strategy to fine-tune pre-trained models on downstream tasks, showing that MLM is not only beneficial for pre-training, but also for the fine-tuning stage.

2 Related Work

Our framework is mostly related to work on curriculum learning (Bengio et al., 2009). Soviany et al. (2022) divide curriculum learning methods into data-level (Chang et al., 2021; Gong et al., 2021; Kocmi and Bojar, 2017; Liu et al., 2018; Nagatsuka et al., 2023), model-level (Croitoru et al., 2025; Sinha et al., 2020), task-level (Liu et al., 2020a; Narvekar et al., 2016), and objective-level (Pathak and Paffenroth, 2019) strategies. Most of existing studies focus on computer vision (Croitoru et al., 2025; Huang et al., 2020; Sinha et al., 2020) and reinforcement learning (Fang et al., 2019; Florensa et al., 2017). Methods in these domains are vaguely related to our approach, with some exceptions that employ curriculum based on masked image modeling (MIM) (Jarca et al., 2024; Madan et al., 2024). In the image domain, the MIM approach was explored from multiple perspectives, which led to the development of adaptive masking strategies based on curriculum learning (Madan et al., 2024), that can produce more robust representations. A notable finding is that an easy-to-hard curriculum works generally well for image masking (Jarca et al., 2024). In contrast, analogous studies focusing on text (Ankner et al., 2024; Yang et al., 2023) suggest that a hard-to-easy curriculum, i.e. using a decaying masking ratio, is more appropriate for text. Our results confirm the observations of Ankner et al. (2024) and Yang et al. (2023), although we reset the masking ratio during training, resulting in a cyclic decaying masking ratio.

Curriculum learning methods specifically designed for text (Gong et al., 2021; Kocmi and Bojar, 2017; Liu et al., 2018, 2020b; Zhan et al., 2021) are also related to our work. Some of the most popular approaches rely on text length (Nagatsuka et al., 2023) and model competence (Platanios et al., 2019) to organize the samples from easy to hard. Recent approaches are based on more complex strategies. For instance, the state-of-the-art curriculum learning method proposed by Poesina et al. (2024) employs data cartography (Swayamdipta et al., 2020) while training a baseline model to obtain the variability and confidence of each sample. The training data is further mapped as easy, ambiguous or hard. The model is then retrained via an easy-to-hard curriculum. To boost performance, the method employs stratified sampling as well as a continuous function to map the data points, resulting in a method called Cart-Stra-CL++.

Different from related curriculum and anti-curriculum learning methods (Ankner et al., 2024; Poesina et al., 2024; Yang et al., 2023), we design an anti-curriculum strategy for the fine-tuning stage of pre-trained language models, leveraging knowledge about the downstream tasks. Moreover, our novel design leads to superior performance on a range of downstream tasks.

3 Method

We propose a novel task-informed anti-curriculum by masking to fine-tune pre-trained language models. Specifically, we employ a cyclic decaying masking ratio which encourages a progressive adaptation of the model over time. In addition, we harness task-specific knowledge to determine which tokens need to be masked, ensuring that the model focuses on the most relevant words in a sentence. By combining a dynamic masking ratio with selective token masking, our strategy can boost performance on complex downstream tasks. We emphasize that relevant words are those that lead to discriminative features. However, deep neural networks can learn co-adapted features, which affects generalization capacity. Masking some of the discriminative features prevents feature co-adaptation (Hinton et al., 2012). TIACBM achieves this effect in a targeted manner, by strategically starting with a higher masking ratio and gradually reducing it, which aligns with theories of curriculum learning and adversarial training.

Prior to the start of the training, we create a vector $\mathbf{r} = \{r_1 \geq \cdots \geq r_K\} \in [0,1]^K$ contain-

Algorithm 1: TIACBM

Input: $\mathbf{x}$ – training sequence of tokens;
$\mathbf{r} = \{r_1, \ldots, r_K\}$ – decaying masking ratio schedule;
`task_relevance` – function that computes the task-specific importance of a word;
t – current training iteration.
Output: $\hat{\mathbf{x}}$ – masked sequence of tokens.

/* 1: Determine the number of masked tokens. */
1. $r_t \leftarrow \mathbf{r}[t \mod K]$;
$N \leftarrow \lfloor |\mathbf{x}| \cdot r_t \rfloor$;

/* 2: Compute the task-specific importance and normalize the values to obtain a probability distribution over all tokens. */
2. $\mathbf{s} \leftarrow \{s_i = \texttt{task\_relevance}(x_i, \mathbf{x}), \forall i = 1, \ldots, |\mathbf{x}|\}$;
$\mathbf{p} \leftarrow \left\{ p_i = \frac{s_i}{\sum_{j=1}^{|\mathbf{s}|} s_j}, \forall i = 1, \ldots, |\mathbf{s}| \right\}$

/* 3: Mask the tokens. */
3. $n \leftarrow 0$;
$\hat{\mathbf{x}} \leftarrow \mathbf{x}$;
while $n < N$ **do**
  $i \sim \text{Categorical}(|\mathbf{x}|, \mathbf{p})$;
  if $\hat{x}_i \notin \{$`[MASK]`, `[SEP]`, `[CLS]`$\}$ **then**
    $\hat{x}_i \leftarrow$ `[MASK]`;
    $n \leftarrow n + 1$;
  end
end
4. **return** $\hat{\mathbf{x}}$

ing K masking ratios, which represents the anti-curriculum schedule. Note that $K \ll T$, where T is the total number of training iterations. Thus, after every K iterations, we reuse the masking ratios starting with r_1. In Algorithm 1, we formally present how the masking is performed for a training text sample, given as a sequence of tokens $\mathbf{x}$. In the first step, we compute the number of tokens to be masked N, based on the sequence length and the masking ratio r_t of the current training iteration t. In the second step, for each token x_i, we call a task-specific function to compute its importance, taking into account the token and the surrounding text. Additionally, in this step, we also normalize the importance scores to obtain probability values. Finally, in the third and last step, we build a categorical distribution from these probabilities and sample N tokens to mask. We further describe and motivate the `task_relevance` function for each task.

Sentiment analysis. For polarity classification, we hypothesize that the most subjective words represent the most important features, and masking them will result in a hard-to-easy curriculum. The core foundation of this approach lies in using the SentiWordNet 3.0 sentiment lexicon (Baccianella et al., 2010). For each sentence, we analyze the most probable synset of each word, using a generic Lesk algorithm, and search for it in the lexicon. This process aims to determine the most likely positive (s_{pos}) and negative scores (s_{neg}) for the current word. Both scores range from 0 to 1, with higher values indicating stronger positive or negative connotations, respectively. We emphasize that these scores are linked together and their sum is lower or equal to 1. Based on these values, Baccianella et al. (2010) determine the objectivity score for each word as:

$$o = 1 - (s_{\text{pos}} + s_{\text{neg}}). \tag{1}$$

In contrast, we leverage the subjectivity score as an importance measure to form the vector $\mathbf{s}$ for a given input $\mathbf{x}$ in Algorithm 1. Accordingly, we compute the importance score s_i for each token x_i as follows:

$$s_i = (s_{\text{pos}}^i + s_{\text{neg}}^i), \forall i = 1, \ldots, |\mathbf{x}|. \tag{2}$$

Text categorization. For text categorization, we hypothesize that content words (nouns, verbs, adjectives, adverbs, and proper names) are more relevant. Consequently, we assign an importance score of 0 to every other part of speech. To compute the importance scores for content words of a given sequence, we also draw upon the knowledge of the pre-trained language model. Before fine-tuning, we extract the attention weights from each attention block and each attention head, given by:

$$A_b^h = \text{softmax}\left(\frac{Q_b^h \cdot (K_b^h)^\top}{\sqrt{d}}\right), \tag{3}$$

where b iterates over the self-attention blocks, h iterates over the attention heads, and $Q_b^h \in \mathbb{R}^{|\mathbf{x}| \times d}$ and $K_b^h \in \mathbb{R}^{|\mathbf{x}| \times d}$ are the query and key matrices of the attention block b and head h. The result, $A_b^h \in \mathbb{R}^{|\mathbf{x}| \times |\mathbf{x}|}$, is a square matrix containing the similarity between each token found in the input sequence $\mathbf{x}$ and all the other tokens. We further compute an importance vector $\mathbf{a}$ by averaging the attention matrices, as follows:

$$\mathbf{a} = \frac{1}{B \cdot H \cdot |\mathbf{x}|} \sum_{h=1}^{H} \sum_{b=1}^{B} \sum_{j=1}^{|\mathbf{x}|} A_{b,j}^h, \tag{4}$$

where B is the number of attention blocks and H is the number of heads. The final importance scores $\mathbf{s}$ employed in Algorithm 1 are computed as:

$$s_i = \begin{cases} a_i, \text{if } x_i \text{ is a content word} \\ 0, \text{otherwise} \end{cases}, \forall i = 1, \ldots, |\mathbf{x}|. \tag{5}$$

Authorship attribution. For the authorship attribution task, we compute the importance score vector $\mathbf{s}$ using a procedure similar to the one described for text classification. However, instead of using the content words for masking, we mask functional words, such as adpositions, determiners, conjunctions, symbols, particles, and punctuation. Authors exhibit consistent writing style patterns, which are reflected in their use of these functional words (Kestemont, 2014). Consequently, Eq. (5) is modified as follows:

$$s_i = \begin{cases} a_i, \text{if } x_i \text{ is a functional word} \\ 0, \text{otherwise} \end{cases}, \forall i = 1, ..., |\mathbf{x}|, \tag{6}$$

where a_i is computed as in Eq. (4).

4 Experiments and Results

4.1 Data sets

We evaluate TIACBM on the three tasks, namely sentiment analysis, text categorization and authorship attribution.

Reuters-21578. Reuters-21578 (Lewis, 1987) is a multi-label text categorization data set containing 12,449 training and 5,458 test instances. The data set gathers documents from 90 categories.

20 Newsgroups. 20 Newsgroups (Lang, 1995) is a multi-class data set for text categorization, which comprises 11,314 training instances and 7,532 test instances belonging to 20 classes.

SST2. The SST2 data set (Socher et al., 2013) is a popular benchmark for sentiment analysis, comprising 67,349 training and 872 validation samples, which are labeled either as positive or negative.

PAN19. For authorship attribution, we use the PAN19 (Kestemont et al., 2019) data set. We report results for Problem 0001 (P1) and Problem 0005 (P5). We discard the unknown files when computing the evaluation metrics. Both problems have 9 authors, with 7 training files each. There are 561 test files for P1, and 264 test files for P5.

4.2 Baselines

We compare TIACBM with five fine-tuning strategies, which are described in detail below.

Conventional. This is the standard fine-tuning approach, which does not involve MLM. It uses the CB-NTR loss (Huang et al., 2021) for Reuters-21578, due to its long-tail distribution.

Constant. This fine-tuning strategy uses a constant masking ratio to mask input tokens. The masking ratio is set to 15%, following Devlin et al. (2019).

Cart-Stra-CL++. This is a state-of-the-art easy-to-hard curriculum approach introduced by Poesina et al. (2024). This method needs to perform data cartography for the baseline fine-tuned with the conventional regime, before employing the curriculum. This essentially doubles the training time.

Decaying Masking Ratio. The decaying masking ratio, a.k.a. anti-curriculum by masking, is proposed by Ankner et al. (2024). This training strategy can be seen as an ablated version of our approach, which is obtained by dropping the cyclical regime ($K = T$) and by discarding task-specific information.

Cyclic Decaying Masking Ratio. This is an ablated version of our approach, which simply discards the task-specific information.

4.3 Experimental Setup

We employ two pre-trained masked language models, $\text{BERT}_{\text{base}}$ (Devlin et al., 2019) and $\text{RoBERTa}_{\text{base}}$ (Liu et al., 2019), in order to evaluate the various fine-tuning strategies. We also experiment with GPT-2 (Radford et al., 2019) to show that TIACBM can be applied beyond masked language models.

We run all experiments three times and report the mean and standard deviation for each experiment. We execute the fine-tuning for 15 epochs for sentiment analysis, using a learning rate of $5 \cdot 10^{-5}$, a batch size of 64 and a max token length of 100. For text categorization, we train for 30 epochs, using a learning rate of 10^{-4}, a batch size of 32 and a max token length of 512. For authorship attribution, we use 30 epochs, a batch size of 8 and a max token length of 512. For PAN19-P1, we use a learning rate of 10^{-4} for BERT and 10^{-5} for RoBERTa and GPT-2. For PAN19-P5, we set the learning rate to $5 \cdot 10^{-5}$ for all language models. We use the cross-entropy loss for all data sets, except on Reuters-21578. In this case, the baseline language model is fine-tuned with the CB-NTR loss (Huang et al., 2021). We keep the same loss for all fine-tuning strategies on Reuters-21578. In terms of optimizers, we use Adamax for the sentiment analysis task, and AdamW for the others. We keep these parameters consistent across all baseline methods and TIACBM. We release our code to reproduce the results at `https://github.com/JarcaAndrei/TIACBM`.

4.4 Results

We compare our approach against the competing fine-tuning strategies in Table 1. Our method improves performance across all downstream tasks, while exhibiting lower variability. Moreover, TIACBM brings significant gains across both

Model	Fine-Tuning Strategy	Reuters	20 News	SST2	PAN19-P1		PAN19-P5	
		Micro F1		Accuracy	Accuracy	Macro F1	Accuracy	Macro F1
BERT$_{\text{base}}$	Conventional	$90.61_{\pm 0.28}$	$84.63_{\pm 0.28}$	$93.38_{\pm 0.14}$	$69.76_{\pm 15.11}$	$58.24_{\pm 6.06}$	$66.10_{\pm 1.56}$	$36.10_{\pm 4.22}$
	Constant	$90.81_{\pm 0.24}$	$84.98_{\pm 0.12}$	$93.94_{\pm 0.15}$	$63.70_{\pm 9.96}$	$47.50_{\pm 7.26}$	$65.76_{\pm 2.92}$	$36.30_{\pm 4.56}$
	Poesina et al. (2024)	$90.72_{\pm 0.13}$	$82.30_{\pm 0.25}$	$94.00_{\pm 0.14}$	$55.23_{\pm 3.48}$	$44.76_{\pm 2.56}$	$68.66_{\pm 2.00}$	$40.72_{\pm 2.64}$
	Ankner et al. (2024)	$90.99_{\pm 0.05}$	$85.39_{\pm 0.23}$	$93.83_{\pm 0.20}$	$54.38_{\pm 16.10}$	$46.03_{\pm 8.53}$	$65.55_{\pm 0.85}$	$36.60_{\pm 4.28}$
	Cyclic Decaying	$90.96_{\pm 0.12}$	$84.88_{\pm 0.08}$	$94.10_{\pm 0.20}$	$50.56_{\pm 15.22}$	$51.94_{\pm 8.99}$	$69.28_{\pm 3.27}$	$42.94_{\pm 2.45}$
	TIACBM (ours)	$\mathbf{91.20}_{\pm 0.20}$	$\mathbf{85.65}_{\pm 0.10}$	$\mathbf{94.61}_{\pm 0.08}$	$\mathbf{77.32}_{\pm 9.33}$	$\mathbf{60.60}_{\pm 7.37}$	$\mathbf{69.94}_{\pm 1.98}$	$\mathbf{44.20}_{\pm 2.67}$
RoBERTa$_{\text{base}}$	Conventional	$90.55_{\pm 0.18}$	$84.49_{\pm 0.11}$	$94.56_{\pm 0.09}$	$89.20_{\pm 3.01}$	$76.92_{\pm 4.64}$	$67.42_{\pm 2.90}$	$38.30_{\pm 4.32}$
	Constant	$90.49_{\pm 0.11}$	$85.10_{\pm 0.30}$	$94.88_{\pm 0.26}$	$92.44_{\pm 0.51}$	$78.84_{\pm 2.61}$	$64.00_{\pm 4.33}$	$33.62_{\pm 5.41}$
	Poesina et al. (2024)	$90.52_{\pm 0.14}$	$79.89_{\pm 0.34}$	$94.81_{\pm 0.23}$	$90.18_{\pm 1.03}$	$78.70_{\pm 2.13}$	$65.16_{\pm 1.26}$	$33.36_{\pm 0.95}$
	Ankner et al. (2024)	$90.42_{\pm 0.09}$	$85.33_{\pm 0.17}$	$94.24_{\pm 0.19}$	$91.76_{\pm 1.53}$	$80.50_{\pm 2.07}$	$65.10_{\pm 0.49}$	$37.64_{\pm 2.78}$
	Cyclic Decaying	$90.70_{\pm 0.14}$	$84.74_{\pm 0.20}$	$94.70_{\pm 0.14}$	$91.36_{\pm 1.24}$	$78.84_{\pm 2.21}$	$67.80_{\pm 2.89}$	$39.36_{\pm 2.63}$
	TIACBM (ours)	$\mathbf{91.06}_{\pm 0.19}$	$\mathbf{85.93}_{\pm 0.18}$	$\mathbf{95.04}_{\pm 0.18}$	$\mathbf{93.98}_{\pm 0.70}$	$\mathbf{83.78}_{\pm 2.55}$	$\mathbf{68.38}_{\pm 1.53}$	$\mathbf{41.86}_{\pm 2.15}$

Table 1: Results on text classification (Reuters-21578, 20 Newsgroups), sentiment analysis (SST2) and authorship attribution (PAN19), with BERT and RoBERTa. Cochran's Q testing confirms that the results of TIACBM are always statistically better than conventional fine-tuning (p-value<0.001). The top scores for each architecture and metric are highlighted in bold.

Model	Fine-Tuning Strategy	SST2	PAN19-P1		PAN19-P5	
		Accuracy	Accuracy	Macro F1	Accuracy	Macro F1
GPT-2	Conventional	$92.35_{\pm 0.30}$	$82.78_{\pm 3.06}$	$67.96_{\pm 3.98}$	$62.88_{\pm 3.64}$	$38.16_{\pm 3.93}$
	Constant	$92.54_{\pm 0.43}$	$59.10_{\pm 1.81}$	$41.58_{\pm 0.76}$	$54.54_{\pm 6.95}$	$32.08_{\pm 2.30}$
	Poesina et al. (2024)	$92.27_{\pm 0.21}$	$72.44_{\pm 6.97}$	$57.60_{\pm 7.02}$	$59.54_{\pm 6.95}$	$32.86_{\pm 2.94}$
	Ankner et al. (2024)	$92.60_{\pm 0.09}$	$83.76_{\pm 1.46}$	$68.70_{\pm 2.13}$	$62.42_{\pm 0.39}$	$36.90_{\pm 2.40}$
	Cyclic Decaying	$92.74_{\pm 0.05}$	$82.38_{\pm 3.51}$	$67.34_{\pm 4.57}$	$65.00_{\pm 3.92}$	$36.46_{\pm 2.81}$
	TIACBM (ours)	$\mathbf{92.96}_{\pm 0.14}$	$\mathbf{85.90}_{\pm 1.96}$	$\mathbf{73.44}_{\pm 3.81}$	$\mathbf{68.20}_{\pm 2.25}$	$\mathbf{42.90}_{\pm 2.56}$

Table 2: Results on sentiment analysis (SST2) and authorship attribution (PAN19), with GPT-2. Cochran's Q testing confirms that the results of TIACBM are always statistically better than conventional fine-tuning (p-value<0.001). The top score for each metric is highlighted in bold.

BERT and RoBERTa. On SST2, TIACBM provides an increase of 1.23% over the baseline, and 0.61% over the state-of-the-art method of Poesina et al. (2024), suggesting that masking the harder (subjective) words towards the easier (objective) words, in a cyclic fashion, improves the performance of the model on sentiment analysis. On Reuters-21578, our approach boosts the micro F_1 score by 0.59%, reaching a top result of 91.48%, surpassing the state-of-the-art model based on CB-NTR (Huang et al., 2021). On 20 Newsgroups, TIACBM outperforms the baseline by 1.44% and Cart-Stra-CL++ (Poesina et al., 2024) by 6.04%, highlighting the effectiveness of content word masking. For PAN19, TIACBM increases both accuracy and macro F_1 score by 2.36% and 15.84%, respectively, demonstrating robust generalization even with limited training data. Compared with the approach of Ankner et al. (2024), we observe that TIACM brings higher performance gains across all tasks, mainly due to the task-specific information harnessed by our approach.

We present additional experiments with GPT-2 in Table 2. The results obtained with GPT-2 are consistent with those obtained with BERT and RoBERTa, confirming that TIACBM leads to significant performance gains. Hence, the results reported in Table 2 indicate that TIACBM is not limited to masked language models, being a generic approach that can be applied to any LLM.

5 Conclusion

We proposed a novel task-informed anti-curriculum by masking approach (TIACBM), and we evaluated its effectiveness on three tasks: sentiment analysis, text categorization by topic, and authorship attribution. The proposed method leverages information about the downstream tasks to decide which tokens to select for masking in a novel anti-curriculum by masking framework. On all the three tasks, our method achieved better results across all experiments, outperforming both baselines and state-of-the-art methods. Moreover, our method performed well on both multi-label and multi-class classification, while also proving resilience against imbalanced data sets, such as Reuters-21578. Additionally, we also showed that TIACBM is effective in scenarios with a low number of training samples, as in the case of PAN19.

Acknowledgments. This research is supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

6 Limitations

We present a novel method to consistently improve the performance of language models on downstream tasks. However, there is no universal anti-curriculum (masking ratio) schedule that can work for all models or data sets, representing an important parameter to be optimized by the user. Still, a general empirical statement proven in our work is that cycling anti-curriculum schedulers are superior in NLP, when it comes to curriculum by masking. Additionally, our method computes token relevance using a task-specific approach and this can be challenging to design for some tasks. We show on two tasks that attention weights can effectively serve this purpose.

7 Ethics Statement

To our knowledge, the proposed method poses no immediate risk. However, it can be adapted for generative modeling, which raises concerns about its potential misuse for malicious purposes, such as fake content generation.

References

Zachary Ankner, Naomi Saphra, Davis Blalock, Jonathan Frankle, and Matthew Leavitt. 2024. Dynamic masking rate schedules for MLM pretraining. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 477–487. Association for Computational Linguistics.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 2200–2024. European Language Resources Association.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 41–48. ACM.

Ernie Chang, Hui-Syuan Yeh, and Vera Demberg. 2021. Does the order of training samples matter? improving neural data-to-text generation with curriculum learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 727–733. Association for Computational Linguistics.

Florinel-Alin Croitoru, Nicolae-Cătălin Ristea, Radu Tudor Ionescu, and Nicu Sebe. 2025. Learning rate curriculum. *International Journal of Computer Vision*, 133(1):291–314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics.

Meng Fang, Tianyi Zhou, Yali Du, Lei Han, and Zhengyou Zhang. 2019. Curriculum-guided Hindsight Experience Replay. In *Proceedings the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 12623–12634. Curran Associates, Inc.

Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. 2017. Reverse curriculum generation for reinforcement learning. In *Proceedings of the 1st Annual Conference on Robot Learning (CoRL)*, volume 78, pages 482–495. PMLR.

Yantao Gong, Cao Liu, Jiazhen Yuan, Fan Yang, Xunliang Cai, Guanglu Wan, Jiansong Chen, Ruiyao Niu, and Houfeng Wang. 2021. Density-based dynamic curriculum learning for intent detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*, pages 3034–3037.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8153–8161. Association for Computational Linguistics.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. 2020. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5900–5909. IEEE.

Andrei Jarca, Florinel-Alin Croitoru, and Radu Tudor Ionescu. 2024. CBM: Curriculum by Masking. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, volume 392, pages 314–321. IOS Press.

Mike Kestemont. 2014. Function Words in Authorship Attribution. From Black Magic to Theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66. Association for Computational Linguistics.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In *Working Notes Papers of the CLEF 2019 Evaluation Labs*, volume 2380 of *CEUR Workshop Proceedings*.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum Learning and Minibatch Bucketing in Neural Machine Translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 379–386. INCOMA Ltd.

Ken Lang. 1995. NewsWeeder: Learning to Filter Netnews. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*, pages 331–339. Morgan Kaufmann Publishers Inc.

David Lewis. 1987. Reuters-21578 Text Categorization Collection. UCI Machine Learning Repository.

Cao Liu, Shizhu He, Kang Liu, and Jun Zhao. 2018. Curriculum Learning for Natural Answer Generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4223–4229. International Joint Conferences on Artificial Intelligence Organization.

Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. 2022. ACPL: Anti-curriculum Pseudo-labelling for Semi-supervised Medical Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20706. IEEE.

Jinglin Liu, Yi Ren, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020a. Task-level curriculum learning for non-autoregressive neural machine translation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 3861–3867. International Joint Conferences on Artificial Intelligence Organization.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020b. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 427–436. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Neelu Madan, Nicolae-Cătălin Ristea, Kamal Nasrollahi, Thomas B Moeslund, and Radu Tudor Ionescu. 2024. CL-MAE: Curriculum-Learned Masked Autoencoders. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2492–2502. IEEE.

Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. 2023. Length-Based Curriculum Learning for Efficient Pre-training of Language Models. *New Generation Computing*, 41(1):109–134.

Sanmit Narvekar, Jivko Sinapov, Matteo Leonetti, and Peter Stone. 2016. Source task creation for curriculum learning. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS)*, pages 566–574. International Foundation for Autonomous Agents and Multiagent Systems.

Harsh Nilesh Pathak and Randy Paffenroth. 2019. Parameter continuation methods for the optimization of deep neural networks. In *Proceedings of the 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1637–1643.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1162–1172. Association for Computational Linguistics.

Eduard Gabriel Poesina, Cornelia Caragea, and Radu Tudor Ionescu. 2024. A Novel Cartography-Based Curriculum Learning Method Applied on RoNLI: The First Romanian Natural Language Inference Corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 236–253. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Samarth Sinha, Animesh Garg, and Hugo Larochelle. 2020. Curriculum by smoothing. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 21653–21664. Curran Associates Inc.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642. Association for Computational Linguistics.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130:1526–1565.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith,

Model	Data Set	r_1	r_K	K
BERT	Reuters	0.15	0.09	3
	20 News	0.15	0.09	3
	SST2	0.15	0.09	3
	PAN19-P1	0.35	0.00	3
	PAN19-P5	0.30	0.00	3
RoBERTa	Reuters	0.15	0.09	3
	20 News	0.15	0.00	3
	SST2	0.15	0.09	3
	PAN19-P1	0.35	0.00	3
	PAN19-P5	0.30	0.00	3
GPT-2	SST2	0.15	0.09	3
	PAN19-P1	0.35	0.00	3
	PAN19-P5	0.30	0.00	3

Table 3: Masking ratios used in our experiments.

and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293. Association for Computational Linguistics.

Alexander Wettig, Tianyu Gao, Zhexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2985–3000. Association for Computational Linguistics.

Dongjie Yang, Zhuosheng Zhang, and Hai Zhao. 2023. Learning better masking for better language model pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7255–7267. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F Wong, and Lidia S Chao. 2021. Meta-Curriculum Learning for Domain Adaptation in Neural Machine Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 14310–14318.

A Hyperparameter Setup

We fix the decaying masking ratio schedule $\mathbf{r} = \{r_1, \ldots, r_K\}$, that is employed in Algorithm 1, through validation. For each architecture, we mention the maximum and minimum masking ratios in Table 3, as well as the length K. We cycle through the schedules every K epochs. We emphasize that the number of masking ratios K determines the cycle length, i.e. the masking ratios represent the curriculum probabilities that are cycled. In other words, the length of the masking ratio vector is equal to the cycle length.

B Ablation Studies

Ablating the masking ratios. In Tables 4 and 5, we ablate the interval of masking ratios r_1-r_K, while maintaining the cycle length $K = 3$. The reported results indicate that TIACBM maintains its performance gains for various intervals. This observation suggests that suboptimal hyperparameter choices can still lead to considerable improvements, attesting the robustness of TIACBM.

Model	r_1-r_K	SST2 Accuracy
BERT$_{\text{base}}$	0.13-0.11	$94.27_{\pm 0.12}$
	0.14-0.10	$94.53_{\pm 0.09}$
	0.15-0.09	$\mathbf{94.61}_{\pm 0.08}$
	0.16-0.08	$94.58_{\pm 0.06}$
RoBERTa$_{\text{base}}$	0.13-0.11	$94.82_{\pm 0.18}$
	0.14-0.10	$94.95_{\pm 0.21}$
	0.15-0.09	$95.04_{\pm 0.18}$
	0.16-0.08	$\mathbf{95.06}_{\pm 0.17}$

Table 4: Ablation study for intervals of masking ratios between r_1 and r_K on SST2. The top score for each architecture is highlighted in bold.

Data Set	Model	r_1-r_K	Accuracy	Macro F1
PAN19-P1	BERT$_{\text{base}}$	0.35-0.00	$77.32_{\pm 9.33}$	$\mathbf{60.60}_{\pm 7.37}$
		0.37-0.02	$76.99_{\pm 8.83}$	$58.89_{\pm 7.64}$
		0.33-0.02	$\mathbf{77.62}_{\pm 8.21}$	$59.94_{\pm 7.23}$
		0.30-0.05	$68.52_{\pm 15.72}$	$57.01_{\pm 8.81}$
	RoBERTa$_{\text{base}}$	0.35-0.00	$\mathbf{93.98}_{\pm 0.70}$	$\mathbf{83.78}_{\pm 2.55}$
		0.37-0.02	$93.23_{\pm 1.36}$	$82.83_{\pm 2.66}$
		0.33-0.02	$92.74_{\pm 1.01}$	$83.15_{\pm 3.21}$
		0.30-0.05	$91.23_{\pm 1.06}$	$81.51_{\pm 2.32}$
PAN19-P5	BERT$_{\text{base}}$	0.30-0.00	$\mathbf{69.94}_{\pm 1.98}$	$\mathbf{44.20}_{\pm 2.67}$
		0.32-0.02	$69.45_{\pm 1.83}$	$43.33_{\pm 2.76}$
		0.28-0.02	$68.92_{\pm 2.12}$	$44.02_{\pm 2.34}$
		0.35-0.05	$68.84_{\pm 2.67}$	$42.20_{\pm 2.05}$
	RoBERTa$_{\text{base}}$	0.30-0.00	$68.38_{\pm 1.53}$	$41.86_{\pm 2.15}$
		0.32-0.02	$\mathbf{69.84}_{\pm 1.54}$	$42.23_{\pm 5.71}$
		0.28-0.02	$68.72_{\pm 1.89}$	$41.33_{\pm 2.48}$
		0.35-0.05	$67.61_{\pm 2.55}$	$\mathbf{42.33}_{\pm 6.03}$

Table 5: Ablation study for intervals of masking ratios between r_1 and r_K on PAN19. The top scores for each architecture and metric are highlighted in bold.

Ablating the cycle length. We perform additional experiments by ablating the cycle length K between 2 and 5, while keeping the minimum and maximum masking ratios fixed. The corresponding results are shown in Table 6. While there are signs of sensitivity to the value of K, the results are generally above the conventional fine-tuning strategy. We conclude that K should be carefully tuned for optimal results.

C Curriculum versus Anti-Curriculum

In Tables 7 and 8, we compare our anti-curriculum method (TIACBM) with a reversed masking ratio schedule, which implements an easy-to-hard curriculum learning. Both approaches benefit from a cyclic schedule and task-specific information. The anti-curriculum approach consistently outperforms

Model	K	SST2	PAN19-P1		PAN19-P5	
		Accuracy	Accuracy	Macro F1	Accuracy	Macro F1
BERT$_{base}$	2	$93.97_{\pm 0.11}$	$66.68_{\pm 14.06}$	$55.52_{\pm 5.95}$	$66.28_{\pm 1.13}$	$42.64_{\pm 1.51}$
	3	$\mathbf{94.61}_{\pm 0.08}$	$\mathbf{77.32}_{\pm 9.33}$	$\mathbf{60.60}_{\pm 7.37}$	$\mathbf{69.94}_{\pm 1.98}$	$\mathbf{44.20}_{\pm 2.67}$
	4	$94.28_{\pm 0.22}$	$73.40_{\pm 10.04}$	$59.87_{\pm 6.97}$	$69.17_{\pm 1.94}$	$43.99_{\pm 1.89}$
	5	$94.55_{\pm 0.12}$	$72.80_{\pm 8.24}$	$58.80_{\pm 8.92}$	$68.78_{\pm 4.10}$	$41.24_{\pm 5.21}$
RoBERTa$_{base}$	2	$94.68_{\pm 0.16}$	$93.37_{\pm 1.18}$	$81.93_{\pm 3.63}$	$67.89_{\pm 2.32}$	$40.98_{\pm 3.31}$
	3	$\mathbf{95.04}_{\pm 0.18}$	$\mathbf{93.98}_{\pm 0.70}$	$\mathbf{83.78}_{\pm 2.55}$	$68.38_{\pm 1.53}$	$41.86_{\pm 2.15}$
	4	$94.91_{\pm 0.21}$	$92.85_{\pm 1.20}$	$83.41_{\pm 2.30}$	$\mathbf{71.30}_{\pm 3.66}$	$\mathbf{45.97}_{\pm 3.15}$
	5	$95.02_{\pm 0.16}$	$92.62_{\pm 1.39}$	$82.38_{\pm 1.56}$	$68.93_{\pm 2.14}$	$41.50_{\pm 1.89}$

Table 6: Ablation study for the hyperparameter K on SST2 and PAN19. The top scores for each architecture and metric are highlighted in bold.

Model	Fine-Tuning Strategy	Reuters	20 News	SST2	PAN19-P1		PAN19-P5	
		Micro F1		Accuracy	Accuracy	Macro F1	Accuracy	Macro F1
BERT$_{base}$	TICBM	$90.53_{\pm 0.15}$	$85.38_{\pm 0.08}$	$94.03_{\pm 0.16}$	$64.43_{\pm 3.19}$	$52.17_{\pm 8.35}$	$63.90_{\pm 2.51}$	$35.23_{\pm 3.56}$
	TIACBM	$\mathbf{91.20}_{\pm 0.20}$	$\mathbf{85.65}_{\pm 0.10}$	$\mathbf{94.61}_{\pm 0.08}$	$\mathbf{77.32}_{\pm 9.33}$	$\mathbf{60.60}_{\pm 7.37}$	$\mathbf{69.94}_{\pm 1.98}$	$\mathbf{44.20}_{\pm 2.67}$
RoBERTa$_{base}$	TICBM	$90.35_{\pm 0.09}$	$85.45_{\pm 0.13}$	$94.92_{\pm 0.14}$	$90.93_{\pm 1.22}$	$78.83_{\pm 0.98}$	$67.92_{\pm 0.95}$	$39.26_{\pm 1.61}$
	TIACBM	$\mathbf{91.06}_{\pm 0.19}$	$\mathbf{85.93}_{\pm 0.18}$	$\mathbf{95.04}_{\pm 0.18}$	$\mathbf{93.98}_{\pm 0.70}$	$\mathbf{83.78}_{\pm 2.55}$	$\mathbf{68.38}_{\pm 1.53}$	$\mathbf{41.86}_{\pm 2.15}$

Table 7: Comparison between curriculum (easy-to-hard) and anti-curriculum (hard-to-easy) approaches applied to BERT and RoBERTa. Both methods benefit from a cyclic schedule and task-specific information. The top score for each metric is highlighted in bold.

its counterpart, validating our choice based on hard-to-easy curriculum.

D TIACBM without Linguistic Priors

Transferring task heuristics. To show that the domain-specific heuristics can generalize across domains, we conduct experiments with the sentiment heuristic (originally applied on SST2) on authorship attribution. In Table 9, we show the corresponding results with BERT and RoBERTa. We observe that the sentiment heuristic is either close to the authorship heuristic, or even surpasses it in performance. This can be seen in the case of PAN19-P5, where the sentiment heuristic capitalizes on the data set design (fandom creative writing) and benefits from the numerous subjective words. Overall, the empirical evaluation indicates that the proposed task-specific heuristics can generalize across tasks. Furthermore, we observe that the sentiment heuristic consistently outperforms the conventional fine-tuning regime, suggesting that TIACBM can bring performance gains even when its heuristic is not aligned with the downstream task.

Generic versus task-specific heuristics. We next employ a generic task-agnostic heuristic, solely based on the attention scores of the fine-tuned model. The generic heuristic masks a number of tokens at each epoch, where the probability of masking a token is proportional to its average weight. We only utilize the attention weights from the first epoch, storing and loading them in later epochs. As shown in Table 10, the generic attention-based approach provides competitive results with our task-specific approaches, even surpassing them in a few cases. Task-specific heuristics provide better overall results for specific tasks, while the generic attention-based heuristic excels in some areas and lacks in others. The generic approach represents an alternative when there are no task-specific priors that can be leveraged.

Overall assessment. In summary, the experiments presented in Table 9 and 10 show that our text-based anti-curriculum by masking does not necessarily depend on prior task-specific information.

E Computational Resources

The experiments are carried out on a machine with 64GB of RAM, an AMD Ryzen 7 7800X3D CPU, and an Nvidia GeForce RTX 4090 GPU. Our most expensive experiments are performed on 20 Newsgroups, with 112 `mins` (3 `mins` per epoch) for TIACBM, and about 180 `mins` for decaying runs, due to the need of masking at every epoch. The masking step takes between 2-3 `mins`. In the case of SST2, the experiments require 6 `mins` per epoch, while the masking step requires 0.5 `mins`. The timetable for PAN19 varies from 0.5 `mins` per epoch to 1 `min` per epoch, regardless of the approach, while the masking takes roughly 3-4 `secs`. Finally, the masking step on Reuters-21578 takes between 1-3 `mins`, depending if non-content words need to be discarded or not. An average epoch takes 1.5 `mins`.

Model	Fine-Tuning Strategy	SST2	PAN19-P1		PAN19-P5	
		Accuracy	Accuracy	Macro F1	Accuracy	Macro F1
GPT-2	TICBM	$92.71_{\pm 0.28}$	$83.62_{\pm 3.73}$	$70.86_{\pm 5.36}$	$66.54_{\pm 3.14}$	$40.28_{\pm 4.53}$
	TIACBM	$\mathbf{92.96}_{\pm 0.14}$	$\mathbf{85.90}_{\pm 1.96}$	$\mathbf{73.44}_{\pm 3.81}$	$\mathbf{68.20}_{\pm 2.25}$	$\mathbf{42.90}_{\pm 2.56}$

Table 8: Comparison between curriculum (easy-to-hard) and anti-curriculum (hard-to-easy) approaches applied to GPT-2. The top score for each metric is highlighted in bold.

Model	Method	PAN19-P1		PAN19-P5	
		Accuracy	Macro F1	Accuracy	Macro F1
$\text{BERT}_{\text{base}}$	Conventional	$69.76_{\pm 15.11}$	$58.24_{\pm 6.06}$	$66.10_{\pm 1.56}$	$36.10_{\pm 4.22}$
	Sentiment Heuristic	$76.62_{\pm 8.05}$	$59.44_{\pm 3.48}$	$\mathbf{70.20}_{\pm 0.36}$	$\mathbf{45.33}_{\pm 2.90}$
	Authorship Heuristic (Original)	$\mathbf{77.32}_{\pm 9.33}$	$\mathbf{60.60}_{\pm 7.37}$	$69.94_{\pm 1.98}$	$44.20_{\pm 2.67}$
$\text{RoBERTa}_{\text{base}}$	Conventional	$89.20_{\pm 3.01}$	$76.92_{\pm 4.64}$	$67.42_{\pm 2.90}$	$38.30_{\pm 4.32}$
	Sentiment Heuristic	$92.06_{\pm 1.52}$	$81.72_{\pm 1.81}$	$\mathbf{69.68}_{\pm 5.50}$	$\mathbf{46.40}_{\pm 5.71}$
	Authorship Heuristic (Original)	$\mathbf{93.98}_{\pm 0.70}$	$\mathbf{83.78}_{\pm 2.55}$	$68.38_{\pm 1.53}$	$41.86_{\pm 2.15}$

Table 9: Results obtained by transferring our sentiment heuristic (originally applied on SST2) to authorship attribution (PAN19-P1 and PAN19-P5), for both BERT and RoBERTa. The task-specific heuristics used by TIACBM can be applied across tasks. The top scores for each architecture and metric are highlighted in bold.

Model	Heuristic	SST2	PAN19-P1		PAN19-P5	
		Accuracy	Accuracy	Macro F1	Accuracy	Macro F1
$\text{BERT}_{\text{base}}$	Conventional	$93.38_{\pm 0.14}$	$69.76_{\pm 15.11}$	$58.24_{\pm 6.06}$	$66.10_{\pm 1.56}$	$36.10_{\pm 4.22}$
	Generic	$94.07_{\pm 0.18}$	$69.78_{\pm 16.83}$	$58.24_{\pm 11.61}$	$\mathbf{71.06}_{\pm 2.46}$	$\mathbf{47.14}_{\pm 3.75}$
	Task-specific	$\mathbf{94.61}_{\pm 0.08}$	$\mathbf{77.32}_{\pm 9.33}$	$\mathbf{60.60}_{\pm 7.37}$	$69.94_{\pm 1.98}$	$44.20_{\pm 2.67}$
$\text{RoBERTa}_{\text{base}}$	Conventional	$94.56_{\pm 0.09}$	$89.20_{\pm 3.01}$	$76.92_{\pm 4.64}$	$67.42_{\pm 2.90}$	$38.30_{\pm 4.32}$
	Generic	$94.83_{\pm 0.16}$	$91.73_{\pm 0.56}$	$80.75_{\pm 1.06}$	$\mathbf{71.96}_{\pm 0.61}$	$\mathbf{42.96}_{\pm 2.97}$
	Task-specific	$\mathbf{95.04}_{\pm 0.18}$	$\mathbf{93.98}_{\pm 0.70}$	$\mathbf{83.78}_{\pm 2.55}$	$68.38_{\pm 1.53}$	$41.86_{\pm 2.15}$

Table 10: Results with the generic (task-agnostic) attention-based heuristic versus task-specific heuristics. The task-agnostic heuristic can be effective when task-specific information is not available. The top scores for each architecture and metric are highlighted in bold.