MASKER: Masked Keyword Regularization for Reliable Text Classification

Seung Jun Moon[*1], **Sangwoo Mo**[*1], **Kimin Lee**[2†], **Jaeho Lee**[1], **Jinwoo Shin**[1]
[1]Korea Advanced Institute of Science and Technology, South Korea
[2]University of California, Berkeley, USA
{june1212,swmo,jaeho-lee,jinwoos}@kaist.ac.kr kiminlee@berkeley.edu

Abstract

Pre-trained language models have achieved state-of-the-art accuracies on various text classification tasks, *e.g.*, sentiment analysis, natural language inference, and semantic textual similarity. However, the *reliability* of the fine-tuned text classifiers is an often underlooked performance criterion. For instance, one may desire a model that can detect out-of-distribution (OOD) samples (drawn far from training distribution) or be robust against domain shifts. We claim that one central obstacle to the reliability is the over-reliance of the model on a limited number of keywords, instead of looking at the whole context. In particular, we find that (a) OOD samples often contain in-distribution keywords, while (b) cross-domain samples may not always contain keywords; over-relying on the keywords can be problematic for both cases. In light of this observation, we propose a simple yet effective fine-tuning method, coined masked keyword regularization (MASKER), that facilitates context-based prediction. MASKER regularizes the model to reconstruct the keywords from the rest of the words and make low-confidence predictions without enough context. When applied to various pre-trained language models (*e.g.*, BERT, RoBERTa, and ALBERT), we demonstrate that MASKER improves OOD detection and cross-domain generalization without degrading classification accuracy. Code is available at https://github.com/alinlab/MASKER.

1 Introduction

Text classification (Aggarwal and Zhai 2012) is a classic yet challenging problem in natural language processing (NLP), having a broad range of applications, including sentiment analysis (Bakshi et al. 2016), natural language inference (Bowman et al. 2015), and semantic textual similarity (Agirre et al. 2012). Recently, Devlin et al. (2019) have shown that fine-tuning a pre-trained language model can achieve state-of-the-art performances on various text classification tasks without any task-specific architectural adaptations. Thereafter, numerous pre-training and fine-tuning strategies to improve the classification accuracy further have been proposed (Liu et al. 2019; Lan et al. 2020; Sanh et al. 2019; Clark et al. 2020; Sun et al. 2019; Mosbach, Andriushchenko, and Klakow 2020; Zhang et al. 2020). However, a vast majority of the works have focused on evaluating the accuracy of the models only and overlooked their *reliability* (Hendrycks et al. 2020), *e.g.*, robustness to out-of-distribution (OOD) samples drawn far from the training data (or in-distribution samples).

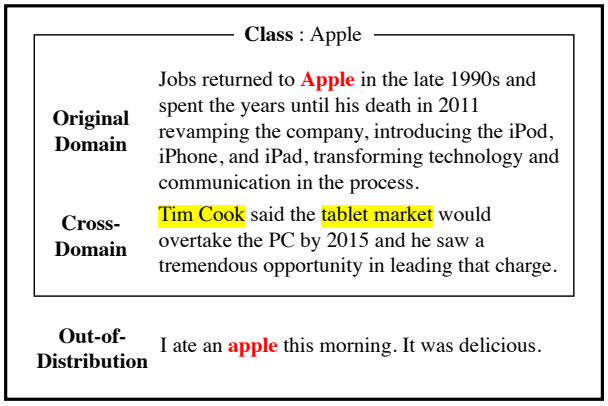
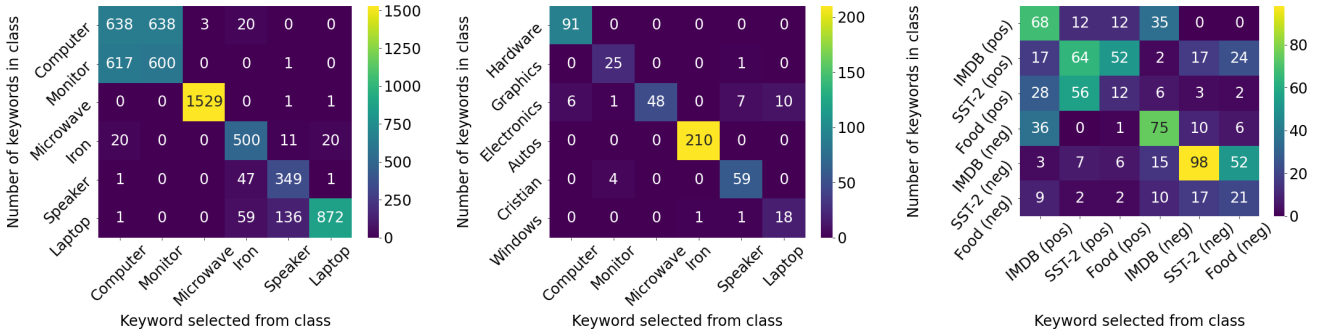


Figure 1: Out-of-distribution (OOD) and cross-domain examples, where class 'Apple' is the original domain. The OOD sample contains the word 'apple' (red) but in a different context. The cross-domain sample does not share the words (*e.g.*, 'Tim Cook') with the original domain, but it still contains some clues (yellow) to guess the correct class.

While *pre-trained* language models are known to be robust in some sense (Hendrycks et al. 2020), we find that *fine-tuned* models suffer from the over-reliance problem, *i.e.*, making predictions based on only a limited number of domain-specific keywords instead of looking at the whole context. For example, consider a classification task of 'Apple' visualized in Figure 1. If the most in-distribution samples contain the keyword 'Apple,' the fine-tuned model can predict the class solely based on the existence of the keyword. However, a reliable classifier should detect that the sentence "I ate an apple this morning" is an out-of-distribution sample (Hendrycks and Gimpel 2017; Shu, Xu, and Liu 2017; Tan et al. 2019). On the other hand, the sentence "Tim Cook said that ..." should be classified as the topic 'Apple' although it does not contain the keyword 'Apple' and the keyword 'Tim Cook' is not contained in the training samples. In other words, the reliable classifier should learn decision rules that

[*]Equal contribution
[†]Work was done while the author was at KAIST



(a) In-distribution (b) Out-of-distribution (c) Cross-domain

Figure 2: Frequency of the keywords selected from the source class (x-axis) in the target class (y-axis). (a) Both the source and target classes are in-distribution, (b) source and target distributions are in- and out-of-distribution, respectively, and (c) source and target distributions are identical but of multiple domains. In (b), one can see that OOD classes often contain the same keywords from the (similar but different) in-distribution classes, *e.g.*, see 'Iron' and 'Autos.' In (c), one can see that the keywords in one domain do not perfectly align to the other domain, *e.g.*, see 'IMDB (neg)' and 'Food (pos).'

generalize across domains (Fei and Liu 2015; Bhatt, Semwal, and Roy 2015; Bhatt, Sinha, and Roy 2016).

This problematic phenomenon frequently happens in real-world datasets. To verify this, we extract the keywords from Amazon 50 class reviews (Chen and Liu 2014) dataset and sentiment analysis datasets (IMDB (Maas et al. 2011); SST-2 (Socher et al. 2013); Fine Food (McAuley and Leskovec 2013)), following the attention-based scheme illustrated in Section 2.1. Figure 2 shows the frequency of the keywords selected from the source class in the target class. Figure 2a shows that the keywords are often strongly tied with the class, which leads the model to learn a shortcut instead of the context. Figure 2b shows the results where the source and target classes are different classes of the Amazon reviews dataset. Here, OOD classes often contain the same keywords from the in-distribution classes, *e.g.*, the class 'Autos' contains the same keywords as the class 'Iron.' On the other hand, Figure 2c shows the results where both source and target classes are sentiments ('pos' and 'neg') classes in IMDB, SST-2, and Fine Food datasets. While the same sentiment shares the same keywords, the alignment is not perfect; *e.g.*, 'IMDB (neg)' and 'Food (pos)' contain the same keywords.

1.1 Contribution

We propose a simple yet effective fine-tuning method coined masked keyword regularization (MASKER), which handles the over-reliance (on keywords) problem and facilitates the context-based prediction. In particular, we introduce two regularization techniques: (a) masked keyword reconstruction and (b) masked entropy regularization. First, (a) forces the model to predict the masked keywords from understanding the context around them. This is inspired by masked language modeling from BERT (Devlin et al. 2019), which is known to be helpful for learning context. Second, (b) penalizes making high-confidence predictions from "cut-out-context" sentences, that non-keywords are randomly dropped, in a similar manner of Cutout (DeVries and Taylor 2017) used for regularizing image classification models. We also suggest two keyword selection schemes, each relying on dataset statistics and attention scores. We remark that all proposed techniques of MASKER can be done in an *unsupervised* manner.

We demonstrate that MASKER, applied to the pre-trained language models: BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and ALBERT (Lan et al. 2020), significantly improves the OOD detection and cross-domain generalization performance, without degrading the classification accuracy. We conduct OOD detection experiments on 20 Newsgroups (Lang 1995), Amazon 50 class reviews (Chen and Liu 2014), Reuters (Lewis et al. 2004), IMDB (Maas et al. 2011), SST-2 (Socher et al. 2013), and Fine Food (McAuley and Leskovec 2013) datasets, and cross-domain generalization experiments on sentiment analysis (Maas et al. 2011; Socher et al. 2013; McAuley and Leskovec 2013), natural language inference (Williams, Nangia, and Bowman 2017), and semantic textual similarity (Wang et al. 2019) tasks. In particular, our method improves the area under receiver operating characteristic (AUROC) of BERT from 87.0% to 98.6% for OOD detection under 20 Newsgroups to SST-2 task, and reduce the generalization gap from 19.2% to 10.9% for cross-domain generalization under Fine Food to IMDB task.

1.2 Related Work

Distribution shift in NLP. The reliable text classifier should detect distribution shift, *i.e.*, test distribution is different from the training distribution. However, the most common scenarios: OOD detection and cross-domain generalization are relatively under-explored in NLP domains (Hendrycks et al. 2020; Marasović 2018). Hendrycks et al. (2020) found that pre-trained models are robust to the distribution shift compared to traditional NLP models. We find that the pre-trained models are not robust enough, and we empirically show that pre-trained models are still relying on undesirable dataset bias. Our method further improves the generalization performance, applied to the pre-trained models.

Shortcut bias. One may interpret the over-reliance problem as a type of shortcut bias (Geirhos et al. 2020), *i.e.*, the model learns an easy-to-learn but not generalizable solution, as the keywords can be considered as a shortcut. The shortcut bias is investigated under various NLP tasks (Sun et al. 2019), *e.g.*, natural language inference (McCoy, Pavlick, and Linzen 2019), reasoning comprehension (Niven and Kao 2019), and question answering (Min et al. 2019). To our best knowledge, we are the first to point out that the over-reliance on keywords can also be a shortcut, especially for text classification. We remark that the shortcut bias is not always harmful as it can be a useful feature for in-distribution accuracy. However, we claim that they can be problematic for unexpected (*i.e.*, OOD) samples, as demonstrated in our experiments.

Debiasing methods. Numerous debiasing techniques have been proposed to regularize shortcuts, *e.g.*, careful data collection (Choi et al. 2018; Reddy, Chen, and Manning 2019), bias-tailored architecture (Agrawal et al. 2018), and adversarial regularization (Clark, Yatskar, and Zettlemoyer 2019; Minderer et al. 2020; Nam et al. 2020). However, most prior work requires supervision of biases, *i.e.*, the shortcuts are explicitly given. In contrast, our method can be viewed as an unsupervised debiasing method, as our keyword selection schemes automatically select the keywords.

2 Masked Keyword Regularization

We first introduce our notation and architecture setup; then propose the keyword selection and regularization approaches in Section 2.1 and Section 2.2, respectively.

Notation. The text classifier $f : \mathbf{x} \mapsto y$ maps a document $\mathbf{x}$ to the corresponding class $y \in \{1, \dots, C\}$. The document $\mathbf{x}$ is a sequence of tokens $t_i \in \mathcal{V}$, *i.e.*, $\mathbf{x} = [t_1, \dots, t_T]$ where $\mathcal{V}$ is the vocabulary set and T is the length of the document. Here, the full corpus $\mathcal{D} = \{(\mathbf{x}, y)\}$ is a collection of all documents, and the class-wise corpus $\mathcal{D}_c = \{(\mathbf{x}, y) \in \mathcal{D} \mid y = c\}$ is a subset of $\mathcal{D}$ of class c. The keyword set $\mathcal{K} \subset \mathcal{V}$ is the set of vocabularies which mostly affects to the prediction.[1] The keyword $\mathbf{k} = [k_1, \dots, k_L]$ of the document $\mathbf{x}$ is given by $\mathbf{k} = [t_i \in \mathbf{x} \mid t_i \in \mathcal{K}]$, where $L \leq T$ is the number of keywords in the document $\mathbf{x}$.

Architecture. We assume the pre-trained language model follows the bi-directional Transformer (Vaswani et al. 2017) architecture, widely used in recent days (Devlin et al. 2019; Liu et al. 2019; Lan et al. 2020). They consist of three components: embedding network, document classifier, and token-wise classifier. Given document $\mathbf{x}$, the embedding network produces (a) a document embedding (for an entire document), and (b) token embeddings, which correspond to each input token. The document and token-wise classifier predict the class of document and tokens, respectively, from the corresponding embeddings. For the sake of simplicity, we omit the shared embedding network and denote the document and token-wise classifier as $f_{\texttt{doc}} : \mathbf{x} \mapsto y$ and $f_{\texttt{tok}} : \mathbf{x} = [t_1, \dots, t_T] \mapsto \mathbf{s} = [s_1, \dots, s_T]$, respectively, where $s_i \in \mathcal{V}$ is a target token corresponds to t_i.

[1] Chosen by our proposed keyword selection (Section 2.1).

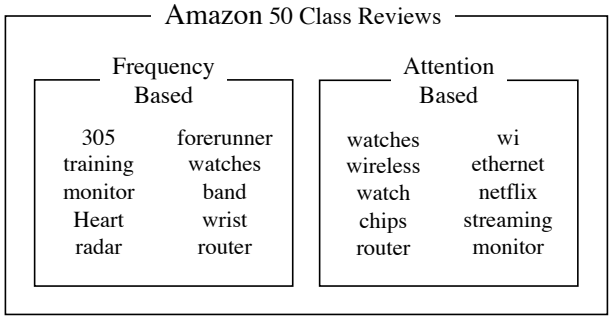


Figure 3: Top 10 keywords chosen from the frequency-based and attention-based selection schemes under the Amazon 50 class reviews dataset. The frequency-based scheme chooses uninformative words (*e.g.*, '305'), while the attention-based scheme chooses more informative ones (*e.g.*, 'watch').

2.1 Keyword Selection Schemes

We consider two keyword selection schemes, based on the dataset statistics (model-free) and trained models. While the former is computationally cheaper, the latter performs better; hence, one can choose its purpose.

Frequency-based. We first choose the keywords using the relative frequency of the words in the dataset. Specifically, we use the term frequency-inverse document frequency (TF-IDF; Robertson (2004)) metric, which measures the importance of the token by comparing the frequency in the target documents (term frequency) and the entire corpus (inverse document frequency). Here, the keywords are defined as the tokens with the highest TF-IDF scores. Formally, let $\mathbf{X}_c$ be a large document that concatenates all tokens in a class-wise corpus $\mathcal{D}_c$, and $\mathbf{D} = [\mathbf{X}_1, \dots, \mathbf{X}_C]$ be a corpus of such large documents. Then, the frequency-based score of token t is given by

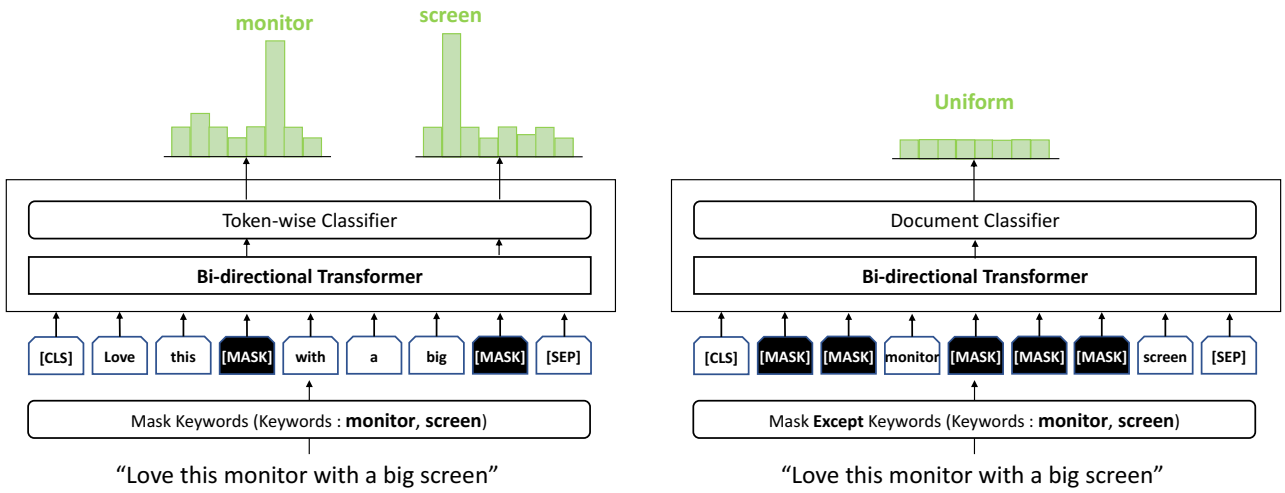
$$s^{\texttt{freq}}(t) = \max_{c \in \{1,\dots,C\}} \texttt{tf}(t, \mathbf{X}_c) \cdot \texttt{idf}(t, \mathbf{D}) \qquad (1)$$

where $\texttt{tf}(t, \mathbf{X}) = 0.5 + 0.5 \cdot n_{t,\mathbf{X}} / \max\{n_{t',\mathbf{X}} : t' \in \mathbf{X}\}$, $\texttt{idf}(t, \mathbf{D}) = \log(|\mathbf{D}|/|\{\mathbf{X} \in \mathbf{D} : t \in \mathbf{X}\}|)$, and $n_{t,\mathbf{x}}$ is number of token t in document $\mathbf{x}$. Note that the frequency-based selection is model-agnostic and easily computed, but does not reflect the contribution of the words to the prediction.

Attention-based. We also choose the keywords using the model attention as it is a more direct and effective way to measure the importance of words on model prediction. To this end, we first train a model with a standard approach using the cross-entropy loss $\mathcal{L}_{\texttt{CE}}$, which leads the model to suffer from the over-reliance (on keywords) issue. Our idea is to use the attention values of the model for choosing the keywords. Here, the keywords are defined as the tokens with the highest attention values. Formally, let $\mathbf{a} = [a_1, \dots, a_T] \in \mathbb{R}^T$ be attention values of the document embedding, where a_i corresponds to the input token t_i. Then, the attention-based score of token t is given by

$$s^{\texttt{attn}}(t) = \sum_{(\mathbf{x},y) \in \mathcal{D}} \frac{1}{n_{t,\mathbf{x}}} \sum_{i \in \{1,\dots,T\}} \mathbb{I}(t_i = t) \cdot \frac{a_i}{\|\mathbf{a}\|} \qquad (2)$$

where $\mathbb{I}$ is an indicator function and $\|\cdot\|$ is ℓ_2-norm.



(a) Masked keyword reconstruction

(b) Masked entropy regularization

Figure 4: Illustration of two portions of our proposed method, MASKER: (a) *Masked keyword reconstruction* masks keyword tokens in input sentences and forces the model to predict the original words in masked tokens. (b) *Masked entropy regularization* masks non-keyword tokens in input sentences and forces the model to print uniform distribution, as regarding it as OOD.

We choose the keywords by picking the top K tokens according to the scores in Eq. (1) and Eq. (2) for each selection scheme, respectively. We also test the class-balanced version, *i.e.*, pick the top K/C tokens for each class, but the class-agnostic one performed better.

Comparison of the selection schemes. We observe that the frequency-based scheme often selects uninformative keywords that uniquely appears in some class. In contrast, the attention-based scheme selects more general keywords that actually influence the prediction. Figure 3 shows the keywords chosen by both selection schemes: the frequency-based scheme chooses uninformative words such as '305' and 'forerunner,' while the attention-based scheme chooses more informative ones such as 'watch' or 'chips.'

2.2 Regularization via Keyword Masking

Using the chosen keywords, we propose two regularization techniques to reduce the over-reliance issue and facilitate the model to look at the contextual information.

Masked keyword reconstruction. To enforce the model to look at the surrounding context, we guide the model to reconstruct the keywords from keyword-masked documents. Note that it resembles the masked language model (Devlin et al. 2019), but we mask the *keywords* instead of random words. Masked keyword reconstruction only regularizes sentences with keywords, and we omit the loss for ones without any keywords. Formally, let $\tilde{\mathbf{k}}$ be a random subset of the full keyword $\mathbf{k}$ (selected as in Section 2.1), that each element is chosen with probability p independently. We mask $\tilde{\mathbf{k}}$ from the original document $\mathbf{x}$ and get the masked document $\tilde{\mathbf{x}} = \mathbf{x} - \tilde{\mathbf{k}}$. Then, the masked keyword reconstruction (MKR) loss is

$$\mathcal{L}_{\texttt{MKR}}(\tilde{\mathbf{x}}, v) := \sum_{i \in \texttt{index}(\tilde{\mathbf{k}})} \mathcal{L}_{\texttt{CE}}(f_{\texttt{tok}}(\tilde{\mathbf{x}})_i, v_i) \qquad (3)$$

where $\texttt{index}(\tilde{\mathbf{k}})$ is the index of the keywords $\tilde{\mathbf{k}}$ with respect to the original document $\mathbf{x}$, v_i is the index of the keywords with respect to the vocabulary set. We remark that the *reconstruction* part is essential; we also test simply augmenting the masked documents, *i.e.*, $\mathcal{L}_{\texttt{CE}}(f_{\texttt{doc}}(\tilde{\mathbf{x}}), y)$, but it performed worse. Choosing proper keywords is also crucial; attention-based keywords performs better than frequency-based or random keywords, as shown in Table 1 and Table 3.

Masked entropy regularization. Furthermore, we regularize the prediction of the *context*-masked documents, that context (non-keyword) words are randomly dropped. The model should not classify the context-masked documents correctly as they lost the original context. Formally, let $\widehat{\mathbf{c}}$ be a randomly chosen subset of the full context words $\mathbf{c} = \mathbf{x} - \mathbf{k}$, where each element is chosen with probability q independently. We mask $\widehat{\mathbf{c}}$ from the original document $\mathbf{x}$ and get the context-masked document $\widehat{\mathbf{x}} = \mathbf{x} - \widehat{\mathbf{c}}$. Then, the masked entropy regularization (MER) loss is

$$\mathcal{L}_{\texttt{MER}}(\widehat{\mathbf{x}}) := D_{\text{KL}}(\mathcal{U}(y) || f_{\texttt{doc}}(\widehat{\mathbf{x}})) \qquad (4)$$

where D_{KL} is the KL-divergence and $\mathcal{U}(y)$ is a uniform distribution. We remark that MER does not degrade the classification accuracy since it regularizes non-realistic context-masked sentences, rather than full documents. Table 1 shows that MER does not drop the classification accuracy in original domain, while Table 3 and Table 4 show that MER improves the cross-domain accuracy. On the other hand, MER differs from the prior sentence-level objectives, *e.g.*, next sentence prediction (Devlin et al. 2019), as our goal is to regularize shortcuts, not learning better in-domain representation.

To sum up, the final objective is given by

$$\mathcal{L}_{\texttt{total}} = \mathcal{L}_{\texttt{CE}} + \lambda_{\texttt{MKR}}\mathcal{L}_{\texttt{MKR}} + \lambda_{\texttt{MER}}\mathcal{L}_{\texttt{MER}} \qquad (5)$$

where $\lambda_{\texttt{MKR}}$ and $\lambda_{\texttt{MER}}$ are hyperparameters for the MKR and MER losses, respectively. Figure 4 visualizes the proposed losses, and the overall procedure is in Appendix B.

Method	Classifier	Keyword	MKR	MER	AUROC ↑	EER ↓	Detection Accuracy ↑	Classification Accuracy ↑
OC-SVM	-	-	-	-	57.01±1.08	41.12±0.90	72.99±4.69	-
OpenMax	-	-	-	-	53.02±3.74	55.21±2.03	74.01±0.58	79.01±0.51
DOC	-	-	-	-	75.12±3.06	25.55±4.12	80.21±0.76	83.14±0.82
BERT	Multi-class	-	-	-	78.46±1.16	28.30±1.11	81.19±0.62	84.76±0.35
	1-vs-rest	-	-	-	80.56±0.84	25.71±0.71	80.78±0.48	84.66±0.26
BERT +MASKER (ours)	1-vs-rest	Random	✓	-	80.52±0.72	25.42±0.38	80.83±0.33	85.06±0.14
	1-vs-rest	Random	-	✓	81.51±0.27	24.17±0.87	82.13±0.58	84.89±0.71
	1-vs-rest	Frequency	✓	-	81.32±0.78	25.12±1.18	81.33±0.67	83.35±1.77
	1-vs-rest	Frequency	-	✓	82.54±0.54	23.88±0.76	82.72±0.58	85.25±0.70
	1-vs-rest	Attention	✓	-	83.33±0.44	24.55±0.67	82.11±0.68	**85.27±0.33**
	1-vs-rest	Attention	-	✓	82.60±0.51	25.02±0.83	80.99±0.65	85.02±0.29
	1-vs-rest	Attention	✓	✓	**85.48±0.54**	**22.30±1.20**	**85.02±0.55**	85.15±0.95

Table 1: Ablation study on OOD detection under the Amazon 50 class reviews. We use 25% of classes as in-distribution, and the rest as OOD. The reported results are averaged over five trials, subscripts denote standard deviations, and the best results are highlighted in bold. All components of our method contribute to the OOD detection performance (%).

ID	OOD	OC-SVM	DOC	Vanilla / Residual / MASKER		
				BERT	RoBERTa	ALBERT
Newsgroup	Amazon	62.1	84.1	85.4/86.7/**87.0**(+1.9%)	85.3/85.9/**87.2**(+2.3%)	86.7/85.4/**89.4**(+3.1%)
	Reuters	53.9	60.0	91.8/93.0/**97.7**(+6.4%)	93.1/92.1/**93.9**(+0.8%)	93.3/92.0/**94.7**(+1.6%)
	IMDB	59.8	88.6	94.6/95.7/**98.5**(+4.1%)	95.2/95.9/**97.7**(+2.7%)	94.5/92.6/**96.6**(+2.2%)
	SST-2	63.0	88.1	87.0/97.0/**98.6**(+13.3%)	94.7/94.9/**98.2**(+3.6%)	95.8/95.6/**98.1**(+2.4%)
	Fine Food	62.8	81.3	85.3/87.3/**93.4**(+9.5%)	88.7/89.4/**92.9**(+4.7%)	77.6/86.6/**91.6**(+18.1%)
Amazon	Newsgroup	61.3	81.3	84.8/83.9/**87.2**(+2.8%)	87.9/86.6/**91.0**(+3.5%)	87.3/85.0/**88.4**(+1.3%)
	Reuters	55.5	79.8	89.7/89.7/**93.5**(+4.2%)	92.3/92.6/**93.6**(+1.4%)	93.1/93.5/**94.5**(+1.5%)
	IMDB	66.2	89.6	93.3/92.8/**95.2**(+2.0%)	90.1/87.0/**93.3**(+3.5%)	89.9/88.6/**95.6**(+6.4%)
	SST-2	60.9	91.5	93.0/88.9/**95.6**(+2.8%)	92.4/94.8/**96.4**(+4.4%)	93.4/91.9/**96.9**(+3.7%)
	Fine Food	51.1	66.8	78.5/77.7/**84.9**(+8.2%)	74.9/80.0/**80.7**(+7.8%)	82.6/86.3/**87.3**(+5.7%)

Table 2: AUROC (%) on various OOD detection scenarios. The reported results are averaged over three trials, and the best results are highlighted in bold. Bracket denotes the relative gain of MASKER over the vanilla model.

3 Experiments

We demonstrate the effectiveness of our proposed method, MASKER. In Section 3.1, we describe the experimental setup. In Section 3.2 and 3.3, we present the results on OOD detection and cross-domain generalization, respectively.

3.1 Experimental Setup

We demonstrate the effectiveness of MASKER, applied to the pre-trained models: BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019) and ALBERT (Lan et al. 2020). We choose $10 \times C$ keywords in a class agnostic way, where C is the number of classes. We drop the keywords and contexts with probability $p = 0.5$ and $q = 0.9$ for all our experiments. We use $\lambda_{\texttt{MKR}} = 0.001$ and $\lambda_{\texttt{MER}} = 0.001$ for OOD detection, and same $\lambda_{\texttt{MKR}} = 0.001$ but $\lambda_{\texttt{MER}} = 0.0001$ for cross-domain generalization, as the entropy regularization gives more gain for reliability than accuracy (Pereyra et al. 2017). We modify the hyperparameter settings of the pre-trained models (Devlin et al. 2019; Liu et al. 2019), specified in Appendix A.

1-vs-rest classifier. Complementary to MASKER, we use 1-vs-rest classifier (Shu, Xu, and Liu 2017) as it further improves the reliability (see Table 1 and Table 3). Intuitively, 1-vs-rest classifier can reject all classes (all prediction scores are low); hence detect OOD samples well.

Baselines. We mainly compare MASKER with vanilla fine-tuning of pre-trained models (Hendrycks et al. 2020), with extensive ablation study (see Table 1 and Table 3). Additionally, we compare with residual ensemble (Clark, Yatskar, and Zettlemoyer 2019), applied to the same pre-trained models. Residual ensemble trains a debiased model by fitting the residual from a biased model. We construct a biased dataset by subsampling the documents that contain keywords. To benchmark the difficulty of the task, we also report the classic non-Transformer models, *e.g.*, one-class support vector machine (OC-SVM, Schölkopf et al. (2000)), OpenMax (Bendale and Boult 2016), and DOC (Shu, Xu, and Liu 2017).

3.2 OOD Detection

We use the highest softmax (or sigmoid) output of the model as confidence score for OOD detection task. We use 20 Newsgroups (Lang 1995) and Amazon 50 class reviews (Chen and Liu 2014) datasets for in-distribution, and Reuters (Lewis

Method	Classifier	Keyword	MKR	MER	Dataset (Train → Test)					
					IMDB → SST-2	IMDB → Food	SST-2 → IMDB	SST-2 → Food	Food → SST-2	Food → IMDB
BERT	Multi-class	-	-	-	85.92 (-7.57)	92.90 (-0.60)	85.74 (-6.74)	87.57 (-4.91)	67.55 (-28.92)	77.31 (-19.16)
	1-vs-rest	-	-	-	84.28 (-8.92)	87.81 (-5.39)	85.34 (-7.46)	84.35 (-8.45)	64.57 (-32.15)	81.34 (-12.16)
BERT +MASKER (ours)	1-vs-rest	Random	✓	-	87.29 (-6.29)	90.52 (-3.06)	86.57 (-7.60)	78.00 (-16.17)	78.79 (-17.51)	84.56 (-12.91)
	1-vs-rest	Random	-	✓	86.84 (-5.97)	90.27 (-2.54)	87.18 (-8.52)	85.91 (-9.79)	79.50 (-17.03)	84.61 (-11.92)
	1-vs-rest	Frequency	✓	-	86.52 (-6.04)	88.41 (-4.15)	87.06 (-7.37)	79.99 (-14.44)	74.72 (-23.73)	80.94 (-17.51)
	1-vs-rest	Frequency	-	✓	86.38 (-6.72)	84.20 (-8.90)	85.31 (-13.55)	88.43 (-10.43)	75.34 (-20.66)	85.34 (-10.66)
	1-vs-rest	Attention	✓	-	87.50 (-5.91)	92.03 (-5.74)	87.78 (-4.92)	**90.12 (-2.58)**	75.57 (-20.82)	79.32 (-17.07)
	1-vs-rest	Attention	-	✓	87.71 (-5.59)	90.39 (-2.63)	84.92 (-7.64)	87.21 (-5.35)	75.80 (-20.87)	82.13 (-14.54)
	1-vs-rest	Attention	✓	✓	**88.02 (-5.44)**	**93.58 (+0.12)**	**88.43 (-3.89)**	89.21 (-3.11)	**80.02 (-16.44)**	**85.57 (-10.90)**

Table 3: Ablation study on cross-domain generalization under sentiment analysis task. The reported results are averaged over five trials, bracketed numbers denote the generalization gap from the training domain accuracy, and the best accuracies are highlighted in bold. All components of our method contribute to the cross-domain accuracy (%).

et al. 2004), IMDB (Maas et al. 2011), and SST-2 (Socher et al. 2013) datasets for out-of-distribution.

Table 1 shows an ablation study on MASKER under the Amazon reviews dataset with a split ratio of 25%. All components of MASKER contribute to OOD detection. Note that MASKER does not degrade the classification accuracy while improving OOD detection. Also, the attention-based selection performs better than the frequency-based or random selection, which implies the importance of selecting suitable keywords. Recall that the attention-based scheme selects the keywords that contribute to the prediction, while the frequency-based scheme often chooses domain-specific keywords that are not generalizable across domains.

Table 2 shows the results on various OOD detection scenarios, comparing the vanilla fine-tuning, residual ensemble, and MASKER. Notably, MASKER shows the best results in all cases. In particular, MASKER improves the area under receiver operating characteristic (AUROC) from 87.0% to 98.6% on 20 Newsgroups to SST-2 task. We find that residual ensemble shows inconsistent gains: it often shows outstanding results (*e.g.*, Newsgroup to SST-2) but sometimes fails (*e.g.*, Amazon to Fine Food). In contrast, MASKER shows consistent improvement over the vanilla fine-tuning.

In Figure 5a and Figure 5b, we visualize the t-SNE (Maaten and Hinton 2008) plots on the document embeddings of BERT and MASKER, under the Amazon reviews dataset with a split ratio of 25%. Blue and red points indicate in- and out-of-distribution samples, respectively. Unlike the samples that are entangled in the vanilla BERT, MASKER clearly distinguishes the OOD samples.

3.3 Cross-domain Generalization

We conduct the experiments on sentiment analysis (IMDB (Maas et al. 2011); SST-2 (Socher et al. 2013); Fine Food (McAuley and Leskovec 2013)), natural language inference (MNLI, Williams, Nangia, and Bowman (2017)), and semantic textual similarity (STS-B, Wang et al. (2019) dataset) tasks, following the settings of Hendrycks et al. (2020).

Table 3 shows an ablation on MASKER study under the sentiment analysis task. The results are consistent with OOD detection, *e.g.*, all components contribute to cross-domain generalization. Notably, while MER is not helpful for the original domain accuracy (see Table 1), it improves the cross-domain accuracy for most settings. In particular, MASKER improves the cross-domain accuracy from 75.6% to 80.0% for Fine Food to SST-2 task. We analyze the most influential keywords (see Appendix D) and find that MASKER extracts the sentiment-related (*i.e.*, generalizable) keywords (*e.g.*, 'astonishing') while the vanilla BERT is biased to some domain-specific words (*e.g.*, 'moonlight').

Table 4 presents the results on sentiment analysis, natural language inference, and semantic textual similarity tasks. We compare MASKER with the vanilla fine-tuning and residual ensemble. The residual ensemble helps cross-domain generalization, but the gain is not significant and often degrades the original domain accuracy. This is because the keywords can be useful features for classification. Hence, naïvely removing (or debiasing) those features may lose the information. In contrast, MASKER facilitates contextual information rather than removing the keyword information, which regularizes the over-reliance in a softer manner.

In Figure 5c and Figure 5d, we provide the t-SNE plots on

Train	Test	OpenMax	DOC	Vanilla / Residual / MASKER		
				BERT	RoBERTa	ALBERT
IMDB	IMDB	87.7	88.0	93.5/92.8/93.5	95.3/94.0/95.6	91.6/91.4/90.8
	SST-2	79.6	77.9	85.9/86.9/**88.1**(+2.6%)	89.7/90.2/**91.8**(+2.3%)	89.8/89.0/**89.9**(+0.1%)
	Fine Food	75.4	78.3	92.9/92.5/**93.6**(+0.8%)	92.6/87.7/**93.0**(+0.4%)	87.1/87.8/**92.1**(+5.7%)
SST-2	SST-2	82.9	83.1	92.5/90.3/92.3	94.5/92.0/94.3	91.9/90.9/91.5
	IMDB	75.3	76.9	85.7/86.2/**88.4**(+3.2%)	86.0/85.7/**87.3**(+1.5%)	83.0/83.0/**83.8**(+1.0%)
	Fine Food	62.2	64.5	87.6/88.2/**89.2**(+1.8%)	86.5/87.8/**89.6**(+3.6%)	79.0/80.6/**84.9**(+7.5%)
Fine Food	Fine Food	93.6	93.3	96.5/94.8/96.5	96.9/95.7/97.1	95.5/95.4/96.6
	IMDB	67.5	67.3	77.3/81.1/**85.6**(+10.7%)	84.1/84.6/**86.6**(+2.9%)	74.7/80.4/**84.8**(+13.5%)
	SST-2	61.9	62.0	67.6/67.8/**80.0**(+18.3%)	78.5/80.1/**83.8**(+6.8%)	71.7/73.2/**83.3**(+16.2%)

(a) Sentiment analysis

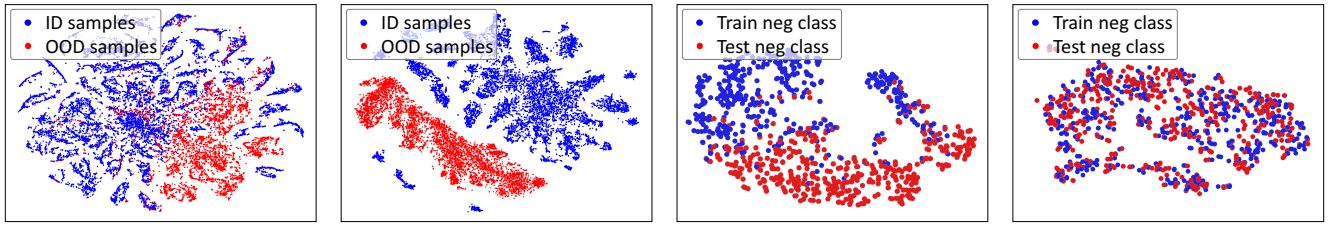
Model	Telephone (ID)	Telephone Letters (OOD)	Telephone Face-to-Face (OOD)
BERT	80.5	77.4	77.5
+MASKER	80.2	**80.4** (+3.9%)	**78.5** (+1.3%)
RoBERTa	84.8	83.6	83.5
+MASKER	86.0	**85.9** (+2.8%)	**83.8** (+0.4%)
ALBERT	82.9	82.8	80.9
+MASKER	82.5	**84.0** (+1.5%)	**86.8** (+7.3%)

(b) Natural language inference

Model	MSRvid (ID)	MSRvid Images (OOD)	MSRvid Headlines (OOD)	Images (ID)	Images MSRvid (OOD)	Images Headlines (OOD)
BERT	91.5	82.0	61.7	88.0	89.7	73.9
+MASKER	91.2	**84.3** (+2.8%)	**66.7** (+8.1%)	88.1	**91.6** (+2.1%)	**75.3** (+1.9%)
RoBERTa	94.2	**88.0**	80.3	91.8	92.9	84.1
+MASKER	93.7	**88.0** (+0.0%)	**84.0** (+4.6%)	91.3	**94.1** (+1.3%)	**85.3** (+1.4%)
ALBERT	92.6	81.2	60.6	90.4	90.9	69.8
+MASKER	93.3	**82.6** (+1.7%)	**68.8** (+13.5%)	90.5	**92.0** (+1.2%)	**78.4** (+12.3%)

(c) Semantic textual similarity

Table 4: Accuracy (%) of original domain and cross-domain on (a) sentiment analysis, (b) natural language inference, and (c) semantic textual similarity tasks, respectively.The reported results are averaged over three trials for sentiment analysis and semantic textual similarity, and a single trial for natural language inference. Bold denotes the best results among the three methods, and bracket denotes the relative gain of MASKER over the vanilla model.



(a) BERT (b) MASKER (ours) (c) BERT (d) MASKER (ours)

Figure 5: t-SNE plots on the document embeddings of BERT and MASKER, on (a,b) OOD detection (Amazon 50 class reviews with split ratio 25%), and (c,d) cross-domain generalization (Fine Food to SST-2). (a,b) Blue and red dots indicate the in- and out-of-distribution samples, respectively. (c,d) Blue and red dots indicate the samples from the same classes ('negative') from training and test domains, respectively. MASKER better distinguishes OOD samples and entangles cross-domain samples.

the document embeddings of BERT and MASKER, under the Fine Food to STS-2 task. Blue and red points indicate original and cross-domain samples, respectively. MASKER better entangles the same classes in training and test datasets (of the different domains) while BERT fails to do so.

4 Conclusion

The reliability of text classifiers is an essential but under-explored problem. We found that the over-reliance on some keywords can be problematic for out-of-distribution detection and generalization. We propose a simple yet effective fine-tuning method, coined masked keyword regularization (MASKER), composed of two regularizers and keyword selection schemes to address this issue. We demonstrate the effectiveness of MASKER under various scenarios.

Acknowledgements

This work was supported by Center for Applied Research in Artificial Intelligence(CARAI) grant funded by Defense Acquisition Program Administration(DAPA) and Agency for Defense Development(ADD) (UD190031RD).

References

Aggarwal, C. C.; and Zhai, C. 2012. A survey of text classification algorithms. In *Mining text data*, 163–222. Springer.

Agirre, E.; Cer, D.; Diab, M.; and Gonzalez-Agirre, A. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 385–393.

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4971–4980.

Bakshi, R. K.; Kaur, N.; Kaur, R.; and Kaur, G. 2016. Opinion mining and sentiment analysis. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 452–455. IEEE.

Bendale, A.; and Boult, T. E. 2016. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1563–1572.

Bhatt, H. S.; Semwal, D.; and Roy, S. 2015. An Iterative Similarity based Adaptation Technique for Cross-domain Text Classification. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 52–61. Beijing, China: Association for Computational Linguistics. doi:10.18653/v1/K15-1006. URL https://www.aclweb.org/anthology/K15-1006.

Bhatt, H. S.; Sinha, M.; and Roy, S. 2016. Cross-domain Text Classification with Multiple Domains and Disparate Label Sets. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1641–1650. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1155. URL https://www.aclweb.org/anthology/P16-1155.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642.

Chen, Z.; and Liu, B. 2014. Mining topics in documents: standing on the shoulders of big data. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1116–1125.

Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; Yih, W.-t.; Choi, Y.; Liang, P.; and Zettlemoyer, L. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2174–2184.

Clark, C.; Yatskar, M.; and Zettlemoyer, L. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4069–4082.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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, Volume 1 (Long and Short Papers)*, 4171–4186.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* .

Fei, G.; and Liu, B. 2015. Social media text classification under negative covariate shift. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2347–2356.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut Learning in Deep Neural Networks. *arXiv preprint arXiv:2004.07780* .

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

Hendrycks, D.; Liu, X.; Wallace, E.; Dziedzic, A.; Krishnan, R.; and Song, D. 2020. Pretrained Transformers Improve Out-of-Distribution Robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2744–2751.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, 331–339. Elsevier.

Lewis, D. D.; Yang, Y.; Rose, T. G.; and Li, F. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research* 5(Apr): 361–397.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment

analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Marasović, A. 2018. NLP's generalization problem, and how researchers are tackling it. *The Gradient* .

McAuley, J. J.; and Leskovec, J. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, 897–908.

McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448.

Min, S.; Wallace, E.; Singh, S.; Gardner, M.; Hajishirzi, H.; and Zettlemoyer, L. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4249–4257.

Minderer, M.; Bachem, O.; Houlsby, N.; and Tschannen, M. 2020. Automatic Shortcut Removal for Self-Supervised Representation Learning. In *International Conference on Machine Learning*.

Mosbach, M.; Andriushchenko, M.; and Klakow, D. 2020. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. *arXiv preprint arXiv:2006.04884* .

Nam, J.; Cha, H.; Ahn, S.; Lee, J.; and Shin, J. 2020. Learning from Failure: Training Debiased Classifier from Biased Classifier. *arXiv preprint arXiv:2007.02561* .

Niven, T.; and Kao, H.-Y. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4658–4664.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. In *ICLR Workshop*.

Reddy, S.; Chen, D.; and Manning, C. D. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics* 7: 249–266.

Robertson, S. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation* .

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*.

Schölkopf, B.; Williamson, R. C.; Smola, A. J.; Shawe-Taylor, J.; and Platt, J. C. 2000. Support vector method for novelty detection. In *Advances in neural information processing systems*, 582–588.

Shu, L.; Xu, H.; and Liu, B. 2017. Doc: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2911–2916.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642.

Sun, C.; Qiu, X.; Xu, Y.; and Huang, X. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, 194–206. Springer.

Tack, J.; Mo, S.; Jeong, J.; and Shin, J. 2020. CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. In *Advances in Neural Information Processing Systems*.

Tan, M.; Yu, Y.; Wang, H.; Wang, D.; Potdar, S.; Chang, S.; and Yu, M. 2019. Out-of-Domain Detection for Low-Resource Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3566–3572.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Williams, A.; Nangia, N.; and Bowman, S. R. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* .

Zhang, T.; Wu, F.; Katiyar, A.; Weinberger, K. Q.; and Artzi, Y. 2020. Revisiting Few-sample BERT Fine-tuning. *arXiv preprint arXiv:2006.05987* .