# De-biased Attention Supervision for Text Classification with Causality

**Yiquan Wu[1*], Yifei Liu[2*], Ziyu Zhao[1], Weiming Lu[1†], Yating Zhang[3], Changlong Sun[3], Fei Wu[1], Kun Kuang[1†]**

[1] College of Computer Science and Technology, Zhejiang University, China
[2] College of Software Technology, Zhejiang University, China
[3] Alibaba Group, China
{wuyiquan,liuyifei,luwm,kunkuang}@zju.edu.cn, benzhao.styx@gmail.com, ranran.zyt@alibaba-inc.com, changlong.scl@taobao.com, wufei@cs.zju.edu.cn,

## Abstract

In text classification models, while the unsupervised attention mechanism can enhance performance, it often produces attention distributions that are puzzling to humans, such as assigning high weight to seemingly insignificant conjunctions. Recently, numerous studies have explored Attention Supervision (AS) to guide the model toward more interpretable attention distributions. However, such AS can impact classification performance, especially in specialized domains. In this paper, we address this issue from a causality perspective. Firstly, we leverage the causal graph to reveal two biases in the AS: 1) Bias caused by the label distribution of the dataset. 2) Bias caused by the words' different occurrence ranges that some words can occur across labels while others only occur in a particular label. We then propose a novel De-biased Attention Supervision (DAS) method to eliminate these biases with causal techniques. Specifically, we adopt backdoor adjustment on the label-caused bias and reduce the word-caused bias by subtracting the direct causal effect of the word. Through extensive experiments on two professional text classification datasets (e.g., medicine and law), we demonstrate that our method achieves improved classification accuracy along with more coherent attention distributions.

| Method | Legal Dataset | | | |
|---|---|---|---|---|
| | Acc | Ma-P | Ma-R | Ma-F |
| BiLSTM+Attention | 85.50 | **81.68** | **80.48** | **80.59** |
| +AS(TF-IDF) | 85.36 | 80.13 | 80.30 | 79.76 |
| +AS(YAKE) | 84.31 | 79.21 | 78.13 | 78.16 |
| +AS(AdaKeyBERT) | **85.55** | 81.16 | 79.89 | 80.07 |
| +AS(VMask) | 84.83 | 78.91 | 79.50 | 78.83 |
| Method | Medical Dataset | | | |
| | Acc | Ma-P | Ma-R | Ma-F |
| BiLSTM+Attention | **77.05** | **67.56** | **62.54** | **63.88** |
| +AS(TF-IDF) | 76.83 | 67.18 | 62.03 | 63.51 |
| +AS(YAKE) | 76.10 | 64.32 | 60.42 | 61.38 |
| +AS(AdaKeyBERT) | 76.25 | 63.06 | 60.83 | 61.01 |
| +AS(VMask) | 75.86 | 62.16 | 60.12 | 60.18 |

Table 1: Pilot study on two professional datasets. The legal (medical) datasets aim to predict the charge (department) given the criminal fact (patient's question). BiLSTM+Attention is a common classification method that uses an unsupervised attention mechanism. TF-IDF, YAKE, AdaKeyBERT, and VMask are four common keyword selection methods.

## Introduction

Text classification stands as a fundamental task in Natural Language Processing (NLP) (Kowsari et al. 2019). When presented with a text, the classification model's objective is to predict the appropriate label. Owing to the prosperity of deep learning, the performance of text classification has been improved significantly (Gasparetto et al. 2022). Within deep learning-based text classification models, the attention mechanism has gained popularity due to its remarkable effectiveness (Du and Huang 2018; Sun and Lu 2020). The attention mechanism intends to give high weight to important information when aggregating the input information, which imitates the way humans make decisions. However, despite its contributions to performance improvement (Hu 2019), the unsupervised training of the attention mechanism frequently produces peculiar outcomes (e.g., assigning high weights to inconsequential connections or nouns). Such outcomes make it incomprehensible to humans and lead it astray from its original purpose (Jain and Wallace 2019).

Recently, several research on Attention Supervision (AS) are conducted to instill a sense of rationality into attention weights across various tasks such as Event Detection (Zhao et al. 2018), VQA (Qiao, Dong, and Xu 2018), and so on. In text classification, the most intuitive AS method is to impart the model the precise attention weight of each token within every sample (Barrett et al. 2018). Such a method necessitates meticulous attention annotation and proves impractical in real-world scenarios. A prevalent and adaptable strategy is to leverage keywords as supervision signals (e.g., construct a keyword vocabulary tailored to the task and assign high weights to these keywords) (Bao et al. 2018; Choi et al. 2020). However, as Fig. 1 and Tab. 1 show, though the attention distribution appears reasonable with AS, the classification performance drops in the professional domains where similar words can lead to different labels[1].

[*]These authors contributed equally.
[†]Corresponding authors.



[1]We use AS(TF-IDF) to represent the AS in the following since

| Dataset: *Legal* \| Label: *Intentional injury* | |
|---|---|
| w/o AS | The Procuratorate alleges that on March 25, 2014, at 9:00 a.m., the defendant A had a dispute with B in his village for some reason and caused a fight. During the fight, the defendant A injured B and his sons C and D. According to the identification of the physical evidence identification office of Public Security Bureau, the injury of B is Grade II minor injury, and the injury of C and D is minor injury. |
| w/ AS | The Procuratorate alleges that on March 25, 2014, at 9:00 a.m., the defendant A had a dispute with B in his village for some reason and caused a fight. During the fight, the defendant A injured B and his sons C and D. According to the identification of the physical evidence identification office of Public Security Bureau, the injury of B is Grade II minor injury, and the injury of C and D is minor injury. |

| Dataset: *Medical* \| Label: *Thoracic Surgery* | |
|---|---|
| w/o AS | I am a patient with bronchiectasis. I cough when I have a cold. Can I drink tea made from mulberry leaves? |
| w/ AS | I am a patient with bronchiectasis. I cough when I have a cold. Can I drink tea made from mulberry leaves? |

Figure 1: The visualization of attention distribution. w/o AS means the attention mechanism is trained without supervision, and w/ AS means the attention supervision. The color is simplified into four levels, the darker the color, the higher the weight. With attention supervision, the attention distribution become reasonable.

In this paper, we delve into this intriguing phenomenon from the perspective of causality (Pearl, Glymour, and Jewell 2016; Imbens and Rubin 2015). By constructing the corresponding causal graph, we reveal two biases in the AS that can affect the performance: **1) Label-caused bias:** The imbalanced distribution of labels will lead to the imbalance of keywords. For instance, if a keyword often co-occurs with high-frequency labels in training, once the keyword occurs in the inference time, the model will be inclined to predict the high-frequency label and thus ignore the low-frequency labels. **2) Word-caused bias:** Some words can occur across different labels, while some words only occur in a particular label. The different occurrence ranges can lead to different attention weights, downplaying the meaning of the word itself. The above two biases are difficult to separate and often ignored in the conventional attention mechanism.

In causal inference, such data-driven biases have been studied for years and several effective methods have been proposed (Pearl, Glymour, and Jewell 2016; Imbens and Rubin 2015). Notwithstanding, how to apply these casual tools within AS for text classification is still an open problem.

To counteract these biases, we propose a novel De-bias Attention Supervision (DAS) method. Specifically, there are two main de-bias operations: 1) Use the backdoor adjustment to cut the edges related to label-caused bias in the causal graph. Besides, the word-caused bias is thus separated out. 2) Capture the direct causal effect of the word on the attention weights as the word-caused bias, and reduce it by subtracting the direct causal effect from the total causal effect. Extensive experiments on two professional text classification datasets prove the effectiveness of our DAS.

The contributions of this paper can be summarized as:

- We investigate the problem of Attention Supervision (AS) for the text classification task from the perspective of causality, and use the causal graph to reveal two biases (e.g., label-caused bias and word-caused bias) behind the AS, which affect the classification performance.
- We propose an end-to-end De-bias Attention Supervision (DAS) method to counteract these biases. DAS first adopts the backdoor adjustment to eliminate the label-caused bias, then it reduces the word-caused bias by subtracting the word's direct causal effect from its total causal effect.
- Experiments on two professional datasets (e.g., medicine and law) show that our method achieves more reasonable attention distribution and better classification accuracy.
- To motivate other scholars to investigate this problem, we make the code and data publicly available [2].

the four keyword selection methods have similar effects for AS.

## Related Work

### Text Classification

Text classification aims to assign the correct label to the input text, which has been widely studied (Hu 2019). With breakthroughs in Natural Language Processing (NLP), the text classification technique has been applied in many applications like spam detection (Tida and Hsu 2022), and comment filtering (J et al. 2021). The methods in the early years are rule-based that require manually extracted features. In recent years, deep learning has been proven to be effective in many domains (Shen et al. 2021, 2022, 2023; Zhou et al. 2022; Wu et al. 2022b, 2023; Liu et al. 2023; Zhang et al. 2023a,b, 2022; Lv et al. 2023b,a, 2022; Li et al. 2022). Thus, deep-learning text classification methods have also been proposed, which require far less labor and achieve better performance (Gasparetto et al. 2022). In this paper, we focus on deep-learning text classification methods that leverage the attention mechanism.

### Attention Mechanism

The attention mechanism, as introduced by (Bahdanau, Cho, and Bengio 2015), has found broad utilization within a diverse array of NLP models. Typically, an attention layer produces a distribution over input text(Sun and Lu 2020). This distribution subsequently serves as the foundation for creating a weighted combination of the input. The intention of the attention mechanism is to give high weights to important words, which mimic the process of human decision-making. In the domain of text classification, the attention mechanism is also widely used (Du and Huang 2018).

[2] https://github.com/6666ev/DAS

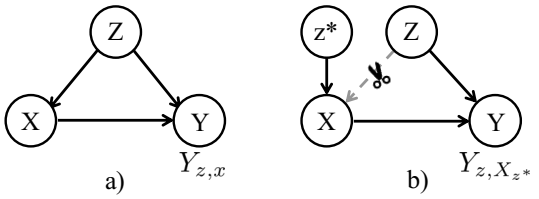


Figure 2: The illustration of the causal graph.

Despite facilitating the performance, unsupervised attention usually shows perplexing distributions (Jiang et al. 2018). To tackle this issue, an intuitive way is to make Attention Supervision (AS). In event detection, (Zhao et al. 2018) takes trigger words as the annotation of AS. In neural machine translation, (Mi, Wang, and Ittycheriah 2016) takes the target sentence as the guidance of AS. In text classification, however, there is no ready-made annotation. Bao et al. (2018) uses human annotated rationales to supervise the attention, which is high-cost and difficult to migrate the method to a new dataset. Choi et al. (2020) derive word importance by modifying its original weight and assessing the resultant impact on predicted output. However, this method is highly dependent on the quality of the original model. A more prevalent and pragmatic AS method is to construct a keyword vocabulary and subsequently utilize these keywords as supervision signals (Nguyen and Nguyen 2018). In this paper, we mainly focus on keyword-based AS.

### Causal Inference

Causal Inference (Pearl 2009) is a powerful statistical modeling technique to remove bias in data (Wu et al. 2022a). That bias might bring a spurious correlation or confounding effect among variables. Recently, many methods have been proposed to remove bias in the literature of causal inference, including counterfactual outcome prediction based on potential outcome framework (Imbens and Rubin 2015) and do-operation based on structure causal model (Pearl 2009; Wu et al. 2020). With do-operation, the backdoor adjustment (Pearl, Glymour, and Jewell 2016) has been proposed for data de-bias. In this paper, we investigate the biases in the AS from the perspective of causal inference and use effective casual methods to reduce these biases.

## Preliminaries

In this section, to be self-contained, we introduce the related concepts of causal inference (Pearl, Glymour, and Jewell 2016; Imbens and Rubin 2015).

**Causal graph** is a directed acyclic graph, where each node denotes a variable and each edge denotes a cause-and-effect relationship. For example, in Fig. 2a, $X \rightarrow Y$ means $X$ has an effect on $Y$.

**Confounding bias** occurs when a common cause affects both the explanatory $X$ and outcome variables $Y$. This path $X \leftarrow Z \rightarrow Y$ is called a **"backdoor"** because it provides an indirect route for bias, confounding the true causal relationship.

**Potential outcomes.** In accordance with the Neyman-Rubin causal model (Rubin 2005), an assignment of a variable is termed as "treatment". This framework introduces the concept of potential outcomes. As illustrated in Fig. 2a, for a given $Z = z$, the potential outcome of $Y$ can be written as $y = Y_{z,x} = Y(X = x, Z = z)$ where $x = X_z = X(Z = z)$. A notable situation is when the variable is subject to "no treatment" (e.g., the absence of drug administration in a drug trial). In these scenarios, it is denoted as $Z = z^*$.

**Causal effect** reflects the differences between two potential outcomes of the same variable under two different treatments. The total effect (TE) of $Z = z$ on $Y$ can be defined as the difference between $Y_{z,X_z}$ and $Y_{z^*,X_{z^*}}$:

$$TE = Y_{z,X_z} - Y_{z^*,X_{z^*}}. \tag{1}$$

TE can be decomposed into natural direct effect $Z \rightarrow Y$ (NDE) and total indirect effect $Z \rightarrow X \rightarrow Y$ (TIE). The NDE of $Z$ is the effect of $Z$ on $Y$ when $X$ is blocked. For example, in Fig. 2b, $X$ is kept as $X_{z^*}$ no matter what $Z$ is. The NDE is expressed as:

$$NDE = Y_{z,X_{z^*}} - Y_{z^*,X_{z^*}}. \tag{2}$$

The TIE is the difference between TE and NDE:

$$TIE = TE - NDE = Y_{z,X_z} - Y_{z,X_{z^*}}. \tag{3}$$

***do* operation** is used to see 'what will happen if we do something'. In Fig. 2a), $X$ is caused by $Z$, but we can $do(X)$ by assigning a value to the $X$ directly to cut off all the edges pointing to $X$ (e.g., $Z \rightarrow X$), which means $X$ is not caused by $Z$ here. The passive observation $P(Y|X)$ is thus changed to active intervention $P(Y|do(X))$, and we can obtain the true causal effect of $X$ on $Y$.

## Methodology

In this section, we first introduce the conventional text classification method with attention mechanism and Attention Supervision (AS). We then analyze the biases behind AS and propose the De-biased Attention Supervision (DAS).

### Text Classification with Attention

Fig. 3 shows the architecture of the conventional text classification model with the attention mechanism, which consists of an encoder, an attention layer, and a predictor.

**Encoder** Given the input text $Input = \{w_t\}_{t=1}^n$, where $n$ is the number of words, the encoder aims to transform it into a sequence of hidden states $H = \{h_t\}_{t=1}^n \in \mathbb{R}^{n \times d}$:

$$H = \text{Encode}(Input), \tag{4}$$

where $d$ is the dimension of the hidden state.

**Attention Mechanism** The attention mechanism aims to assign a weight to each token according to the query $q$. Given the hidden states $H$, the attention distribution $a \in \mathbb{R}^{n \times 1}$ is calculated as follow:

$$e_i = \tanh(h_i W_q + b_{attn}), \tag{5}$$

$$a = \text{softmax}(e), \tag{6}$$

where $W_q \in \mathbb{R}^{d \times 1}$, $b_{attn} \in \mathbb{R}^{n \times 1}$ are learnable parameters, and $W_q$ is the implementation of the query $q$. Next, the final representation of input $R \in \mathbb{R}^d$ is produced as $R = \sum_i a_i h_i$.

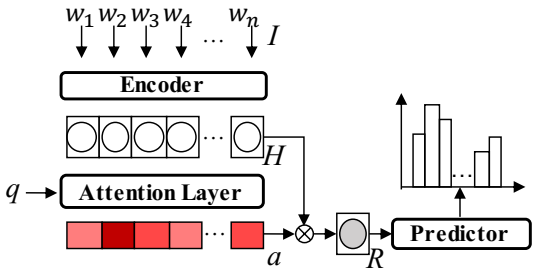


Figure 3: The architecture of conventional text classification model with the attention mechanism.

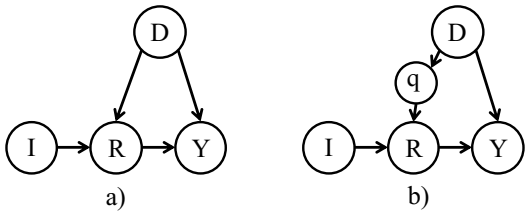


Figure 4: Causal graphs that illustrate label-caused bias.

**Predictor** For the predictor, given the representation of the input $R$, it will output the predicted probability $\hat{y} \in \mathbb{R}^m$, where $m$ is the number of labels. $\hat{y}$ is calculated as follow:

$$\hat{y} = \text{softmax}(W_p R + b_p), \tag{7}$$

where $W_p \in \mathbb{R}^{m \times d}$ and $b_p \in \mathbb{R}^{m \times 1}$ are the trainable parameters of the fully connected layer that calculates the logits of each label, and the softmax operation is used to transfer the logits into probability.

**Training and Inference** The cross-entropy loss has been proved suitable for text classification, which is calculated as:

$$\mathcal{L}_c = -\sum_{i=1}^{m}(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)), \tag{8}$$

where $\hat{y}_i \in [0, 1]$ is the predicted probability, and $y_i \in \{0, 1\}$ indicates the ground truth.

In the inference, the label with the highest predicted probability will be chosen as the output.

## Attention Supervision (AS)

As Fig. 1 shows, without supervision, attention will produce confusing distribution, and attention supervision is used to alleviate this phenomenon. Since the attention annotation of each token in every sample is hard to access, keyword-based supervision is a more practical method.

**Keyword Selection** The keyword is the word that has an impact on the results for a certain task. To obtain the keywords, there are several methods including TF-IDF, YAKE (Campos et al. 2020), AdaKeyBERT (Priyanshu and Vijay 2022), VMask (Chen and Ji 2020), and so on. With the pilot study shown in Tab. 1, these keyword selection methods have similar effects in AS, so we use take TF-IDF as the keyword selection method due to its simplicity and robustness.

The TF-IDF algorithm depends on two factors: term frequency (TF) and inverse document frequency (IDF). For the word $w$, given the sample $S_i$, the TF of the $w$ is defined as $TF_i^w = \frac{n_i^w}{n_i}$, where $n_i^w$ is the number of occurrences of the $w$ in $S_i$ and $n_i$ is the number of words in $S_i$. The inverse document frequency (IDF) is defined as $IDF^w = log\frac{|S|}{|\{i:w \in S_i\}|}$, where $|S|$ is the total number of samples and $|\{i : w \in S_i\}|$ is the number of samples that $w$ appears. The TF-IDF score of the word $w$ for sample $S_i$ is then obtained:

$$TF\text{-}IDF_i^w = TF_i^w * IDF^w. \tag{9}$$

The importance score of the word $w$ is $\sum_i TF\text{-}IDF_i^w$, and the keyword vocabulary $V$ can be obtained by selecting the high-score words.

**Supervision Signal** Next, we need to translate the keyword vocabulary $V$ to the supervision signals. Given the input text $I$, we check whether each word appears in $V$ and obtain the indication sequence $A \in \mathbb{R}^{n \times 1}$, where $A_i \in \{0, 1\}$ indicates the appearance of $i$-th word. The loss of attention supervision is calculated as:

$$\mathcal{L}_{AS} = -\sum_{i=1}^{n}((A_i \log(a_i) + (1 - A_i) \log(1 - a_i))). \tag{10}$$

The final loss $\mathcal{L}$ is the weighted sum of $\mathcal{L}_c$ and $\mathcal{L}_{AS}$:

$$\mathcal{L} = \mathcal{L}_c + \lambda * \mathcal{L}_{AS}, \tag{11}$$

where $\lambda$ is the preference weight.

## De-biased Attention Supervision (DAS)

As analyzed above, there exist biases in the AS, which can affect the performance of classification. In this section, we introduce how to address these biases from the view of causality and propose our DAS method.

### Label-caused Bias

**Causal Look** Fig. 4a models the label-caused bias. $D$ denotes the label distribution in the dataset. $I$ and $Y$ are the representation of input text and predicted labels, and $R$ is the representation of input after the attention mechanism.[3] From the figure, we can find a backdoor path: $R \leftarrow D \rightarrow Y$, which hinders the attainment of accurate results. To fully resolve the confounding bias, we have to block the edge from $D \rightarrow R$. Such dependence is difficult to cut off directly, but thanks to the query of attention mechanism, we find a handle. Taking the query $q$ into consideration, the causal graph is changed to Fig. 4b, where the cut off of $D \rightarrow R$ can be achieved by cutting off $D \rightarrow q$.

The query $q$ can be interpreted as the assumption of the model before classification. If we directly adopt AS, the assumption will favor labels with high frequency thus affecting

[3]The token embeddings are pre-trained from a task-independent corpus, so the label distribution $D$ does not influence the embedding of the input text $I$.

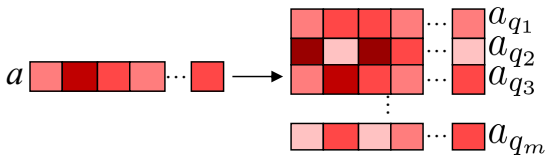


Figure 5: The implementation of the Equation 12.

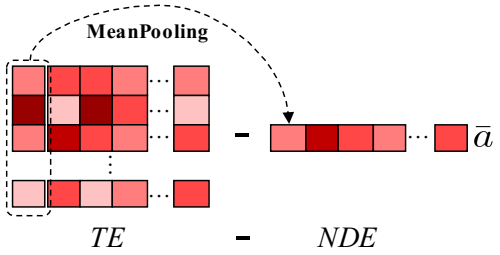


Figure 6: The implementation of the Equation 15.

the attention distribution through $q$. To remove such label-caused bias in AS, we make a backdoor adjustment, which is a main de-confounding technique (Pearl, Glymour, and Jewell 2016). Specifically, a *do* operation is taken on $R$ through $q$ by computing the interventional posterior:

$$P(Y|do(R)) = \sum_{q \in Q} P(Y|R,q)P(q), \tag{12}$$

where $Q \in \{q_1, q_2, ..., q_m\}$ represents the assumptions for each label. $P(q)$ is assigned to $\frac{1}{m}$ for each $q$.

**Implementation** The implementation of Equation 12 is depicted in Figure 5. Since $R$ is computed based on attention distributions, we utilize $q_1$ to $q_m$ to concurrently generate $m$ attention distributions $\{a_{q_1}, a_{q_2}, ..., a_{q_m}\} \in \mathbb{R}^{n \times m}$. Consequently, $P(Y|R,q)$ can be derived from the predictor. During training, in order to link each $q$ with a specific label in the context of AS, for a sample with label $y$, the model computes its attention supervision loss $\mathcal{L}_{AS}$ only through $a_{q_y}$ and masks the remaining $m-1$ attention distributions.

### Word-caused Bias

**Causal Look** For a word $w$ in $I$, the total effect (TE) of the $w$ on the $Y$ can be written as ($I$ is omitted here for the writing simplicity):

$$TE = Y_{w,a_w} - Y_{w^*,a_w^*}, \tag{13}$$

where $a_w$ is the attention weight of $w$.

Since the word-caused bias is produced by word frequency, it is precisely represented by the natural direct effect (NDE) of $w$:

$$NDE = Y_{w,a_w^*} - Y_{w^*,a_w^*}, \tag{14}$$

which can be viewed as the effect of $w$ on $Y$ when the attention is fixed. With the help of backdoor adjustment, we get the attention weight of $w$ for each label. To obtain $Y_{w,a_w^*}$ we can take a meanpooling on $a(w)$, where $a \in \{a_{q_1}, a_{q_2}, ..., a_{q_m}\}$. The reduction of the word-caused bias is represented as:

$$TIE = TE - NDE = Y_{w,a_w} - Y_{w,a_w^*}. \tag{15}$$

**Algorithm 1:** The pseudocode of DAS.

**Input :** Training dataset $D$, Keyword vocabulary $V$, Hyperparameter $\lambda$.
**Output:** Model parameters $\theta$.

**for** *(I, y)* $\in D$ **do**
  Initialize attention signal $A$ from $V$;
  $H$ = Encode($I$) ;
  **for** $i = 1$ **to** $m$ **do** ▷ Equ. 12
    $e_{q_i} = \tanh(HW_{q_i} + b_{q_i})$
    $a_{q_i} = \text{softmax}(e_{q_i})$
  **end**
  $\bar{a} = \text{MeanPooling}(\{a_{q_1}, a_{q_2}, ..., a_{q_m}\})$;
  **for** $i = 1$ **to** $m$ **do** ▷ Equ. 15
    $\tilde{a}_{q_i} = a_{q_i} - \bar{a}$
    $\tilde{a}_{q_i}$ = softmax($\tilde{a}_{q_i}$);
    $\tilde{R}_i = \sum_{j=1}^{n} h_j * \tilde{a}_{q_i(j)}$;
  **end**
  $\tilde{R} = \{\tilde{R}_1, \tilde{R}_2, ..., \tilde{R}_m\}$ ;
  $\hat{y}$ = softmax(FC($R$)) ▷ Equ. 7;
  $\theta$ =argmin$_\theta \mathcal{L}(y, \hat{y}, A, a_{q_y}; \theta)$ ▷ Equ. 11;
**end**

| Type | Legal Verdict | Medical Triage |
|---|---|---|
| # of Samples | 164,194 | 1,773,497 |
| # of Labels | 138 | 185 |
| Gini coefficient of Labels | 0.47 | 0.77 |
| Avg. Tokens in Input | 239.18 | 41.41 |

Table 2: Statistics of the two datasets.

**Implementation** As Fig. 6 shows, we implement Equ. 15 by subtracting the average attention weight $\bar{a} \in \mathbb{R}^{n \times 1}$ from original attention weight. In other words, each value is subtracted from the average value of its column. After this subtraction, words that have high attention weights across all labels become less significant, while words with high attention weights for only a few labels retain their importance.

The pseudocode of DAS is shown in Alg. 1.

## Experiments

### Datasets

**Legal Verdict**[4]**.** This dataset is released by Chinese AI and Law Challenge (CAIL2018) (Zhong et al. 2018), and it has been widely used in LegalAI research. Each sample consists of a factual description and a corresponding charge. The fact description serves as the input, while the charge is output.

**Medical Triage**[5]**.** This dataset collects medical conversations. The input is patients' questions and the output is the corresponding department.

The statistics of the two datasets are presented in Tab. 2. The Gini coefficient shows the imbalance of label distributions within each dataset. To ensure fair evaluations, we partition each dataset randomly into training, validation, and test sets, maintaining an 80%:10%:10% ratio.

[4]https://github.com/thunlp/CAIL

[5]https://github.com/liangsbin/Chinese-medical-dialogue-data

| Methods | Legal Verdict | | | | | Medical Triage | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Ma-P | Ma-R | Ma-F | Ma-F low25% | Acc | Ma-P | Ma-R | Ma-F | Ma-F low25% |
| BOW | 80.27 | 73.96 | 69.59 | 70.57 | 50.99 | 74.83 | 67.24 | 58.78 | 61.62 | 56.04 |
| SVM | 80.10 | 81.87 | 76.16 | 76.47 | 66.94 | 74.17 | 71.73 | 57.36 | 61.07 | 54.53 |
| Transformer | 82.29 | 79.64 | 76.87 | 77.41 | 67.30 | 75.88 | 67.89 | 63.20 | 64.32 | 57.42 |
| Transformer+ATT | 84.66 | 79.87 | 78.87 | 78.73 | 68.69 | 75.99 | 68.15 | 61.76 | 63.79 | 57.29 |
| Transformer+AS | 84.50 | 81.77 | 77.40 | 78.08 | 65.12 | 76.43 | 67.41 | 61.73 | 63.14 | 57.48 |
| Transformer+DAS | **85.92** | **82.36** | **80.50** | **80.84** | **70.68** | **77.04** | **68.26** | **64.54** | **65.14** | **60.16** |
| BiLSTM | 83.89 | 76.85 | 77.20 | 76.41 | 61.76 | 76.55 | 66.38 | 61.58 | 62.87 | 53.65 |
| BiLSTM+ATT | 85.50 | 81.68 | 80.48 | 80.59 | 69.37 | 77.05 | **67.56** | 62.54 | 63.88 | 55.54 |
| BiLSTM+AS | 85.36 | 80.13 | 80.30 | 79.76 | 67.76 | 77.33 | 67.18 | 62.03 | 63.51 | 54.97 |
| BiLSTM+DAS | **86.33** | **82.97** | **81.83** | **82.22** | **73.26** | **77.34** | **67.56** | **64.40** | **65.17** | **58.84** |

Table 3: Results of classification on two datasets, and ATT denotes unsupervised attention mechanism.

## Baselines

We implement the following baselines:

**BoW** is a simplifying representation, the text is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. **SVM** is a robust prediction methods based on statistical learning frameworks. **Transformer** (Vaswani et al. 2017) is a deep learning model that adopts the mechanism of self-attention (distinct from the attention mechanism in this paper) to process sequential input data. **BiLSTM** (Hochreiter and Schmidhuber 1997) is an artificial neural network used in deep learning, featuring feedback connections.

To mitigate the encoder's impact and effectively evaluate the performance of our De-biased Attention Supervision (DAS), we have applied the attention mechanisms **(ATT)**, attention supervision **(AS)**, and de-biased attention supervision **(DAS)** on both the Transformer and BiLSTM models.

## Evaluation Metric

**Automatic Evaluation** To evaluate the performance of classification, we calculate the Accuracy (Acc), Macro-P(Ma-P), Macro-R(Ma-R), and Macro-F1(Ma-F).

**Human Evaluation** For evaluating the rationality of attention distributions, we carry out human evaluations. We randomly select 500 test cases from each dataset. For each case, we visualize the attention distributions, as shown in Figure 1, generated by different methods (e.g., ATT, AS, DAS). We present these visualizations to 5 human annotators[6], who are then asked to identify the distribution that appears most reasonable as the 'better one'. A case is considered valid only if at least 3 annotators select the same distribution. Then, we calculate the proportion of each method being selected as the 'better one' to make the comparison.

## Implementation Details

We conducted our experiments using two V100 GPUs. We use Gensim (Řehůřek and Sojka 2010) with a large generic corpus to initialize the word embeddings of deep-learning models, which is in the dimension of 300. The size of the keyword vocabulary is set to 1000. The setting for $\lambda$ is 0.15.

[6]To ensure fairness, we randomize the order of results.

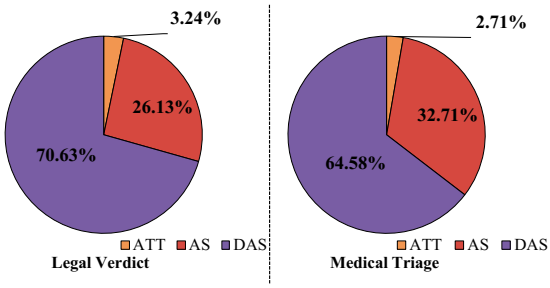


Figure 7: Results of human evaluation.

## Experimental Results

We analyze the experimental results in this section.

**Performance of classification:** Examining Tab. 3, we can derive the following conclusions: 1) In both datasets, the performance of AS drops compared to ATT (e.g., Ma-P for BiLSTM+AS declines from 81.68% to 80.13% in the legal verdict dataset). This decline underscores the impact of biases within AS on classification accuracy. 2) The application of DAS leads to enhanced performance by the models (e.g., Transformer and BiLSTM) compared to other attention mechanisms (e.g., ATT, AS) on both datasets. This observation suggests that the de-bias operations in DAS contribute to improved classification performance. 3) Notably, the improvement in performance is more pronounced on low-frequency labels (e.g., Ma-F for Transformer+DAS increases from 65.12% to 70.68% on legal dataset and from 57.48% to 60.16% on medical dataset), which proves that DAS has the ability to make the models more robust. 4) The medical triage is a more challenging dataset, possibly due to its larger label set. 5) The performance of deep-learning models surpasses that of classic models.

**Rationality of attention distribution:** After certification, there remain 463 samples for Legal Verdict and 480 samples for Medical Triage, and the results are depicted in Fig. 7. In the majority, the attention distributions of DAS are picked as the better ones (e.g., 70.63% for the legal verdict dataset and 64.58% for the medical triage dataset). This provides evidence that DAS generates more coherent attention distributions than ATT and AS. Compared to ATT, AS enhances the rationality of attention distribution, however, it brings no benefit to the classification performance as Tab. 3 shows.

| Dataset: *Legal Verdict* \| Label: *False Accusation* | | Predicted Label |
|---|---|---|
| ATT | On May 6, defendant A pretended to be injured in the waist by B because of a dispute over arable land with B. He also accused B of injuring himself to the public security organ with the intention of falsely accusing B. Appraised by the Physical Evidence Appraisal Center of the Public Security Department, the fracture of the left transverse process of A's 2nd to 4th lumbar vertebrae was old and it was not suitable to assess the degree of injury. | *False Accusation* ✅ |
| AS | On May 6, defendant A pretended to be injured in the waist by B because of a dispute over arable land with B. He also accused B of injuring himself to the public security organ with the intention of falsely accusing B. Appraised by the Physical Evidence Appraisal Center of the Public Security Department, the fracture of the left transverse process of A's 2nd to 4th lumbar vertebrae was old and it was not suitable to assess the degree of injury. | *Intentional Injury* ❌ |
| DAS | On May 6, defendant A pretended to be injured in the waist by B because of a dispute over arable land with B. He also accused B of injuring himself to the public security organ with the intention of falsely accusing B. Appraised by the Physical Evidence Appraisal Center of the Public Security Department, the fracture of the left transverse process of A's 2nd to 4th lumbar vertebrae was old and it was not suitable to assess the degree of injury. | *False Accusation* ✅ |

Figure 8: Case study. The attention weight is simplified into four levels.

| Methods | Legal Verdict | | Medical Triage | |
|---|---|---|---|---|
| | Acc | Ma-F | Acc | Ma-F |
| AS | 85.36 | 80.30 | 77.33 | 63.51 |
| DAS | **86.33** | **82.22** | **77.34** | **65.17** |
| w/o wb | 85.81 | 80.98 | 77.25 | 64.04 |

Table 4: Results of ablation experiment.

## Analysis Study

We conducted an ablation experiment as follows[7]:

- **w/o wb** removes the operation on the word-caused bias, and only employs the adjustment on the label-caused bias.

The results presents in Tab. 4 reveal the following insights: 1) The performance gap between DAS and w/o wb underscores that the eliminating word-caused bias in attention supervision can contribute positively to the classification. 2) In comparison to AS, w/o wb proves the effectiveness of backdoor adjustment in mitigating label-caused bias.

Fig. 8 shows an intuitive comparison among the methods. In this case, we can observe the following: 1) ATT correctly predicts the label but generates an inadequate attention distribution. For instance, it assigns high weights to the word "himself". 2) AS improves the attention distribution quality, but the predicted label is incorrect. In this dataset, the label *Intentional Injury* is high-frequency, and the word "injury" is distributed across multiple labels. This introduces biases into AS and hampers accurate prediction. 3) DAS, by mitigating biases (both label-caused and word-caused) achieves a weight distribution that prioritizes the word "falsely accusing", leading to the correct label prediction. This case study underscores the effectiveness of DAS in enhancing both attention distribution quality and classification accuracy.

Fig. 9 depicts the relationship between the classification performance (e.g., Ma-F1) and the size of the keyword vocabulary. From it, we have the following observations: 1) DAS consistently outperforms AS across varying sizes. 2) With the size grows, the performance tends to imporve. 3) Notably, there are dips in performance for both datasets. This phenomenon may be because the quality of keywords can vary with the size of the vocabulary.

[7]Unless specified otherwise, we utilize BiLSTM as the encoder for DAS in the subsequent sections.

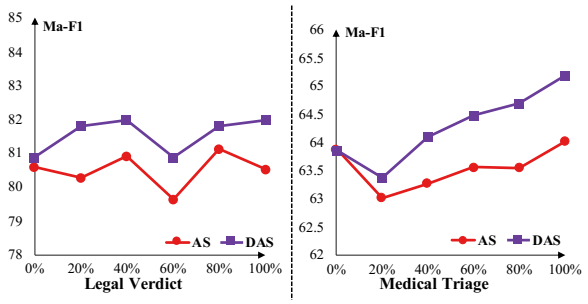


Figure 9: The relevance of the Ma-F1 and the size of keyword vocabulary in two datasets.

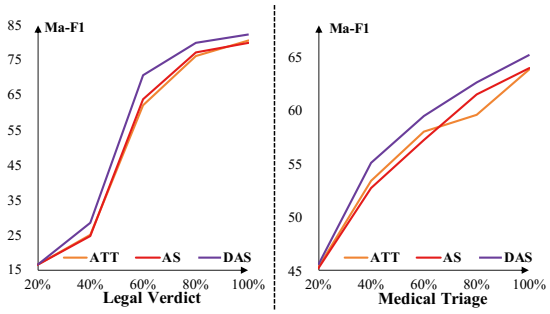


Figure 10: The relevance of Ma-F1 and training set size.

Turning to Fig. 10, we find that DAS gains a higher training efficiency for both datasets, which proves the effectiveness of DAS from an alternative perspective.

## Conclusion and Future Work

In this paper, our investigation centers on attention supervision (AS) for text classification. We conduct a series of experiments on two professional datasets (e.g., medicine and law). We identify two biases behind the AS: label-caused bias and word-caused bias. To eliminate these two biases, we propose a novel De-biased Attention Supervision (DAS) algorithm for text classification. Extensive experiments demonstrate that DAS achieves better classification accuracy alongside more reasonable attention distributions. In the future, besides attention supervision, we intend to address other types of bias, such as biases in the pretraining corpus, to further enhance the efficacy of text classification.

## Acknowledgments

This work was supported in part by National Key Research and Development Program of China (2021YFC3340300), National Natural Science Foundation of China (No. 62376243, 62037001, U20A20387), Academic Rising Star Program of Zhejiang University, the StarryNight Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010), Project by Shanghai AI Laboratory (P22KS00111), Program of Zhejiang Province Science and Technology (2022C01044), the Project of the Donghai Laboratory (Grant no. DH-2022ZY0013) and the Key Research and Development Program of Zhejiang Province, China (No. 2023C01152).

## References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bao, Y.; Chang, S.; Yu, M.; and Barzilay, R. 2018. Deriving Machine Attention from Human Rationales. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 1903–1913. Association for Computational Linguistics.

Barrett, M.; Bingel, J.; Hollenstein, N.; Rei, M.; and Søgaard, A. 2018. Sequence Classification with Human Attention. In Korhonen, A.; and Titov, I., eds., *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, 302–312. Association for Computational Linguistics.

Campos, R.; Mangaravite, V.; Pasquali, A.; Jorge, A.; Nunes, C.; and Jatowt, A. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509: 257–289.

Chen, H.; and Ji, Y. 2020. Learning variational word masks to improve the interpretability of neural text classifiers. *arXiv preprint arXiv:2010.00667*.

Choi, S.; Park, H.; Yeo, J.; and Hwang, S. 2020. Less is More: Attention Supervision with Counterfactuals for Text Classification. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 6695–6704. Association for Computational Linguistics.

Du, C.; and Huang, L. 2018. Text Classification Research with Attention-based Recurrent Neural Networks. *Int. J. Comput. Commun. Control*, 13(1): 50–61.

Gasparetto, A.; Marcuzzo, M.; Zangari, A.; and Albarelli, A. 2022. A Survey on Text Classification Algorithms: From Text to Predictions. *Inf.*, 13(2): 83.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Hu, D. 2019. An Introductory Survey on Attention Mechanisms in NLP Problems. In Bi, Y.; Bhatia, R.; and Kapoor, S., eds., *Intelligent Systems and Applications - Proceedings of the 2019 Intelligent Systems Conference, IntelliSys 2019, London, UK, September 5-6, 2019, Volume 2*, volume 1038 of *Advances in Intelligent Systems and Computing*, 432–448. Springer.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

J, A. K.; Abiramy, S.; Trueman, T. E.; and Cambria, E. 2021. Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit. *Neurocomputing*, 441: 272–278.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 3543–3556. Association for Computational Linguistics.

Jiang, M.; Liang, Y.; Feng, X.; Fan, X.; Pei, Z.; Xue, Y.; and Guan, R. 2018. Text classification based on deep belief network and softmax regression. *Neural Comput. Appl.*, 29(1): 61–70.

Kowsari, K.; Meimandi, K. J.; Heidarysafa, M.; Mendu, S.; Barnes, L. E.; and Brown, D. E. 2019. Text Classification Algorithms: A Survey. *Inf.*, 10(4): 150.

Li, M.; Wang, T.; Zhang, H.; Zhang, S.; Zhao, Z.; Miao, J.; Zhang, W.; Tan, W.; Wang, J.; Wang, P.; et al. 2022. End-to-End Modeling via Information Tree for One-Shot Natural Language Spatial Video Grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8707–8717.

Liu, Y.; Wu, Y.; Zhang, Y.; Sun, C.; Lu, W.; Wu, F.; and Kuang, K. 2023. ML-LJP: Multi-Law Aware Legal Judgment Prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1023–1034.

Lv, Z.; Chen, Z.; Zhang, S.; Kuang, K.; Zhang, W.; Li, M.; Ooi, B. C.; and Wu, F. 2023a. Ideal: Toward high-efficiency device-cloud collaborative and dynamic recommendation system. *arXiv preprint arXiv:2302.07335*.

Lv, Z.; Wang, F.; Zhang, S.; Kuang, K.; Yang, H.; and Wu, F. 2022. Personalizing Intervened Network for Long-tailed Sequential User Behavior Modeling. *arXiv preprint arXiv:2208.09130*.

Lv, Z.; Zhang, W.; Zhang, S.; Kuang, K.; Wang, F.; Wang, Y.; Chen, Z.; Shen, T.; Yang, H.; Ooi, B. C.; et al. 2023b. DUET: A Tuning-Free Device-Cloud Collaborative Parameters Generation Framework for Efficient Device Model Generalization. In *Proceedings of the ACM Web Conference 2023*, 3077–3085.

Mi, H.; Wang, Z.; and Ittycheriah, A. 2016. Supervised Attentions for Neural Machine Translation. In Su, J.; Carreras, X.; and Duh, K., eds., *Proceedings of the 2016 Confer-*

*ence on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2283–2288. The Association for Computational Linguistics.

Nguyen, M.; and Nguyen, T. H. 2018. Who is Killed by Police: Introducing Supervised Attention for Hierarchical LSTMs. In Bender, E. M.; Derczynski, L.; and Isabelle, P., eds., *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, 2277–2287. Association for Computational Linguistics.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Priyanshu, A.; and Vijay, S. 2022. AdaptKeyBERT: An Attention-Based approach towards Few-Shot\& Zero-Shot Domain Adaptation of KeyBERT. *arXiv preprint arXiv:2211.07499*.

Qiao, T.; Dong, J.; and Xu, D. 2018. Exploring Human-Like Attention Supervision in Visual Question Answering. In McIlraith, S. A.; and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 7300–7307. AAAI Press.

Řehůřek, R.; and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA. http://is.muni.cz/publication/884893/en.

Rubin, D. B. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469): 322–331.

Shen, K.; Ju, Z.; Tan, X.; Liu, Y.; Leng, Y.; He, L.; Qin, T.; Zhao, S.; and Bian, J. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Shen, K.; Leng, Y.; Tan, X.; Tang, S.; Zhang, Y.; Liu, W.; and Lin, E. 2022. Mask the correct tokens: An embarrassingly simple approach for error correction. *arXiv preprint arXiv:2211.13252*.

Shen, K.; Wu, L.; Tang, S.; Zhuang, Y.; Ding, Z.; Xiao, Y.; Long, B.; et al. 2021. Learning to generate visual questions with noisy supervision. *Advances in Neural Information Processing Systems*, 34: 11604–11617.

Sun, X.; and Lu, W. 2020. Understanding Attention for Text Classification. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 3418–3428. Association for Computational Linguistics.

Tida, V. S.; and Hsu, S. H. Y. 2022. Universal Spam Detection using Transfer Learning of BERT Model. In *55th Hawaii International Conference on System Sciences, HICSS 2022, Virtual Event / Maui, Hawaii, USA, January 4-7, 2022*, 1–9. ScholarSpace.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wu, A.; Kuang, K.; Xiong, R.; and Wu, F. 2022a. Instrumental Variables in Causal Inference and Machine Learning: A Survey. *arXiv preprint arXiv:2212.05778*.

Wu, Y.; Kuang, K.; Zhang, Y.; Liu, X.; Sun, C.; Xiao, J.; Zhuang, Y.; Si, L.; and Wu, F. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 763–780.

Wu, Y.; Liu, Y.; Lu, W.; Zhang, Y.; Feng, J.; Sun, C.; Wu, F.; and Kuang, K. 2022b. Towards Interactivity and Interpretability: A Rationale-based Legal Judgment Prediction Framework. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 4787–4799.

Wu, Y.; Lu, W.; Zhang, Y.; Jatowt, A.; Feng, J.; Sun, C.; Wu, F.; and Kuang, K. 2023. Focus-aware response generation in inquiry conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 12585–12599.

Zhang, R.; Shen, J.; Liu, T.; Liu, J.; Bendersky, M.; Najork, M.; and Zhang, C. 2023a. Do not blindly imitate the teacher: Using perturbed loss for knowledge distillation. *arXiv preprint arXiv:2305.05010*.

Zhang, R.; Yu, Y.; Shen, J.; Cui, X.; and Zhang, C. 2023b. Local Boosting for Weakly-Supervised Learning. *arXiv preprint arXiv:2306.02859*.

Zhang, R.; Yu, Y.; Shetty, P.; Song, L.; and Zhang, C. 2022. Prboost: Prompt-based rule discovery and boosting for interactive weakly-supervised learning. *arXiv preprint arXiv:2203.09735*.

Zhao, Y.; Jin, X.; Wang, Y.; and Cheng, X. 2018. Document Embedding Enhanced Event Detection with Hierarchical and Supervised Attention. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, 414–419. Association for Computational Linguistics.

Zhong, H.; Xiao, C.; Guo, Z.; Tu, C.; Liu, Z.; Sun, M.; Feng, Y.; Han, X.; Hu, Z.; Wang, H.; and Xu, J. 2018. Overview of CAIL2018: Legal Judgment Prediction Competition. *CoRR*, abs/1810.05851.

Zhou, S.; Liu, Y.; Wu, Y.; Kuang, K.; Zheng, C.; and Wu, F. 2022. Similar case based prison term prediction. In *CAAI International Conference on Artificial Intelligence*, 284–297. Springer.