International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-020-00231-3

REGULAR PAPER


# Mitigating sentimental bias via a polar attention mechanism

Tao Yang[1] · Rujing Yao[1] · Qing Yin[1] · Qiang Tian[2] · Ou Wu[1]

Received: 14 October 2019 / Accepted: 25 July 2020




**Abstract**

Fairness in machine learning has received increasing attention in recent years. This study focuses on a particular type of machine learning fairness, namely sentimental bias, in text sentiment analysis. Sentimental bias occurs on words (or phrases) when they are distributed distinctly in positive and negative corpora. It results in that an excessively proportion of words carry negative/positive sentiment in learned models. This study proposed a new attention mechanism, called polar attention, to mitigate sentimental biases. It consists of two modules, namely polar flipping and distance measurement. The first module explicitly models word sentimental polarity and can prevent that neutral words flip positively or negatively. The second module is used to attend negative/positive words. In the experiments, three benchmark data sets are used, and supplementary testing sets are compiled. Experimental results verify the effectiveness of the proposed method.

**Keywords** Sentiment analysis · Sentimental bias · Attention · Polar flipping

## 1 Introduction

The fairness issue in machine learning refers to that artificial intelligence applications (exactly learned models) that systematically discriminate against specific populations [27]. For example, some facial recognition systems misclassify gender more frequently when presented with dark-skinned women than with light-skinned men [17]. Model biases result in certain population groups being unfairly denied loans, insurance, and employment opportunities [16].

Biases in population models occur when training data are skewed on certain (group) variables related to gender, race, or culture. Recent literature has presented proposals to mitigate model biases. Liu et al. [8] conducted a theoretical analysis of fairness in machine learning and emphasized that a fairness criterion is crucial for a debiasing algorithm design. Edizel et al. [19] addressed the problem of algorithmic bias in recommender systems. Some studies have improved model fairness by imposing additional constraints or conditions [3,9,20,22,26]. As previously stated, although numerous achievements have been made, almost all the existing studies aim to mitigate biases caused by specific population variables.

In contrast with existing studies, the current study focuses on a particular type of model bias in text sentiment analysis, namely sentimental bias, in which neutral words or phrases flipped to a truly polar (negative/positive) sentiment in the learned model. Figure 1 shows the generation of sentimental bias on the word "***express***." The sentimental labels of the first three sentences are "negative" and the fourth is "positive." If such skew on negative/positive labels for sentences containing the word "***express***" exists on the whole training data, then the model learned using conventional techniques is highly likely to consider "***express***" a strongly negative word. Consequently, the learned model is likely to label the test sample (its true label is "neutral") in Fig. 1 as "negative." The word "***express***" itself definitely has no sentimental polarity. In contrast with existing model biases, no certain variable is directly related to sentimental bias, and thus, existing fairness criteria and debiasing methods are not directly applicable. Hence, new methods should be investigated.

✉ Ou Wu
wuou@tju.edu.cn

Tao Yang
yangtao087@gmail.com

Rujing Yao
rjyao@tju.edu.cn

Qing Yin
qingyin@tju.edu.cn

Qiang Tian
tianqiang@tjnu.edu.cn

[1] The Center for Applied Mathematics, Tianjin University, Tianjin, China

[2] Tianjin Normal University , Tianjin, China





Journal: 41060 MS: 0231 TYPESET DISK LE CP Disp.:2020/8/4 Pages: 10 Layout: Large

| | |
|---|---|
| 1. The *express* delivery is too slow. | --- Negative |
| 2. The *express* service is poor. | --- Negative |
| 3. The *express* goods were all broken. | --- Negative |
| 4. The *express* delivery is timely. | --- Positive |
| Test: This *express* company is called FedEx. --- | ? |

**Fig. 1** The first three training sentences are labeled "negative." The distribution between negative and positive sentences is skewed (3:1)

Indeed, sentiment analysis mainly rely on truly polar (negative/positive) words. Sentimental bias leads to that many neutral words are mistakenly flipped into negative or positive state in learned models. As a result, classification errors are prone to occur when sentences contain these neutral words. Ideally, the polarities of neutral words (e.g., "express") should not be flipped while those of truly polar words should be.

This paper firstly proposed a polar flipping module. This module assumes that each word initially has a neutral polarity and this polarity can finally flip either positively or negatively. Considering that neutral words usually have less importance for classification, the distance between the initial and finally flipped polarities of each word is then regarded as an attention score. These two modules consist of a new attention mechanism, called polar attention. A small flipping rate can avoid unnecessary flips for neutral words and does not restrain necessary flips for truly polar words. The flipping proportion for neutral words is then reduced while the truly polar words still receive higher attention. Consequently, sentimental bias can be alleviated.

To our knowledge, this work is the first to consider sentimental bias, which differs from conventional model biases in terms of specific population variables. The experimental results show the initial success of our new attention mechanism. All resources are available at https://github.com/absa-nlp/PA-Net.

## 2 Related work

Existing studies on learning fairness either proposed new fairness metrics for demographic groups or presented mitigation strategies to improve fairness. A number of metrics (e.g., statistical parity [3], equalized odds [9], and predictive parity [15]) have been used to determine the fairness of classifiers to specific sensitive attributes, such as race and gender. These metrics can be imposed as constraints or incorporated into a loss function.

Two types of mitigation methods are typically investigated. The first type reorganizes training data using conventional methods, such as removing population variables[9]. The second type protects sensitive attributes via adversarial learning [4]. This type of methods tries to learn a predictor that can classify correctly and an adversary that fails to predict sensitive attributes. Previous attention so far for learning fairness in NLP has been primarily on word embedding. Bolukbasi et al. [14] observed that word embeddings trained on Google News articles contains gender bias and some embeddings pinpoint sexism implicit in training texts. To mitigate gender-bias, Zhao et al. [5] proposed to preserve gender information in certain dimensions of word vectors while compelling other dimensions to be free of gender influence. Dixon et al. [7] firstly investigated untended bias in text classification. Nevertheless, their work still focuses on the bias related to demographic groups.

Unlike existing studies, our work focuses on another simple yet foundational bias in sentiment analysis which is called sentimental bias. In contrast with existing model biases, words or phrases rather than population variables are to be considered in sentimental bias. New debiasing algorithms should be investigated.

## 3 Methodology

This section describes our polar attention mechanism and the entire network called PA-Net shown in Fig. 2.

The input is a sentence with $n$ words, and each word is associated with a word embedding $w_i \in R^d$, where $i$ is the word index in the context and $d$ is the embedding dimension. We map the input words into their vector representations using a pretrained embedding table. Then, assuming that the polar labels of each word are divided into three categories, namely positive, neutral, and negative, which are represented by $P_1$, $P_2$, and $P_3$, respectively. We initially assign a neutral label for each word as the input of the polar flipping module. This module outputs the final polar labels for each word. $P_1$, $P_2$, and $P_3$ can be characterized by one-hot vectors:

$$\begin{aligned} P_1 &= [1, 0, 0] \\ P_2 &= P_{initial} = [0, 1, 0] \\ P_3 &= [0, 0, 1] \end{aligned} \tag{1}$$

Subsequently, a bidirectional long short-term memory (Bi-LSTM) is utilized to capture the hidden representation of each word in a sentence. The output of Bi-LSTM at time $t$ is calculated as follows:

$$H_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{2}$$

where $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are the corresponding hidden vectors from the forward and backward LSTMs.

Note that all the initial polarities are set as neutral. The final polarities of neutral words (e.g., "express") should







Journal: 41060 MS: 0231 TYPESET DISK LE CP Disp.:2020/8/4 Pages: 10 Layout: Large

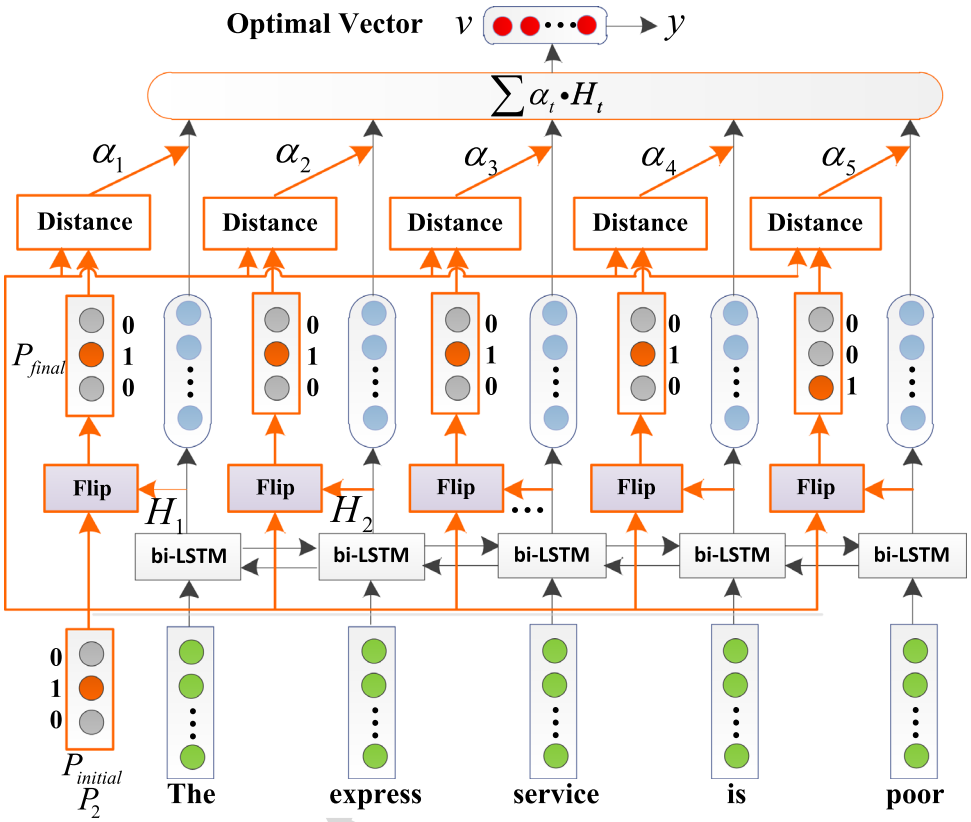



**Fig. 2** Overall of PA-Net. The orange lines represent the polar attention, which contains the polar flipping and the distance measurement modules. In the given example sentence, only the true polar word "poor" encourages flipping to the negative

remain neutral, whereas those of truly polar words should be flipped. We utilize the semantic information of context to determine whether a word is flipped positively or negatively to achieve a reasonable flip. The polar flipping module is described as follows:

$$P_{final} = \sigma_1(H_t)P_2 + [1-\sigma_1(H_t)] \{\sigma_2(H_t)P_1 + [1-\sigma_2(H_t)]P_3\} \quad (3)$$

where $\sigma_1$ and $\sigma_2$ are sigmoid functions and used to calculate the probabilities that the final polarity will remain $P_2$, or will flip to $P_1$ or $P_3$ based on the context information $H_t$. Therefore, the final output of Eq. (3) is the weighted combination of $P_1$, $P_2$, and $P_3$. In particular, if $\sigma_1(H_t) == 1$, then $P_{final} = P_2$. If $\sigma_1(H_t) == 0$ and $\sigma_2(H_t) == 1$, $P_{final} = P_1$. If $\sigma_1(H_t) == 0$ and $\sigma_2(H_t) == 0$, $P_{final} = P_3$.

Intuitively, if the polarity of a word is flipped to positive or negative, then this word is usually quite useful for sentiment classification (e.g., "*poor*" shown in Fig. 2). Alternatively, flipped words should receive more attention. Accordingly, the Euclidean distance between the initial and final polarities is used as the attention score of a word. The measurement is:

$$\alpha_t = \left\| P_{final} - P_{initial} \right\|_2 \quad (4)$$

where $\|\cdot\|_2$ refers to the Euclidean norm. Eq. (4) indicates that words that remain neutral will receive less attention, while truly polar words are encouraged to flip. Particularly, if the final polarity of a word is positive ([1, 0, 0]) or negative ([0, 0, 1]), the attention weight calculated by Eq. (4) is "1;" on the contrary, if the final polarity remains neutral ([0, 1, 0]), the attention weight calculated by Eq. (4) is "0."

In the output layer, the outputs of the Bi-LSTM are summed with the associated attention scores to produce the final dense feature vector as follows:

$$v = \sum_{t=1}^{n} \alpha_t \cdot H_t \quad (5)$$

Then, the following softmax function is used to predict the final category:

$$y = softmax(W^T v + b) \quad (6)$$

Considering that the number of truly polar words is usually limited and to prevent the flipping of neutral words, a flipping regularizer is added and defined as follows:

$$\Omega = \sum_{t=1}^{n} R_t \quad (7)$$

$$R_t = \begin{cases} \sigma_1(H_t) + [1-\sigma_2(H_t)], & if\ L_t = Positive \\ \sigma_1(H_t) + \sigma_2(H_t), & if\ L_t = Negative \\ 1-\sigma_1(H_t), & if\ L_t = Neutral \end{cases} \quad (8)$$







Journal: 41060 MS: 0231 TYPESET DISK LE CP Disp.:2020/8/4 Pages: 10 Layout: Large

**Table 1** Details of the benchmark data sets

| Data sets | Train | | Dev | | Test | |
|---|---|---|---|---|---|---|
| | Pos. | Neg. | Pos. | Neg. | Pos. | Neg. |
| MR | 4318 | 4318 | 480 | 480 | 533 | 533 |
| SST-2 | 3610 | 3310 | 444 | 428 | 909 | 912 |
| IMDB | 9992 | 10008 | 1265 | 1235 | 1243 | 1257 |

where $L_t$ is the polarity label of the $t$-th word, which obtained from Wu et al. [1] [24]. This regularizer penalizes the polarity of words appearing in the sentiment lexicon to be incorrectly flipped. Besides, if a word does not appear in the sentiment lexicon, we also regard the polarity of this word as neutral in regularizer since only a few words can express sentimental tendencies.

PA-Net is trained with the following objective function:

$$J = \frac{1}{N}\sum_{i=1}^{N}[-y_i \log p(y_i) + \lambda\Omega] \tag{9}$$

where $N$ is the number of training samples; $y_i$ indicates the true label; $p(y_i)$ denotes the prediction probability for the true label; $\Omega$ signifies the regularizer of Eq.( 7); and $\lambda$ is the regularizer weight and searched via cross validation.

# 4 Experiments

## 4.1 Data sets and evaluation

We perform experiments on three benchmark English sentiment analysis data sets: MR [23], SST2 [12], and IMDB [2]. All data sets consist of reviews with positive and negative classes. To facilitate the repeatability of experiment, we adopt the same data splitting process used in previous work [11,12]. Table 1 presents the details of the three data sets. The samples in all partitions are basically balanced.

In this work, we evaluate our proposed polar attention mechanism in terms of classification performance and effectiveness in mitigating sentimental bias. To verify the debiasing process, three supplementary testing sets, denoted as MR-S, SST2-S, and IMDB-S, are compiled for the three original benchmark sets. Each supplementary testing sample is neutral, because sentimental bias mostly affects neutral samples. The compilation strategy for each supplementary testing set is as follows. We select biased words from all neutral words by calculating the distribution of a word in the positive and negative categories. Then, each bias word is utilized to construct a neutral sentence. Additional details are provided in Sect. 4.2.

The classification accuracy on the original benchmark set reflects the overall performance, whereas the accuracy on the supplementary testing set reflects the debiasing capability.

## 4.2 Details of supplementary testing sets

In this subsection, we will describe the details of compiling the supplementary testing sets. The construction process is as follows.

- First, we counted all the words, with word vectors that can be found in the pre-trained embedding table in the three benchmark data sets. A total of 18616 words were obtained.
- Second, five annotators manually picked the clearly neutral words among the 18616 words (e.g., names of people, counties, and cities). We did not pick vague neutral words because their polarity is context sensitive, such as "rainbow" and "flower." A corpus of candidate biased words containing 7796 words was collected from the annotation results of the five annotators by voting.
- Third, we counted the times each candidate biased word appeared in the positive and negative samples in each training data set. Two evaluation metrics, namely bias rate and sum of numbers, were utilized to determine whether a candidate biased word is more likely to be a true biased word. This selection strategy is as follows:

$$bias = \begin{cases} 1 & r > 2 \cup s > 5 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $r$ is bias rate and $s$ is sum of numbers, which are defined as follows:

$$r = \frac{\max\left(n_{pos}, n_{neg}\right)}{\max\left\{\min\left(n_{pos}, n_{neg}\right), 1\right\}} \tag{11}$$

$$s = n_{pos} + n_{neg} \tag{12}$$

where $n_{pos}$ and $n_{neg}$ denote the number of times that a word appears in the positive and negative samples,

**Table 2** Examples of supplementary testing sets (biased words are in bold type)

| Samples of Supplementary Sets | Label |
|---|---|
| serving **sara** is a movie. | Neutral |
| the boy is named for his **grandfather**. | Neutral |
| he was born at **indian**. | Neutral |
| **schaeffer** is a director. | Neutral |
| **clooney** is a hollywood actor. | Neutral |

[1] https://github.com/Tju-AI/two-stage-labeling-for-the-sentiment-orientations







Journal: 41060 MS: 0231 TYPESET DISK LE CP Disp.:2020/8/4 Pages: 10 Layout: Large

**Table 3** Details of hyperparameter Settings

| Models | Hyperparameters | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **LSTM Hidden units** | | **Dropout** | | **Batch** | | λ | |
| Bi-LSTM | 50/50/100 | | 0.5 | | 32/16/64 | | - | |
| ATT-Bi-LSTM | 50/50/100 | | 0.5 | | 32/16/64 | | - | |
| PA-Bi-LSTM | 50/50/100 | | 0.5 | | 32/16/64 | | 2/1.5/3 | |
| | **Regions** | **Conv Size** | **Filters** | **Pooling** | **Repeats** | **Dropout** | **Batch** | λ |
| DPCNN | [3,4,5] | 3 | 64 | 3 | 2/2/4 | 0.3 | 32/32/64 | - |
| PA-DPCNN | [3,4,5] | 3 | 64 | 3 | 2/2/4 | 0.3 | 32/32/64 | 1.5/2/1.5 |
| | **Filters** | **Kernel Size** | **Stacks** | | **Dilations** | **Dropout** | **Batch** | λ |
| TCN | 128/128/256 | 3 | 2/2/4 | | [1,2,4] | 0.4 | 32/16/64 | - |
| PA-TCN | 128/128/256 | 3 | 2/2/4 | | [1,2,4] | 0.4 | 32/16/64 | 0.2 |
| | **Blocks** | **Heads** | **Hidden Units** | | **Feed Size** | **Dropout** | **Batch** | λ |
| Transformer | 1/1/2 | 4 | 32/32/64 | | 256/256/512 | 0.5 | 32/32/64 | - |
| PA-Trans | 1/1/2 | 4 | 32/32/64 | | 256/256/512 | 0.5 | 32/32/64 | 2.5/0.5/2.5 |

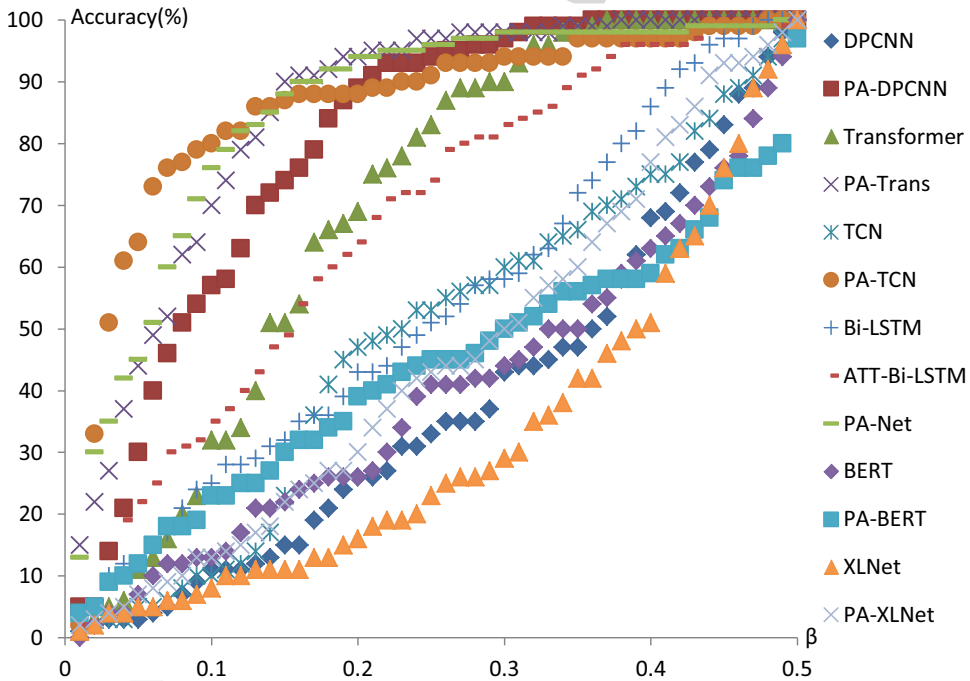



**Fig. 3** A qualitative comparison that we set the step size of $\beta$ to 0.01 and test all models on MR-S

respectively. Based on our initial analysis of candidate biased words, the higher the bias rate, the more likely sentimental bias will occur. To this consideration, we set the threshold of bias rate is 2 to keep an unbalanced ratio more than twice. Besides, we found that words with a small total number are not easy to cause stable bias, even if the bias rate is large. Therefore, we set the threshold of sum of words is 5 to obtain a sufficient amount of stable biased words. Hence, words with bias rates more than two and total numbers more than five are included in the high-probability biased words corpus. Consequently, we obtained three high-probability biased words corpus from MR, SST2, and IMDB data sets containing 403, 365, and 1276 words, respectively.

– Fourth, we randomly sampled three times from each high-probability biased word corpus. For each random sample, 100 bias words were selected, and each word was utilized to generate a neutral sentence manually. Then, we obtained three supplementary testing sets from the three benchmark data sets. For the supplementary testing set of each benchmark data sets, it contains three groups of neutral sentences generated by three randomly selected bias words. The average accuracy on these three groups was used as the final accuracy on the supplementary testing



Journal: 41060 MS: 0231 TYPESET DISK LE CP Disp.:2020/8/4 Pages: 10 Layout: Large

set. Some examples of the supplementary testing samples are listed in Table 2.

### 4.3 Competing models

Theoretically, the proposed polar attention mechanism can be integrated into most existing deep models. The following models are selected.

- *Bi-LSTM* [6]: This model uses bidirectional long short-term memory to capture sequential information.
- *Attention-based Bi-LSTM (ATT-Bi-LSTM)* [10]: This model introduces Bi-LSTM with an attention mechanism to automatically select features that have a decisive effect on classification.
- *Deep pyramid convolutional neural network (DPCNN)* [21]: In this model, a low-complexity word-level deep convolutional neural network architecture is adopted to represent long-range associations in sentences.
- *Temporal convolutional network (TCN)* [13]: This model combines dilations and residual connections with causal convolutions to model sequence information.
- *Transformer* [1]: This model proposed by Google is based solely on self-attention mechanisms for machine translation. We only use the encoder part in this work.
- *BERT* [18]: This model leverages the vanilla BERT pre-trained weights and fine-tunes on different data sets.
- *XLNet* [25]: This model integrates the idea of autoregressive models and bi-directional context modeling of BERT. XLNet make use of a permutation operation which achieved by using a special attention mask in Transformers during pre-training. We utilize the pre-trained weights of XLNet and fine-tune on different data sets in this work.
- *PA-Net, PA-DPCNN, PA-TCN, and PA-Trans*: These models are obtained by integrating our polar attention mechanism into Bi-LSTM, DPCNN, TCN, and Transformer, respectively. The polar attention mechanism is only used in the last hidden layer of the corresponding models.
- *PA-BERT*: BERT pre-trained model is used. The output of the last layer of BERT is used to replace the word embeddings of the PA-Net model.
- *PA-XLNet*: This model integrates the polar attention into the last hidden layer of pre-trained XLNet and fine-tunes on different data sets.

### 4.4 Hyperparameter settings

To conduct a fair comparison, the model with polar attention adopts the same parameter settings as those of the original model.

We use the 300-dimensional word embeddings from GloVe in the experiments. All models are trained using Adam. Additional hyperparameter $\lambda$ is searched from [0, 3] with a step size of 0.5. We list the hyperparameters of different models for **MR/SST2/IMDB** data set in Table 3.

### 4.5 Overall competing results

In order to evaluate the ability of the model to mitigate bias, we map the prediction probability in supplementary testing sets in the range of $[0.5 \pm \beta]$ into the neutral category. A qualitative comparison that we set the step size of $\beta$ to 0.01 and test all models on MR-S is shown in Fig. 3. We observe that all models with polar attention are superior to the original models in mitigating sentimental bias. In the following

**Table 4** Results on three benchmark data sets and supplementary testing sets regarding average accuracy (%). We map the prediction probability in the range of $[0.5 \pm \beta]$. In **B1**, **B2** and **B3**, the mapping ranges are $[0.5 \pm 0.05]$, $[0.5 \pm 0.10]$, and $[0.5 \pm 0.15]$, respectively

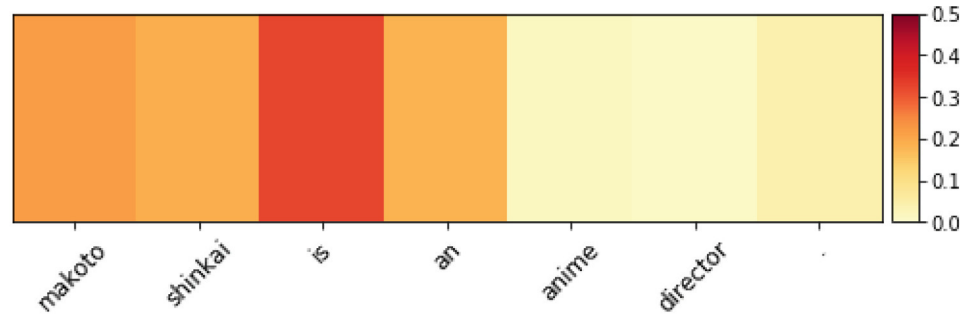
| Methods | MR | MR-S | | | SST2 | SST2-S | | | IMDB | IMDB-S | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Test | B1 | B2 | B3 | Test | B1 | B2 | B3 | Test | B1 | B2 | B3 |
| DPCNN | 77.8 | 13.7 | 25.4 | 40.3 | 83.0 | 5.2 | 12.0 | 21.7 | 86.0 | 8.3 | 14.4 | 21.7 |
| PA-DPCNN | 80.0 | 23.7 | 46.6 | 67.1 | 83.6 | 15.5 | 32.7 | 46.3 | 87.4 | 33.5 | 68.2 | 83.0 |
| Transformer | 77.2 | 15.3 | 23.7 | 39.3 | 82.1 | 13.9 | 26.7 | 45.3 | 85.9 | 8.4 | 17.7 | 31.0 |
| PA-Trans | 79.0 | 24.5 | 42.3 | 69.4 | 82.4 | 27.3 | 40.3 | 51.7 | 86.3 | 17.1 | 43.4 | 57.8 |
| TCN | 80.1 | 33.3 | 48.9 | 64.7 | 82.8 | 13.7 | 32.3 | 46.6 | 85.4 | 46.7 | 76.2 | 84.9 |
| PA-TCN | 82.1 | **45.8** | **65.5** | 73.2 | 84.4 | 27.5 | 45.2 | 60.3 | 87.3 | **58.8** | **82.3** | **89.7** |
| Bi-LSTM | 80.7 | 20.4 | 38.5 | 56.3 | 83.1 | 15.2 | 30.5 | 45.3 | 87.2 | 16.2 | 35.3 | 47.3 |
| ATT-Bi-LSTM | 81.5 | 24.0 | 44.9 | 61.2 | 84.1 | 15.1 | 35.7 | 49.4 | 86.6 | 10.7 | 17.1 | 28.9 |
| PA-Net | 82.8 | 36.9 | 60.4 | **78.3** | 85.7 | **33.2** | **57.7** | **77.1** | 87.8 | 34.3 | 60.6 | 74.5 |
| BERT | 88.2 | 8.7 | 15.0 | 22.3 | 91.1 | 5.3 | 7.2 | 10.7 | 93.9 | 10.5 | 22.3 | 33.6 |
| PA-BERT | 88.7 | 16.0 | 23.7 | 32.6 | 91.3 | 11.7 | 20.7 | 31.0 | 94.5 | 25.0 | 47.7 | 68.7 |
| XLNet | 89.3 | 4.1 | 6.5 | 9.6 | 91.8 | 4.4 | 8.3 | 12.0 | 95.1 | 3.7 | 7.7 | 15.6 |
| PA-XLNet | **89.5** | 7.5 | 15.4 | 27.8 | **92.0** | 6.8 | 12.7 | 18.2 | **95.5** | 10.1 | 23.0 | 42.8 |



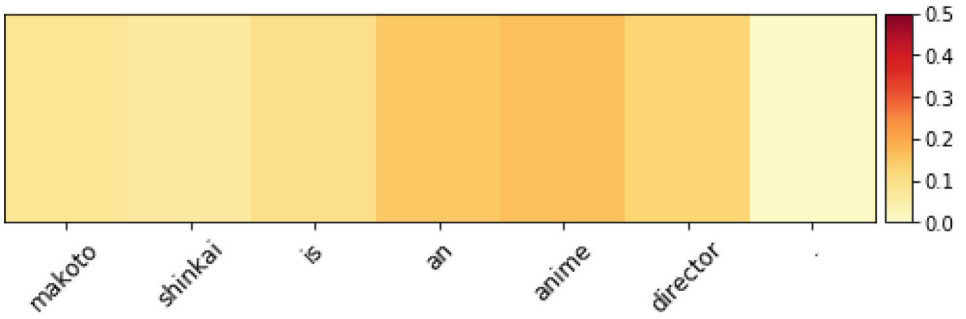




Journal: 41060 MS: 0231 TYPESET DISK LE CP Disp.:2020/8/4 Pages: 10 Layout: Large




**(a)** Attention heat map by ATT-Bi-LSTM




**(b)** Attention heat map by PA-Net

**Fig. 4** Visualizations of two different attention mechanisms in the same sentence

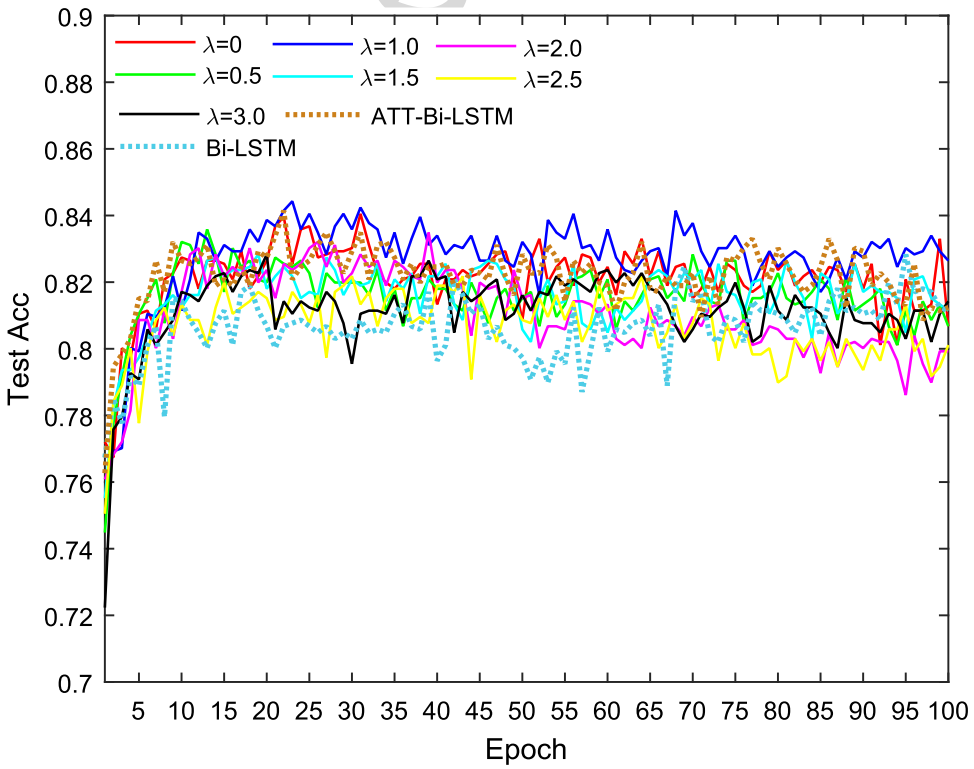



**Fig. 5** Test accuracy of different models during training on the MR data set



Journal: 41060 MS: 0231 TYPESET DISK LE CP Disp.:2020/8/4 Pages: 10 Layout: Large

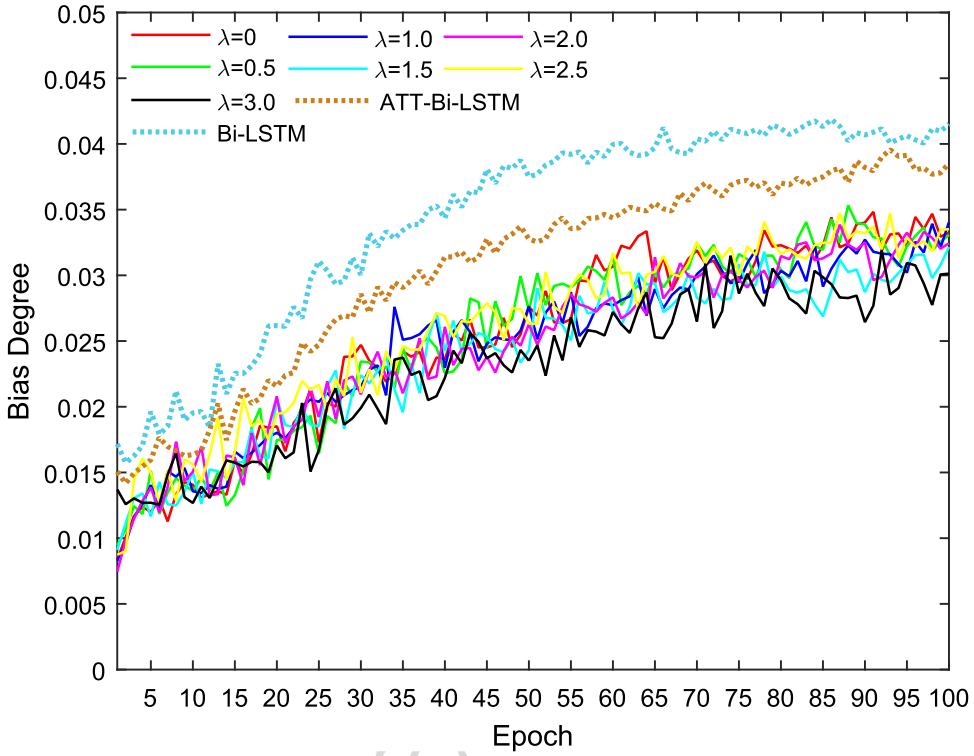



**Fig. 6** Bias degrees of different models during training on MR-S



experiments, we set the $\beta$ to 0.05, 0.10, and 0.15 for quantitative evaluation.

To facilitate comparison, all models are divided into six groups. The experimental results are presented in Table 4, and each result is the average value of five runs with random initialization. The best results are in bold type. We observe that all models with the polar attention achieved the best results on all benchmark testing sets and supplementary testing sets compared to original models. In particular, the models with polar attention are slightly better than the original models on the benchmark data sets. However, significant improvements are obtained by the polar attention-based models on the supplementary testing sets. The average increments are 15.2%, 23.3%, and 27.0% under three mapping methods (i.e., **B1**, **B2**, and **B3**) from the prediction probability to the neutral label. In addition, BERT and XLNet can effectively enhance the performance of classification. PA-XLNet which integrates the polar attention into pre-trained XLNet further improved the performance of XLNet and achieved the best results on all benchmark testing sets. The proposed polar attention mechanism cannot only improve classification performance, but also significantly mitigate sentimental bias.

## 4.6 Analysis

To verify whether our proposed method can effectively focus on polar words and limit sentimental bias, we visualize the attention weights of PA-Net and compare them with those of ATT-Bi-LSTM. The visualization results are presented in Fig. 4.

The test sentence in Fig. 4 is "***makoto shinkai is an anime director .***." We observe that the neutral word "***is***" has high weight in Fig. 4a, but the weight of "***is***" can be effectively reduced via our polar attention mechanism as shown in Fig. 4b. Figure 4b also illustrates that other neutral words such as "***makoto***" and "***shinkai***," have relatively lower weights than those in Fig. 4a. The results verify that our polar attention can remarkably degrade the sentimental relevance of neutral words.

## 4.7 Control experiments

In this section, we firstly investigate the impact of λ on test accuracy. The results are presented in Fig. 5.

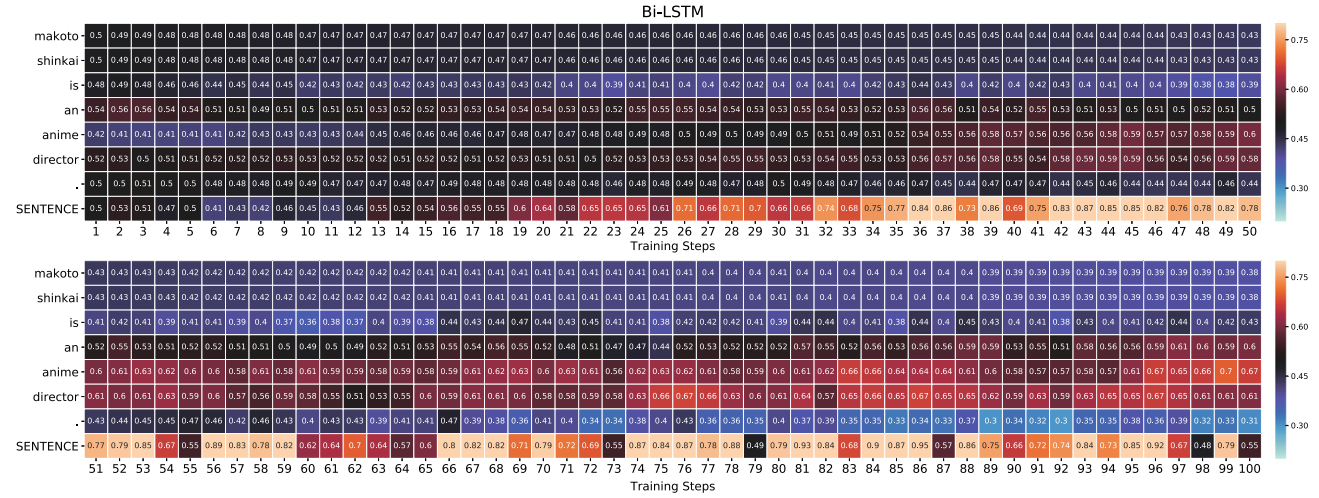
Figure 5 illustrates that the Bi-LSTM model achieves the lowest test accuracy on the MR data set. Although ATT-Bi-LSTM and PA-Net exhibit nearly the same performance, the test accuracy of PA-Net can be better than that of ATT-Bi-LSTM by adjusting different λ values. However, larger λ values may reduce test accuracy.

To verify that the polar attention mechanism can mitigate the sentimental biases during training, the bias degree (i.e., mean square error for the true biased words corpus) in MR-S is utilized to evaluate Bi-LSTM, ATT-Bi-LSTM, and PA-Net with different λ values at each training step. The results are presented in Fig. 6.

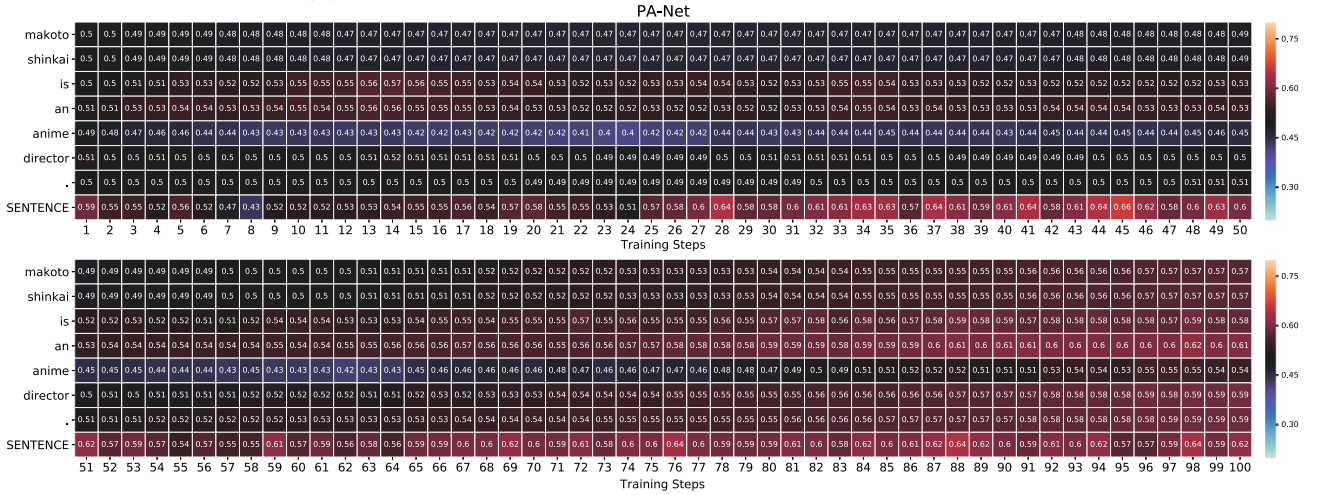




Journal: 41060 MS: 0231 TYPESET DISK LE CP Disp.:2020/8/4 Pages: 10 Layout: Large




**(a)** Changes in sentimental scores during Bi-LSTM training




**(b)** Changes in sentimental scores during PA-Net training

**Fig. 7** A comparison of the changes in sentimental scores of Bi-LSTM and PA-Net for a given sentence and each word it contains. The "**SENTENCE**" means that the sentence "***makoto shinkai is an anime director***"

In the figure, PA-Net remarkably reduces bias degrees compared with Bi-LSTM and ATT-Bi-LSTM. Compared with Bi-LSTM, ATT-Bi-LSTM can slightly mitigate the increasing trends of bias degrees. In our search ranges for hyperparameter $\lambda$, the performance of the polar attention mechanism is extremely stable.

Unexpectedly, the bias degrees of all the models increased with an increase in training steps. In order to observe this phenomenon intuitively, we present an enlightening example. Considering the same neutral sentence "***makoto shinkai is an anime director .***," we can obtain the sentimental scores of the sentence and each word it contains in each training step, and visualize the sentimental scores accordingly. Figure 7a, b present a comparison of the changes in sentimental scores of Bi-LSTM and PA-Net during training. As shown in Fig. 7, both models are accurate in predicting test samples at the beginning of training. With the increase of training steps, obvious sentimental bias appeared in Bi-LSTM, especially on the words "***makoto***," "***shinkai***," "***is***," "***anime***," "***director***," and the symbol ".." The predicted result of this neutral sentence tends to be positive. While our proposed PA-Net, as shown in Fig. 7b, has a significant mitigating effect on sentimental bias. The predicted results of all test samples are stable and remain in the range of 0.4 to 0.6. This comparison verifies the effectiveness of the polar attention mechanism in mitigating sentimental bias.

## 5 Conclusion

The skewed distribution of words on different sentimental categories incurs sentimental bias in sentiment analysis. Our work is the first attempt to mitigate word sentimental bias. A novel polar attention mechanism is proposed to explicitly







Journal: 41060 MS: 0231 TYPESET DISK LE CP Disp.:2020/8/4 Pages: 10 Layout: Large

model the word-level polarities together with a distance-based attention scoring module. It can reduce the extent of sentimental bias for neutral words while truly polar words are still attended. The experimental results on three benchmark data sets and their corresponding supplementary testing data verify the effectiveness of the proposed mechanism.

**Acknowledgements** This work is supported by Tianjin NFS (19JCZDJC 31300), AI Key Project of Tianjin (19ZXZNGX0050), and Frontier Science and Technology Innovation Projict (2019QY2404).

## References


1. Vaswani, A. et al.: Attention is all you need. In: NIPS2017, pp. 5998–6008 (2017)
2. Maas, A.L. et al.: Learning word vectors for sentiment analysis. In: ACL2011, ACL, pp. 142–150 (2011)
3. Dwork, C. et al.: Fairness through awareness. In: ITCS2012, ACM, pp. 214–226 (2012)
4. Goodfellow, I. et al.: Generative adversarial nets. In: NIPS2014, pp. 2672–2680 (2014)
5. Zhao, J. et al.: Learning gender-neutral word embeddings. arXiv:180901496(2018)
6. Tai, K.S. et al.: Improved semantic representations from tree-structured long short-term memory networks. arXiv:150300075 (2015)
7. Dixon, L. et al.: Measuring and mitigating unintended bias in text classification. In: AAAI2018, ACM, pp. 67–73 (2018)
8. Liu, L.T. et al.: Delayed impact of fair machine learning. arXiv:180304383 (2018)
9. Hardt, M. et al.: Equality of opportunity in supervised learning. In: NIPS2016, pp. 3315–3323 (2016)
10. Zhou, P. et al.: Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. arXiv:161106639 (2016)
11. Qian, Q. et al.: Linguistically regularized lstms for sentiment classification. arXiv:161103949 (2016)
12. Socher, R. et al.: Recursive deep models for semantic compositionality over a sentiment treebank. In: EMNLP, pp. 1631–1642 (2013)
13. Bai, S. et al.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:180301271 (2018)
14. Bolukbasi, T. et al.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: NIPS2016, pp. 4349–4357 (2016)
15. Dieterich, W. et al.: Compas risk scales: demonstrating accuracy equity and predictive parity. Northpoint Inc (2016)
16. Binns, R.: Fairness in machine learning: lessons from political philosophy. arXiv:171203586(2017)
17. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, pp. 77–91 (2018)
18. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805 (2018)
19. Edizel, B., Bonchi, F., Hajian, S., Panisson ,A., Tassa, T.: Fairecsys: mitigating algorithmic bias in recommender systems. Int. J. Data Sci. Anal. pp. 1–17 (2019)
20. Feldman, M., Friedler, SA., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 259–268 (2015)
21. Johnson, R., Zhang, T.: Deep pyramid convolutional neural networks for text categorization. ACL **1**, 562–570 (2017)
22. Joseph, M., Kearns, M., Morgenstern, JH., Roth, A.: Fairness in learning: Classic and contextual bandits. In: Advances in Neural Information Processing Systems, pp. 325–333 (2016)
23. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: ACL, pp. 115–124 (2005)
24. Wu, O., Yang, T., Li, M., Li, M.: hot lexicon embedding-based two-level lstm for sentiment analysis (2018)
25. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, RR., Le, QV.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems, pp. 5754–5764 (2019)
26. Zafar, MB., Valera, I., Gomez Rodriguez, M., Gummadi, KP.: Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In: WWW2017, pp. 1171–1180 (2017)
27. Zou, J., Schiebinger, L.: Design ai so that it's fair. Nature **559**(7714), 324–326 (2018)