The Journal of Supercomputing
https://doi.org/10.1007/s11227-021-03652-4

Entity-level sentiment prediction in Danmaku video interaction

Qingchun Bai, et al. ***[full author details at the end of the article]***

Accepted: 22 January 2021

Abstract
Sentiment analysis in Danmaku video interaction aims at measuring public mood in respect of the video, which is helpful for the potential applications in behavioral science. Once these sentiments are discovered, this feedback can help video creators improve the video quality and greatly enhance online users' watching experience. Predicting these entity-level sentiments is challenging because there is no publicly available dataset about entity-level sentiment analysis of Danmaku-enabled video comments. Furthermore, the targeted entity with skewed unbalance distribution in real-world scenarios, making the task more challenging, especially when the target entity only has positive (negative) emotional comments. In this case, applying previous aspect-level sentiment analysis models directly will introduce entity bias. In this paper, we propose a large-scale Chinese video comments dataset containing time-sync Danmaku comments and traditional video comments, targeting multiple entities and sentiments associated with each entity from popular video websites. We also propose a framework of entity-level sentiment analysis with two de-biasing models: hard-masking de-bias model and soft-masking de-bias model. This framework is defined by parallel neural networks to learn the representation of comments sentences. Based on the representations, our model learns a masking strategy for entity words to avoid overfitting and mitigate the bias. Our experiments on Danmaku-enabled video datasets show that the soft-masking model significantly outperforms comparable baselines, with a relative $F1$-score improvement of 9.33% compared to AEN-BERT and a relative $F1$-score improvement of 45.77% compared to Td-LSTM. Furthermore, experiments on different distribution bias of entity demonstrate that our proposed model can achieve competitive performances. The findings of this research have implications for measuring public sentiment for entities mentioned in a specific video domain. It can also be used as a benchmark dataset for aspect entity sentiment detection methods.

Keywords Sentiment analysis · Video comments · Danmaku · Aspect-level sentiment analysis

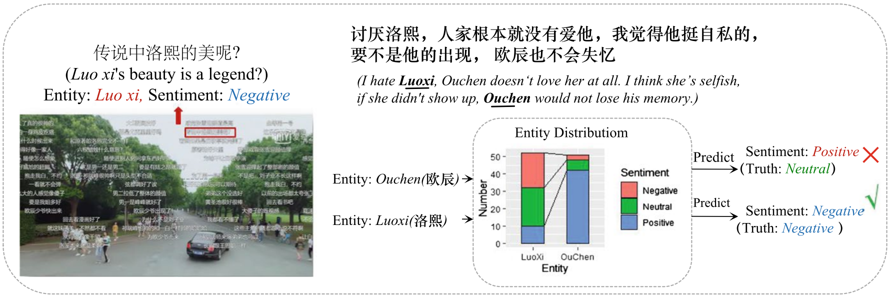


Fig. 1 An illustration of our task. Left is an example of a Danmaku-enabled video. Users can overlay anonymous comments on the video content, and comments are ranged in time order. Entity-level sentiment analysis focuses on predicting sentiment pertinent to a given entity that appeared in a sentence. Right is an example of sentiment distribution for an entity

1 Introduction

Danmaku-enabled videos present scrolling marquee comments overlaid directly on top of the videos, which is considered information-seeking and emotion-venting channel [23]. Recent researches [6, 14] have seen a growing interest in understanding users' comments on Danmaku-enabled videos. For example, users' interactions with Danmaku-enabled videos can provide immediate feedback to the video creators to improve video quality [23, 43], but also greatly enhance the online users' watching experience. These comments express people's emotions and tendencies, such as joy, anger, sorrow, joy, criticism and praise. Based on this analysis, potential users can browse these subjective comments to learn what the public thinks about an event or movie star. Therefore, fine-grained sentiment analysis on video is helpful for the potential applications in behavioral science.

As a subtask of fine-grained aspect-based sentiment analysis (ABSA), entity-level sentiment analysis focuses on predicting sentiment pertinent to a given entity that appeared in a sentence. For example, given that a user posts the following comment in a Danmaku-enabled video, "I don't like Luoxi because the girl doesn't like him. Also, I think he is quite selfish. If not for him, Ouchen would not lose his memory.", the goal of entity-level sentiment analysis is to predict this user's effects (positive, negative or neutral) of the two entities: *Luoxi* and *Ouchen* (see Fig. 1 for illustration). To achieve fine-grained analysis on Danmaku-enabled videos, a variety of reviews corpora [2, 14, 21, 38] has been released in video-based research to analyze the specific content of the video. Within this source, one can find discussions on video content, actors' performance, or personal information, all of which can be relevant to the behavioral science [40]. However, most previous work analysis comments directly without considering entity-level analysis. Meanwhile, there is no existing publicly available dataset for entity-level sentiment analysis about Danmaku. To achieve better sentiment representation, current research has developed a neural network-based model to capture the importance of contextual information related to affect dimensions in entities; for example, a recent work proposed a framework that enables the neural network model to pay more attention to the local context words leveraging BERT-shared layer [42]. While this line of work has achieved the state-of-the-art performance results, the self-attention-based model such as BERT may reinforce bias pertinent to an unbalanced entity from the dataset. In Danmaku-enabled video, the targeted entity with skewed unbalance distribution, simply applying these BERT-based models in entity-level sentiment analysis might reinforce bias (and even introduce additional bias) embedded in the subjective datasets such as users' comments on Danmaku-enabled videos (for example, in Fig. 1b).

In this paper, we propose a new framework for entity-level sentiment prediction in Danmaku video comments. In our framework, we learn the representation of comments sentences by using parallel neural networks. Based on the representations, our model learns a masking strategy for entity words to avoid overfitting and mitigate the bias. The primary idea is that if we can mask the word directly, sentiment prediction bias of entities could be mitigated accordingly. However, this approach ignores the entity semantic directly. Hence, we can use a dynamic masking strategy to avoid paying too much attention to a specific entity. Therefore, in this task, we propose and experiment with two de-biasing models: hard-masking de-bias model and soft-masking de-bias model, intending to mitigate sentiment prediction bias of entities that appeared in users' comments. Specifically, hard masking de-bias model uses a formidable entity masking strategy to mask a specific entity. The soft-masking de-bias model uses a dynamic masking strategy that is designed to avoid paying too much attention to a particular entity. To compare our proposed models with the state-of-the-art models, we construct real-world Danmaku video datasets and conduct experiments on these data. Overall, our work contributes to the following area:

- We present and release a new Chinese movie review dataset[1] consisting of two contemporary data types: comments and Danmaku with annotated entity-level sentiment labels. These datasets provide a new perspective for entity-based sentiment classification, making the dataset ideal for exploring the interaction of sentiment classification on video comments.
- We propose and experiment with two de-bias models (hard-masking de-bias model and soft-masking de-bias model) to absorb entity unbalance and prevent overfitting.
- We conduct extensive experiments on two kinds of Danmaku video datasets. The results show that our model achieves competitive performance and can serve as a baseline for future studies on de-biasing entity-level sentiment analysis tasks.

2 Related work

In this section, we review studies on Danmaku video interactions, aspect-based sentiment classification and related datasets to identify the research gap in the area of entity sentiment prediction in Danmaku video interaction.

[1] https://github.com/adableau/DanSentiment.

2.1 Danmaku video interaction

Recent research on Danmaku video has focused on a variety of topics. For example, some research has focused on the user experience of Danmaku video by analyzing linguistics patterns in Danmaku comments [7, 28, 39, 40]. There has also been research on video tagging [38], content analysis [2, 14], highlights detection [37], recommendation [6] and comment generation [22, 29]. Since video comments contain slang, emoji and other noise, learning semantic representation has been a challenge in analyzing these comments. To address this challenge, researchers have also focus on learning specializing embeddings [21].

These previous works have provided insights for linguistics patterns expressed in Danmaku video interactions and methods to detect and generate comments automatically. Within this source, one can find discussions on video content, actors' performance or personal information, all of which can be relevant to the behavioral science [40]. However, little research has focused on the application of entity-level sentiment analysis in Danmaku comments. Meanwhile, there is no existing publicly available dataset for entity-level sentiment analysis about Danmaku.

2.2 Aspect-based sentiment analysis

Past research on entity sentiment analysis has primarily focused on aspect-based sentiment analysis (ABSA) task [3–5, 8, 20, 41, 45–47]. ABSA Task focuses on modeling the interaction between sentences and aspect words to achieve competitive results [44]. For example, an earlier work proposed a deep memory network model to capture the importance of context word when inferring the sentiment polarity of an aspect [35]. In more recent work, research has focused on learning better representational performance and intuitive interpretation of sentences [1, 33]. For example, local context focus [42] mechanism has been proposed for aspect-based sentiment classification based on multihead self-attention. The work of [1] helps improve system efficiency, analyzed using experimental results of error rate, precision, recall and accuracy. The work of [48] focuses on training a pretraining model on cross-domain sentiment analysis tasks, which achieves competitive performance. As a fundamental subtask of aspect-level sentiment analysis (ABSA), entity-level sentiment analysis attracted increasing attention recently due to its broad applications. Existing research on entity-level sentiment analysis has not considered machine bias caused by unbalancing labels. As suggested in recent works, the bias in neural network-based models can be problematic since it can reinforce existing stereotyping in the groups' representation, such as gender bias. To address this issue, one line of research in AI bias has focused on quantifying machine bias and methods to mitigate the risk of bias during model training [13, 25–27]. For example, research [17] has proposed a template-based method to quantify gender bias in BERT. Recent researchers have also focused on developing de-biasing methods such as leveraging counterfactual data substitution and names intervention [26]. These previous works have insights for developing quantifying, detecting and reducing biases in entity-level sentiment analysis tasks. Few studies have focused on de-bias entity sentiment in Danmaku video interaction. Therefore, in this work, we focus on de-biasing entity-level sentiment prediction in Danmaku video interactions. Past research on entity sentiment analysis has primarily focused on aspect-based sentiment analysis (ABSA) task However, most previous work analysis comments directly without considering entity-level analysis. Meanwhile, there is no existing publicly available dataset for entity-level sentiment analysis about Danmaku.

2.3 Danmaku sentiment analysis

There is also some corpus proposed for the task of ABSA. Corpora can consider as source of data (e.g., Twitter [11]) or domain-specific such as SemEval-2014,[2] SemEval-2015[3] and SemEval-2016.[4] Fine-grained entity-level sentiment is also proposed such as Sentihood [18, 30] and Mitchell,[5] which is extracted from a question answering (QA) platform. However, in the area of the movie review domain, previous works only have focused on sentiment analysis such as IMDB Corpus [24]. Studying Danmaku video interaction is essential, and a variety of reviews corpora [2, 14, 21, 38] has been released in video-based research to analyze the specific content of the video. Within this source, one can find discussions on video content, actors performance or personal information, all of which can be relevant to the behavioral science [40]. Overall, there are two drawbacks of existing publicly available datasets. First, almost all the mentioned works of textual movie reviews such as IMDB[6] are annotated as positive and negative, without entity-level annotation [24]. Second, a dataset combining Chinese movie reviews textual sentiment is absent. Therefore, we propose a Chinese movie review dataset consisting of two contemporary data types: comments and Danmaku with annotated entity-level sentiment labels. Our analysis of DanSentiment indicates the targeted entity with skewed unbalances distribution, making the dataset ideal for exploring interpretability of sentiment classification.

3 Danmaku corpus

Comparing with traditional video reviews (e.g., movie reviews), the unique properties of Danmaku are time-synchronized and fine-grained. Danmaku can be seen as a (noisy) user-generated script of the underlying video stream.

Preprocessing. According to the prepossess of [15, 36], annotate the large-scale Danmaku and comments data directly will cause huge human resource consumption. We randomly sample some popular videos of both comments and Danmaku

[2] http://alt.qcri.org/semeval2014/task4/.

[3] http://alt.qcri.org/semeval2015/task12/.

[4] http://alt.qcri.org/semeval2016/task5/.

[5] http://www.m-mitchell.com/code/index.html.

[6] https://www.imdb.com/.

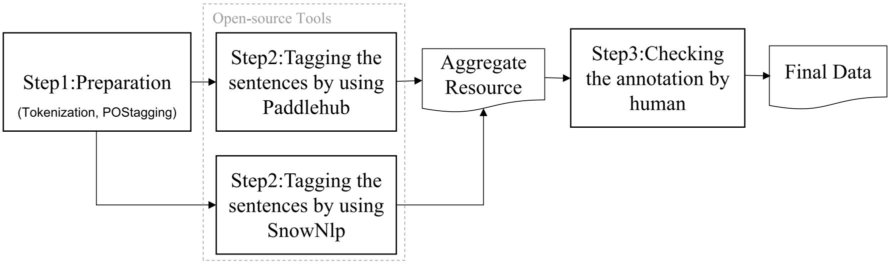


Fig. 2 Preprocessing annotation steps interaction flow

and then filter out the repeat and without entities comments. The preprocessing framework flow is defined in Fig. 2, which involves four steps.:

- We have collected two kinds of movie review corpus, namely Danmaku and comments from popular video website iQiYi.[7] We preprocess the data by Tokenization, POStagging as preparation data.
- We use open-source tools and algorithms to label the data automatically (e.g., Paddlehub,[8] SnowNlp[9]). Then, we aggregate the results to reduce checking complexity.
- We recruit six undergraduate students in a Chinese research university who love watching TV to check the annotation tagging results. For this case, we use the following litmus test: If the Danmaku or comment contains negative emotions for the given entity (e.g., dislike, angry, hate, abuse), it belongs to the negative class. If the Danmaku or comment contains positive emotions for the given entity (e.g., like, happy, praise), then it belongs to the positive class. Otherwise, there is no emotional tendency, and it belongs to neutral. For example, the following comment "*Zhangsan* is not good at acting, *Lisi* is so beautiful" can be classified onto the negative class for "*Zhangsan*" (Entity) and positive class for "*Lisi*" (Entity).
- We finally collect the labeled corpus as datasets for the entity-level sentiment classification task.

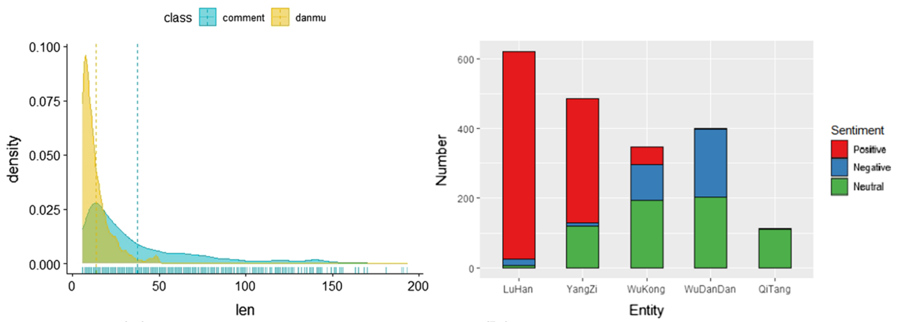
Table 1 shows the coding book for categorizing different sentiment polarity of Danmaku and comments. We use the majority voting rules [12] to choose the correct label. If more than five annotators voted the same label, we set the tag as ground truth. Finally, we receive 13,903 comments and 3415 Danmaku as our experiment datasets.

Data statistics. Table 2 summarizes the statistic information compared with other related corpora. In this section, we describe our observations to present the feature

[7] https://iqiyi.com.
[8] https://github.com/PaddlePaddle/PaddleHub.
[9] https://pypi.org/project/snownlp/.

Table 1 Coding book for entity-level sentiment analysis on Danmaku and comments

Entity (sentiment)	Selection criteria	Example
Kunkun (neutral)	(Hint) Hint the keyframe of the video	*Kunkun* will come, hahahahah
Zhou Ziyang (neutral)	(Need) Ask some questions	Who can tell me how to contact with *Zhou Ziyang*
Eason Chan (positive)	(Thought) Positive opinion for the entity	*Eason Chan* is so cute
Zheng Zhi (negative)	(Thought) Negative opinion for the entity	I hate *Zheng Zhi*, he is not honest
Lisa (neutral)	(Quarrel) Users quarrel with each other, but without negative opinion for the entity	Are you stupid? Why don't you like Lisa. A bunch of keyboard-warriors came out, XiaoZhan's fans



(a) Length distribution. **(b)** Entity-level Sentiment Distribution.

Fig. 3 Data statistics. In (**a**), five of the most frequent entities sentiment distribution are counted. In (**b**), five of the most frequent entities sentiment distribution are counted

Table 2 Comparison among our corpus and other human-labeled ABSA task dataset

Dataset	Pos/Neu/Neg	#Aspect	#Avg.len	Train/Test	All
Mitchell-en	695/2259/269	2355	18.44	–	2350
Twitter	1734/3491/1733	195	18.83	6248/710	6958
SemEval-2015	1652/98/749	744	13.63	1645/845	2499
SemEval-2016	2268/145/953	960	13.52	2507/859	3366
Laptop14	1328/994/629	1328	16.77	2631/638	2951
Restaurants14	2360/829/1533	1711	15.97	1462/1120	4722
Danmaku	1230/1643/542	576	14.05	2732/683	3415
Comments	4808/6820/2275	7392	42.64	11122/2781	13903

Pos/Neu/Neg represents positive, neutral and negative, respectively. Single-Entity (Multi-Entity) represents the number of single entities (multiple entities) in the dataset, #Entity/Aspect represents the total number of entity (aspect words), and #Avg.len represents the average tokens (words in English corpus) per sentence of our corpora. First, our corpus is the first large-scale manually annotated corpus targeted in the area of video comments. Second, in our corpus, Danmaku sentences are ephemeral and short (avg. 14.05 tokens) compared to another corpus. This setting makes the task challenging to discover sentiment to a specific entity. Third, commentators are often affected by the actors or actresses, and the emotional distribution of an entity is often too unbalanced in an online video website. We show an example in Fig. 3.

4 Methodology

In this section, we focus on entity-level sentiment classification on video comments. Formally, in a Danmaku-enabled video, let $C = \{w_1, w_2, \dots, w_{|C|}\}$ be the an input time-sync comment sequence with $|C|$ words and $E = \{w_s, \dots, w_e\}$ be an entity that is a span in the sentence C, where s and e are the start and end indices of the entity. Each comment C contains one or more entities. This task aims to predict the sentiment polarity $y \in Y$ of the entity E in sentence C, where $Y = \{0, 1, 2\}$ represents the set of sentiment labels, including negative, neutral and positive. Our primary goal is to learn a probabilistic sentiment classifier, $p(y \mid x, e)$.

Unlike existing datasets of this task, the sentiment distribution of entities in Danmaku video comments is extremely unbalanced, which leads to the sentiment bias embedded in the data. For example, entity "Luoxi" (Fig. 1) has 62 comments co-occurring with negative but 0 with positive. The base model (Sect. 4.1) trained on such skewed comments data tends to predict a negative sentiment for entity "Luoxi" without fully considering the semantic knowledge of context. Additionally, we find that the model performs poorly for entities with few or no occurrences. Our method operates on the premise that the training data distribution corresponding to a few entities suffers from entity bias.

Based on these observations, we propose two approaches, namely hard-masking de-bias model and soft-masking de-bias model. In this section, we first introduce the base model for entity-level sentiment analysis. We refer to this model as local-based model (Sect. 4.1). Then, we propose a hard-masking de-bias model which simply masks the entity via a hard entity masking strategy (Sect. 4.2). While this model is efficient and straightforward to implement, it ignores the entity information completely. Therefore, we propose a soft-masking de-bias model, where a dynamic masking strategy is designed to avoid paying too much attention to the entity (Sect. 4.3). In this section, we discuss the tasks for entity-level sentiment classification based on Danmaku corpus.

4.1 Base model

To learn representations of comments and entity, we focus on the recent work of bidirectional encoder representation from transformers (BERT) model to learn useful features. Our approach heavily borrows from past work BERT-SPC [32], which regards sentence and entity as a pair and inputs them into the BERT model. We do

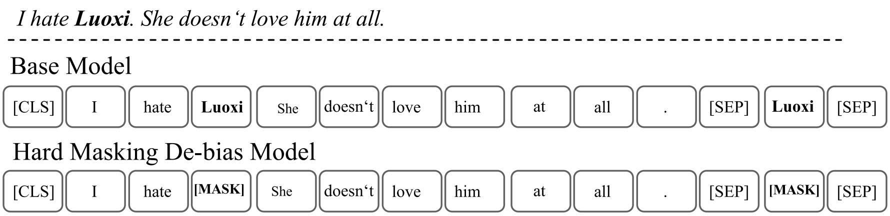


Fig. 4 An example of hard entity masking. Different from the entity-aware model, the entity "Luoxi" is replaced by the specific token [MASK] as the input to reduce the influence of bias

so to learn the representations and then use the learned representations to calculate the interaction between the sentence and entity for a classifier to predict entity-level sentiment label. In particular, we concatenate the comment C and entity E to obtain the input sequence.

$$x = \{[CLS], w_1, w_2, \dots, w_{|C|}, [SEP], w_s, \dots, w_e, [SEP]\}$$

Then, we feed the input x into BERT and obtain the last hidden of [CLS], $h_{[CLS]}$, for sentiment classification. Finally, we use a full connection layer to predict the sentiment probability of the given entity $P(y|C,E) = \text{Softmax}(rW^T + b)$, where $W \in \mathbb{R}^{|Y|*d_h}$ and $b \in \mathbb{R}^{|Y|}$ represent the weighted parameter. The off-the-shelf pre-trained model with whole word masking for Chinese BERT[10] is used here [9].

We refer to this model as a local-based model, which focuses on the local information (e.g., entity) and suffers from sentiment bias. While this model can learn the local information about a given entity, it tends to ignore the global information (e.g., contextual information).

4.2 Hard-masking de-bias model

Intuitively, to avoid the model consider entity words as emotional features, a simple and efficient approach is to ignore the entity in the comments. Inspired by BERT architecture, which is trained by masked language modeling to predict the masked token, we propose a hard entity masking strategy. In particular, the entity E is replaced with a special token [MASK] to avoid bias. Given an input sentience x=$\{[\text{CLS}], w_1, w_2, \dots, w_{|C|}, [\text{SEP}], w_s, \dots, w_e, [\text{SEP}]\}$, we replace the $E = \{w_s, \dots, w_e\}$ by a special token [MASK] and obtain the corrupted sentence

$$x' = \{[\text{CLS}], w_1, \dots, w_{s-1}, [\text{MASK}], w_{e+1}, \dots, w_{|C|}, [\text{SEP}], w_s, \dots, w_e, [\text{SEP}]\}$$

Then, we feed x' into BERT with fine-tuning to learn the representation. To be specific, the sentiment polarity prediction for a given entity is calculated in the following steps:

[10] https://github.com/ymcui/Chinese-BERT-wwm.

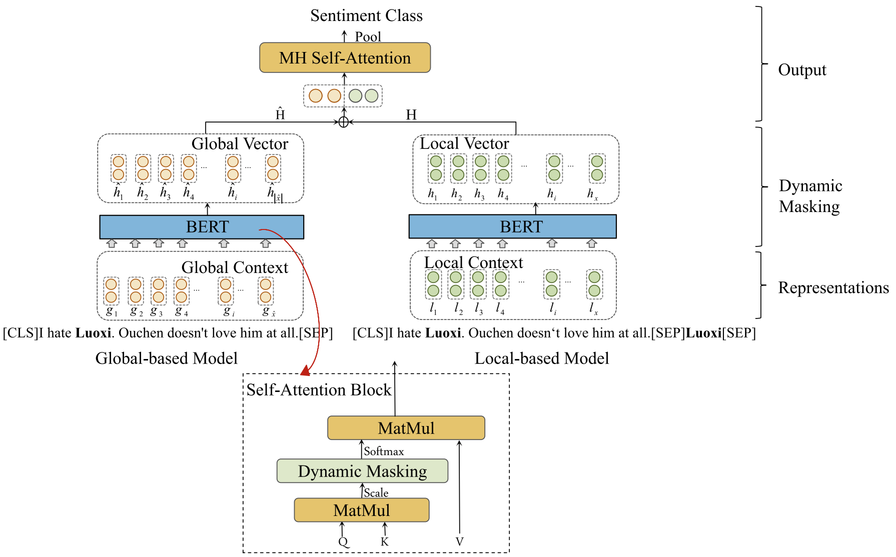


Fig. 5 Overview architecture of our proposed soft-masking de-bias model

1. We replace entity words in x with a [MASK] token to obtain the corrupted sentence x'. For example, in Fig. 4, for comment "*I hate Luoxi , She doesn't love her at all.*", we obtain the corrupted comment "[CLS] *I hate* [MASK], *She doesn't love her at all.* [SEP] [MASK] [SEP]."
2. We feed the masked sentence $\hat{x}$ into BERT model to learn sentence representation. In particular, the last hidden of [CLS] is adopted as the feature representation of the sentence $r \in \mathbb{R}^{d_h}$, where d_h is the hidden state's dimension of BERT.
3. We use a full connection layer to predict the sentiment probability of the masked entity $P(y|C, E) = \text{Softmax}(rW^T + b)$, where $W \in \mathbb{R}^{|Y| * d_h}$ and $b \in \mathbb{R}^{|Y|}$ represent the weighted parameter.

Hard entity masking strategy can reduce the bias of entity effectively by replacing the entity with [MASK] directly. However, this approach suffers information loss pertinent to the entity. To address this problem, we propose a soft-masking de-bias model, which considers both the entity-aware and entity-unaware models by dynamic masking (Fig. 5).

4.3 Soft-masking De-bias model

In this section, we propose a modified soft-masking BERT model (soft-masking) to absorb unbalance from the entity based on two observations: (1) preventing overfitting from the training data since reviewers prefer to give negative sentiment to the negative character. This phenomenon makes the sentiment distribution of entities in video comments incredibly unbalanced. (2) "Masking" the entity words is a simple and intuitive approach way to absorb unbalance problem. This approach will lose too much information (defined as hard masking)). To tackle this problem, we first introduce two parts, the local-based model and the global-based model, to capture the entity and context information simultaneously. Formally, given a sentence C and entity E, we construct two sequences to learn the local-based model and global-based features. We define local-based model inputs x as $\{[\text{CLS}], C, [\text{SEP}], E, [\text{SEP}]\}$. Note that local-based model is used to learn the representation of context and entity information. The definition of inputs x is entity sensitive. Global-based model is used to capture context information totally, which is entity insensitive by inputting the $\hat{x} = \{[CLS], w_1, w_2, \dots, w_{|C|}, [SEP]\}$. In the Global-based model, the representation will focus on global information because it does not take the entity as input. Then, we define the model of **D**ynamic **M**asking (DM) mechanism in scaled dot-product attention by masking the entity in input sequence dynamically. From this way, we can learn the entity information, which is defined as $\text{Self-Attention}(Q, K, V) = \text{softmax}(\frac{QK}{\sqrt{d_k}} + \text{DyMask}(x, p))V$. Here, DyMask is a one-like vector with length $|x|$ and the value of entity's position is set as 0 with a probability p. Finally, we train a classifier to learn the parameters with the cross-entropy loss.

Local and Global Representations First, we obtain the sentence representations of local-based and global-based inputs. In the case of enables multiple contents inputs, we use BERT model as backbone network to get the representations of sentence [10]. The same setting as base model, we input the x into entity-aware model to obtain local representation $H = [h_1, h_2, \dots, h_{|x|}]$ and input the $\hat{x}$ into global-based model to obtain global representation $\hat{H} = [\hat{h}_1, \hat{h}_2, \dots, \hat{h}_{|\hat{x}|}]$. In particular, local representation H focuses too much on the entity, while global representation $\hat{H}$ ignores the entity information completely.

Dynamic Masking. To force global-based and local-based model learn the entity and context information, respectively, a dynamic masking strategy is proposed to combine them adaptive. Therefore, we design local dynamic masking and global dynamic masking for local-based and global-based models, respectively, to softly mask the entity. We force the local-based model to learn the context information and force the global-based model to learn entity information via these dynamic masking strategies. Different from the standard multihead model, we integrate the local dynamic masking strategy into attention mechanism for local-based model. Similarly, we integrate the global dynamic masking strategy into attention mechanism for global-based model. Specifically, for local dynamic masking in local-based model, we mask the entity in x with an probability p via our masking function $\text{DyMask}(x, p) = (1 - \text{att\_mask}) * -10{,}000.0$. Here, att_mask is a one-like vector with length $|x|$ and the value of entity's position is set as 0 with a probability p. This is effectively the same as removing the masked entity. Then, we calculate the attention score by integrating dynamic masking into dot product between multihead attention, which is defined as

$$
\begin{aligned}
&\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\cdots,\text{head}_n)W_h \\
&\text{head}_i = \text{Self-Attention}(QW_i^Q, KW_i^K, VW_i^V) \\
&\text{Self-Attention}(Q,K,V) = \text{softmax}\left(\frac{QK}{\sqrt{d_k}} + \text{DyMask}(x,p)\right)V
\end{aligned}, \tag{1}
$$

where $Q = K = V = H$. d_k is the dimension of Q and K, n is the number of heads. W_i^Q, W_i^K and W_i^V are the learnable parameters. The final local representation $H = \text{MultiHead}(Q,K,V)$.

Similar to local-based model, we train dynamic masking in global-based model. We compute the attention weight by integrating the global masking and obtain the final global representation $\hat{H}$. Finally, we obtain the last hidden states in local-based and global-based models which are modeled by BERT as sentence representations.

Output Layer. In the output layer, we concatenate the local representation H and global representation $\hat{H}$, $H^+ = [h_1^+, h_2^+, \dots, h_{|x|}^+] = [h_1;\hat{h}_1, h_2;\hat{h}_2, \dots, h_{|x|};\hat{h}_{|x|},]$. Then, we input it into a standard multihead attention to learn the final representation.

$$
H^+ = \text{MultiHead}(H^+, H^+, H^+), \tag{2}
$$

Then, the state vector of [CLS], $r^+ = h_1^+$, is obtained as the sentence representation to predict the sentiment class of each entity. For the entity E, we compute the probability of sentiment class as

$$
P(y|C,E) = \text{softmax}(W_c r^+ + b_c), \tag{3}
$$

where $W \in \mathbb{R}^{|Y|*d_h}$ and $b \in \mathbb{R}^{|Y|}$ represent the weighted parameter.

Training The model is trained by the standard gradient descent algorithm with the cross-entropy loss, which is defined as:

$$
\mathcal{L} = -\log P(y|C,E). \tag{4}
$$

where C represents sentence context and E represents entity span context.

5 Experiment

5.1 Baselines and experimental setting

To validate the effectiveness of the proposed model, we compare with the state-of-the-art aspect-level sentiment classification models in previous studies, namely MemNet [35], Td-LSTM [34], TNet [19], BERT-SPC [32], AEN-BERT [32] and LCF-BERT [42].

Due to space limitation, we do not introduce all the methods here. The default hyperparameter settings in baseline models are as follows: The word embedding layer in the baseline methods was in English; hence, we apply pretrained 200-dimension Chinese vector representations [31] instead of Glove. We set training batch size = 128 and hidden dim = 300, epochs = 100. All models are trained on two datasets, Danmaku and comments.

The default hyperparameter settings in our model as follows: In detail, in validation process, we split the training data into five new nonoverlapping datasets randomly, In the training process, we select four datasets as the training set and the remaining one as the test set. Then, we feed them into train model with mini-batch. The cross-validation was repeated for K times ($k = 5$ in experiments), and the average accuracy of K times was taken as the evaluation index of the final model. $E = \frac{1}{10}\sum_{i=1}^{10} E_i$ we use Adam stochastic optimization optimizer [16] with the setting learning rate $2e-5$ to train the model. The model is trained by mini-batch with batch size 64, and epochs = 20. We add dropout with rate 0.6 to avoid overfitting. All the parameters are l_2-regularized, with the hyper-parameter $\lambda = 10^{-2}$. We utilize the Chinese-Roberta model [9] to initialize the parameters, which are based on whole word masking tricks that mask the whole word instead of masking Chinese characters. The dim in hidden layer is 768, 12-heads, and the number of all parameters is 110M. All the hyperparameter settings for hard masking and soft masking are the same.

5.2 Results

Experiments in this section intend to investigate the following questions.

1. How about the sentiment classification performance of our proposed methods on Danmaku-enabled video comments?
2. Can we mitigate the bias caused by the unbalanced entity-level sentiment distribution?

To answer question 1, we evaluate performance compared with the state-of-the-art baselines. For question 2, we design experiments on synthetic data with unbalanced sentiment label for the entity: we do so because different distribution of entity sentiment can better evaluate the de-bias performance of our method.

5.2.1 Comparison with state of the art

For the sentiment classification task, we use Accuracy (Acc.), Macro-Precision (P), Macro-Recall (R), and Macro-$F1$ (F) as evaluation metrics to evaluate the performance of our approach and baselines. The metric accuracy calculates the correctly classified ratio for test data set, where $\mathrm{Acc} = \frac{\#\mathrm{correct}}{\#\mathrm{total}}$. Macro-average metrics calculate the precision, recall, $F1$ of each category separately, then average the scores out to get the final Macro-PRF value. In automatic classification metrics, macro-averages (Macro-Precision, Macro-Recall and Macro-F) are benefit for unbalance label problem of the dataset. For each category, we first calculate PRF scores separately and then average them as the final PRF value, which is defined as,

Table 3 Macroscore of sentiment classification experiment results on comments datasets

Comments data		Acc.	P	R	$F1$
Baselines	MemNet [35]	0.7488	0.7444	0.6981	0.7153
	Td-LSTM [34]	0.7074	0.7306	0.5963	0.6024
	TNet [19]	0.7819	0.7776	0.7351	0.7518
	BERT-SPC [32]	0.8670	0.8521	0.8712	0.8606
	AEN-BERT [32]	0.8233	0.8224	0.7890	0.8031
	LCF-BERT [42]	0.8749	0.8626	0.8841	0.8720
Our models	Hard-Masking	0.7931	0.8052	0.7364	0.7601
	Hard-masking (entity mask in test-set)	0.8849	0.8762	0.8753	0.8757
	Soft-Masking	$0.8888^{‡}$	$0.8766^{‡}$	$0.8799^{‡}$	$\mathbf{0.8781}^{†‡}$

Models in the first part are baseline methods which include neural network-based and BERT-based methods. The second parts are BERT-based methods of our work. The marker † refers to $p < 0.05$ by comparing with BERT-SPC in paired t test, and the marker ‡ refers to $p < 0.05$ by comparing with Td-LSTM in paired t test

Table 4 Macroscore of sentiment classification experiment results on Danmaku datasets

Danmaku		Acc.	P	R	$F1$
Baselines	MemNet [35]	0.6652	0.6466	0.6344	0.6399
	Td-LSTM [34]	0.6691	0.6486	0.6537	0.6512
	TNet [19]	0.7734	0.7551	0.7307	0.7412
	BERT-SPC [32]	0.8255	0.8740	0.8804	0.8770
	AEN-BERT [32]	0.8676	0.8579	0.8588	0.8572
	LCF-BERT [42]	0.8859	0.8728	0.8690	0.8703
Our models	Hard-Masking	0.8684	0.8493	0.8535	0.8507
	Hard masking (entity mask in test set)	0.8713	0.8586	0.8570	0.8572
	Soft masking	$0.8888^{‡}$	$0.8805^{‡}$	$0.8799^{‡}$	$\mathbf{0.8798}^{‡}$

The marker † refers to $p < 0.05$ by comparing with BERT-SPC in paired t test, and the marker ‡ refers to $p < 0.05$ by comparing with Td-LSTM in paired t test

$$\begin{aligned}\text{Macro-P} &= \frac{1}{n}\sum_{i=1}^{n} P_i \\ \text{Macro-R} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\ \text{Macro-F1} &= \frac{2\times\text{Macro-P}\times\text{Macro-R}}{\text{Macro-P}+\text{Macro-R}}.\end{aligned} \tag{5}$$

where n represents the segment parts of a video. Table 3 summarizes the overall experimental results on comments datasets of these models, and Table 4 reports the results on Danmaku datasets with all baselines.

Based on the experimental results, we have the main observations as follows.

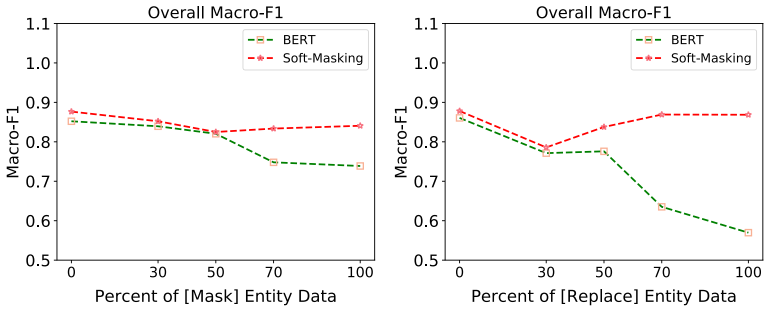
1. The soft-masking approach significantly outperforms all baselines on both Danmaku and comments datasets, indicating that our method can better learn the contextualized representations. Meanwhile, the dynamic masking attention mechanism can effectively avoid overfitting of entity, which demonstrates that our proposed model can better deal with entity bias problem in Danmaku video interaction comments data.
2. As listed at Table 4, BERT-based approaches perform better than embedding-based approaches, which can capture more structural information without requiring handcrafted features, thus showing their benefit for this task.
3. As shown, BERT-based approaches perform better than embedding-based approaches, which can capture more structural information without requiring handcrafted features and show their benefit for this task.
4. Compared with comments data, the BERT-based performance in Danmaku shows significantly superiority, which shows that comments sentiments is more complex than Danmaku in this task.
5. We also notice that LCF-BERT model also achieves competitive performance compared with other baselines on comments dataset, showing that the design of local context focus can capture long-term internal dependencies. On the contrary, in Danmaku dataset, the performance of BERT-SPC (normal BERT with fine-tune model) is slightly better than LCF-BERT, showing that the design of local context attention is not good at modeling short noisy Danmaku text.

5.2.2 Performance on different distribution of entity

To model the unbalance bias in Danmaku comments, we test extension experiments on synthetic examples. We do so because we have observed the effectiveness of our proposed entity-level sentiment classification in real Danmaku and comments. However, in order to test the robustness of the de-bias model, and to facilitate a more profound understanding, we conduct additional experiments to analyze the performance when entities with unbalance distribution in the test set. By testing on unbalance data, we aim to verify that our proposed model can mitigate the entities' bias and predict authentic opinion expressions.

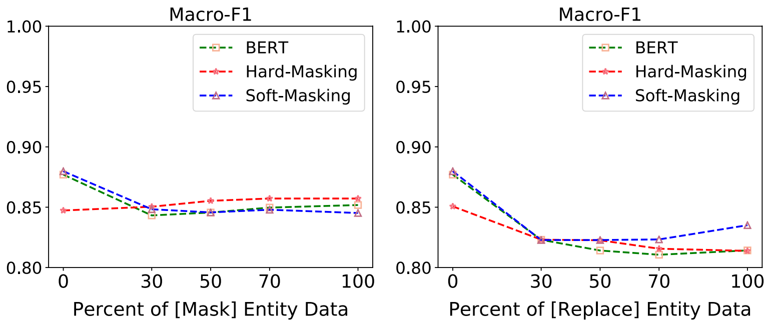
Synthetic data generation The duplication or similarity of comments will produce biased corpora since Danmaku and comment data have a high cover of entity sentiment label. In this section, we describe how to generate synthetic data in the scene of entity label unbalance. We use two intervention strategies on both Danmaku and comments datasets: masking entity and replacing entity. The details of the strategies are defined as follows:

1. *Masking entity*. For entity words in the training corpus, we replace them with a unique token [MASK] on each sentence. We increase the masked entities with different percent in dataset. By masking supervision about co-occurring entities, we intend to understand better what is the model focus.



(a) Masking entity **(b)** Replacing entity

Fig. 6 Macro-$F1$ score of entity-level sentiment classification on comments dataset. X-axis represents the percent of substitution entity, where 30% represents the composition ratio of the test data is 30% replaced (masked) entity data and 70% with standard data. Y-axis represents Macro-$F1$ score

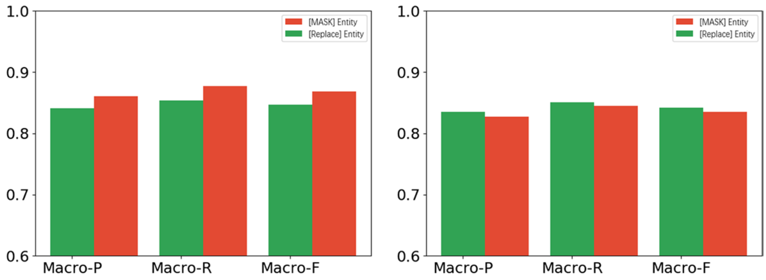


(a) Masking entity **(b)** Replacing entity

Fig. 7 Entity substitution results on Danmaku dataset. Macro-$F1$ score of entity-level sentiment classification on comments dataset. X-axis represents the percent of substitution entity, where 30% represents the composition ratio of the test data is 30% replaced (masked) entity data and 70% standard data. Y-axis represents macro-$F1$ score

2. *Replacing entity.* For entity words in the testing corpus, We replace them with a random token instead of masking it directly, where the replaced entities (random token) do not occur in the training set. In this way, the goal is to test the robustness of the performance on an unseen entity.

Analysis on comments data. We conduct experiments to determine whether the hard-masking de-bias model is useful for detecting entity sentiment on video comments data. The main results based on bias entity substitution are shown in Fig. 6. In comments data, with the increase in unknown entity in test set, the performance of both BERT and soft masking is dipped. For the BERT-based model, the dipped percent of the masking strategy is slightly less than the replacing approach, which shows that traditional BERT puts forward much attention on the entity. By comparing the results of masking and replacing strategy (Fig. 8),



(a) Comments Dataset. **(b)** Danmaku Dataset.

Fig. 8 Comparison of substitution strategies. The green bar represents the replacing entity strategy, and the red bar represents the replacing masking strategy. Masking strategy performs better than replacing strategy in comments data. Contrary to comments data, replacing strategy weakly performs better than masking entity in Danmaku data

we found that replacing strategy performs better than masking entity in comments data.

Analysis on Danmaku data. Different from comments data, Danmaku data are relatively short, with a more intuitive sentiment. We conduct further experiments on Danmaku data to explore the entity influence in short, noisy data. The main results based on bias entity substitution can be found in Fig. 7. **First**, in masking strategy (Fig. 7a), hard-masking de-bias model (hard masking in the figure) achieved state-of-the-art performance when masked data increased. **Second**, in replacing strategy (Fig. 7b), the performance of all models decreases when entity substitution incorporated and soft-masking de-bias model (soft masking) performance better than other methods. **Finally**, by comparing the results of masking and replacing strategy (Fig. 6b), we found that replacing strategy weakly performs better than masking entity in video Danmaku data (Fig. 8).

These experimental results further illustrate that using the current aspect-level sentiment classification model directly will cause entity bias in entity-level sentiment analysis task, especially in the case of an unbalance entity.

6 Conclusions and future work

In this work, we study an entity-level sentiment classification task for Danmaku-enabled video comments and propose a large-scale entity-level sentiment video interaction corpus. We believe that this corpus can provide a benchmark in the area of Chinese movie review analysis. Different existing corpus, this is the first video area entity-level sentiment prediction dataset. The corpus contains more than 15k entity-level sentiment sentences totally, with manual labeling. We further analyzed the entity unbalance phenomenon in Danmaku and comments and designed two entity-aware de-bias model to absorb entity unbalance and prevent overfitting. Experiments on real-world Danmaku and comments dataset show that our model outperforms state-of-the-art baselines on entity-level sentiment classification. Meanwhile, to mitigate the bias problem caused by an unbalanced entity, additional experiments are conducted to demonstrate the effectiveness of our proposed approach. Finally, tests on different distribution biases of entities show that our proposed model can achieve competitive performance.

In our empirical comparison, we observe that although the BERT fine-tune-based model performs well on the test set, there are still several limitations remaining unsolved. First, the syntactic information of the sentence is important for this task, while this work can only deal with semantic information and cannot model the syntactic dependency tree with complex structure. We plan to combine syntactic information into the pretrained model to get a better representation of sentences. Second, the domain knowledge is not studied well for entity-level sentiment analysis on video comments. We plan to add domain knowledge to learn video area commonsense knowledge. Third, the existing models are used for the pretrained model (embedding-based or BERT-based), which lack explanation about the model. Future research can be improved in the following aspects. We plan to predict how entity words drive the prediction of the model. We should be able to query our model and identify potential feature interactions to understand which features may be important in the model's decision strategy.

Acknowledgements The computation is performed in ECNU Multifunctional Platform for Innovation (001). The authors would like to thank Shijun Zhang and Kai Song for excellent technical support and data annotation support. The authors would also like to thank the editors and the reviewers for their high-quality and constructive comments. These suggestions and comments are important and useful.

References

1. Alarifi A, Tolba A, Al-Makhadmeh Z, Said W (2020) A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks. J Supercomput 76(6):4414–4429. https://doi.org/10.1007/s11227-018-2398-2
2. Bai Q, Hu Q, Fang F, He L (2018) Topic detection with danmaku: a time-sync joint NMF approach. DEXA 11030:428–435
3. Cai Y, Wan X (2019a) Multi-domain sentiment classification based on domain-aware embedding and attention. In: IJCAI-19, IJCAI, pp 4904–4910
4. Cai Y, Wan X (2019b) Multi-domain sentiment classification based on domain-aware embedding and attention. In: Kraus S (ed) IJCAI. ijcai.org, pp 4904–4910
5. Cao Y, Xu H (2020) Satnet: Symmetric adversarial transfer network based on two-level alignment strategy towards cross-domain sentiment classification (student abstract). In: AAAI, pp 13763–13764
6. Chen Y, Gao Q, Rau PL (2017b) Watching a movie alone yet together: understanding reasons for watching Danmaku videos. In: International Journal of Human–Computer Interaction
7. Chen X, Zhang Y, Ai Q, Xu H, Yan J, Qin Z (2017a) Personalized key frame recommendation. In: ACM SIGIR, 2017. ACM, pp 315–324
8. Cho K, van Merrienboer B, Gülçehre Ç, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: EMNLP, pp 1724–1734
9. Cui Y, Che W, Liu T, Qin B, Yang Z, Wang S, Hu G (2019) Pre-training with whole word masking for chinese bert. arXiv preprint arXiv:190608101
10. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805
11. Dong L, Wei F, Tan C, Tang D, Zhou M, Xu K (2014) Adaptive recursive neural network for target-dependent twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of

the Association for Computational Linguistics, ACL 2014, June 22–27, 2014, Baltimore, MD, USA, Volume 2: Short Papers. The Association for Computer Linguistics, pp 49–54. https://doi.org/10.3115/v1/p14-2009
12. Felsenthal DS, Machover M (2001) The treaty of nice and qualified majority voting. Soc Choice Welf 18(3):431–464
13. Gonen H, Goldberg Y (2019) Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In: NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies—Proceedings of the Conference, pp 609–614
14. He M, Ge Y, Wu L, Chen E, Tan C (2016) Predicting the popularity of DanMu—enabled videos: a multi-factor view. Springer, Berlin
15. Khot T, Clark P, Guerquin M, Jansen P, Sabharwal A (2020) QASC: a dataset for question answering via sentence composition. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, pp 8082–8090
16. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: ICLR 2015
17. Kurita K, Vyas N, Pareek A, Black AW, Tsvetkov Y (2019) Measuring bias in contextualized word representations. CoRR abs/1906.07337. arxiv:1906.07337
18. Lei J, Zhang Q, Wang J, Luo H (2019) BERT based hierarchical sequence classification for context-aware microblog sentiment analysis. In: Gedeon T, Wong KW, Lee M (eds) ICONIP, Lecture Notes in Computer Science, vol 11955, pp 376–386
19. Li X, Bing L, Lam W, Shi B (2018) Transformation networks for target-oriented sentiment classification. arXiv preprint. arXiv:180501086
20. Lin C, Zhao S, Meng L, Chua T (2020) Multi-source domain adaptation for visual sentiment classification. In: AAAI. AAAI Press, pp 2661–2668
21. Lv G, Tong X, Chen E, Yi Z, Yi Z (2016) Reading the videos: temporal labeling for crowdsourced time-sync videos based on semantic embedding. In: AAAI, pp 3000–3006
22. Maas AL, Daly RE, Pham PT, Huang D, Potts C (2011) Learning word vectors for sentiment analysis. In: Meeting of the Association for Computational Linguistics: Human Language Technologies
23. Ma X, Cao N (2017) Video-based evanescent, anonymous, asynchronous social interaction: motivation and adaption to medium. In: ACM CSCW, pp 770–782
24. Ma S, Cui L, Dai D, Wei F, Sun X (2019) Livebot: generating live video comments based on visual and textual contexts. In: AAAI
25. Maudslay RH, Gonen H, Cotterell R, Teufel S (2019) It's all in the name: mitigating gender bias with name-based counterfactual data substitution. EMNLP
26. Maudslay RH, Gonen H, Cotterell R, Teufel S (2020) It's all in the name: mitigating gender bias with name-based counterfactual data substitution. EMNLP, pp 5267–5275
27. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2019) A survey on bias and fairness in machine learning. CoRR abs/1908.09635. arxiv:1908.09635
28. Qian X, Liu X, Ma X, Lu D, Xu C (2016) What is happening in the video? Annotate video by sentence. IEEE Trans Circuits Syst Video Technol 26(9):1746–1757
29. Ruan S, Zhang K, Wang Y, Tao H, He W, Lv G, Chen E (2020) Context-aware generation-based net for multi-label visual emotion recognition. In: ICME, pp 1–6
30. Saeidi M, Bouchard G, Liakata M, Riedel S (2016) Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In: Calzolari N, Matsumoto Y, Prasad R (eds) COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11–16, 2016, Osaka, Japan, ACL, pp 1546–1556. https://www.aclweb.org/anthology/C16-1146/
31. Song Y, Shi S, Li J, Zhang H (2018) Directional skip-gram: explicitly distinguishing left and right context for word embeddings. In: ACL, pp 175–180
32. Song Y, Wang J, Jiang T, Liu Z, Rao Y (2019) Attentional encoder network for targeted sentiment classification. arXiv preprint. arXiv:190209314
33. Su Y, Hu W, Jiang J, Su R (2020) A novel LMAEB-CNN model for Chinese microblog sentiment analysis. J Supercomput 76(11):9127–9141. https://doi.org/10.1007/s11227-020-03198-x
34. Tang D, Qin B, Feng X, Liu T (2015) Effective lstms for target-dependent sentiment classification. arXiv preprint. arXiv:151201100

35. Tang D, Qin B, Liu T (2016) Aspect level sentiment classification with deep memory network. arXiv preprint arXiv:160508900
36. Wang B, Yao T, Zhang Q, Xu J, Wang X (2020) Reco: A large scale Chinese reading comprehension dataset on opinion. In: AAAI, pp 9146–9153
37. Xian Y, Li J, Zhang C, Liao Z (2015) Video highlight shot extraction with time-sync comment. In: International Workshop on Hot Topics in Planet-Scale Mobile Computing and Online Social Networking, pp 31–36
38. Yang X, Binglu W, Junjie H, Shuwen L (2017b) Natural language processing in "bullet screen" application. In: ICSSSM. IEEE, pp 1–6
39. Yang W, Ruan N, Gao W, Wang K, Ran W, Jia W (2017a) Crowdsourced time-sync video tagging using semantic association graph. In: ICME, 2017. IEEE, pp 547–552
40. Yao Y, Bort J, Huang Y (2017) Understanding Danmaku's potential in online video learning. CHI 2017:3034–3040
41. Yu J, Jiang J (2019) Adapting bert for target-oriented multimodal sentiment classification. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, IJCAI, pp 5408–5414
42. Zeng B, Yang H, Xu R, Zhou W, Han X (2019) LCF: A Local context focus mechanism for aspect-based sentiment classification. Applied Sciences (Switzerland) 9(16), https://doi.org/10.3390/app9163389
43. Zhao Y, Peng X, Tang J, Song S (2017) Understanding young people's we-intention to contribute in Danmaku websites: motivational, social, and subculture influence. In: Conference 2017 Proceedings
44. Zhou J, Chen Q, Huang JX, Hu QV, He L (2020a) Position-aware hierarchical transfer model for aspect-level sentiment classification. Inf Sci 513:1–16
45. Zhou J, Huang JX, Hu QV, He L (2020b) Is position important? Deep multi-task learning for aspect-based sentiment analysis. Appl Intell 50(10):3367–3378
46. Zhou J, Huang JX, Hu QV, He L (2020d) SK-GCN: modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification. Knowl Based Syst 205:106292
47. Zhou J, Huang JX, Hu QV, He L (2020c) Modeling multi-aspect relationship with joint learning for aspect-level sentiment classification. In: DASFAA, pp 786–802
48. Zhou J, Tian J, Wang R, Wu Y, Xiao W, He L (2020e) Sentix: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In: Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8–13

Authors and Affiliations

Qingchun Bai[1] · Kai Wei[2] · Jie Zhou[1] · Chao Xiong[3] · Yuanbin Wu[1] · Xin Lin[1] · Liang He[1]

✉ Qingchun Bai
qc_bai@foxmail.com

1 Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, China

2 University of Pittsburgh, Pittsburgh, PA, USA

3 iQiYi. Inc., Beijing, China