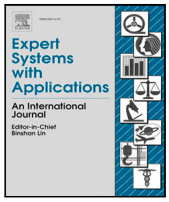
Expert Systems With Applications 213 (2023) 119194

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Generating effective label description for label-aware sentiment classification

Xiaofei Zhu [a,*], Zhanwang Peng [a], Jiafeng Guo [b], Stefan Dietze [c,d]

[a] College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China
[b] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[c] Knowledge Technologies for the Social Sciences, Leibniz Institute for the Social Sciences, Cologne 50667, Germany
[d] Institute of Computer Science, Heinrich-Heine-University Düsseldorf, Düsseldorf 40225, Germany

ARTICLE INFO

Keywords:
Sentiment classification
Text summarization
Attention network
Sentiment analysis

ABSTRACT

Sentiment classification aims to predict the sentiment label for a given text. Recently, several research efforts have been devoted to incorporate matching clues between text words and class labels into the learning process of text representation. However, these methods heavily rely on the availability of label content. Moreover, they simply capture the label-specific signals to measure each word's contribution by either implicitly employing a learnable label representation or explicitly leveraging the interaction between text words and labels via the interaction mechanism. To deal with these issues, in this paper, we propose a novel framework called Label-Guided Dual-view Sentiment Classifier (LGDSC). We first introduce a new strategy for generating an effective label description and then design a novel Dual-Channel Label-guided Attention Network (DLAN) to learn a text representation via two different channels. DLAN will be further leveraged to learn label-guided text representations from two different views. Extensive experimental results on four real-world datasets demonstrate that LGDSC consistently outperforms the state-of-the-art baseline methods.

1. Introduction

Sentiment classification (also known as opinion mining) (Cambria, 2016; Cambria, Li, Xing, Poria, & Kwok, 2020; Lin, Fu, Li, Cai, & Zhou, 2021; Zhu, Zhu, Guo, & Dietze, 2022; Zhu, Zhu, Guo, Liang, & Dietze, 2021) has emerged as a powerful strategy for understanding consumer opinion towards a product, which is of significant importance for organizations during their decision making process. In this context, sentiment classification aims to predict the sentiment label for a given review written by consumers. Conventional approaches typically learn text representation based on the input review text, and after that a fully-connected (FC) layer at the topmost of the network is used to make the final prediction. Tang, Qin, and Liu (2015) learn text representations with a hierarchical neural network. It first produces sentence representations with convolutional neural network or long short-term memory, then leverages the obtained sentence vectors to learn review-level representation. Ma, Sun, Lin, and Ren (2018) propose a hierarchical end-to-end model for joint learning of text summarization and sentiment classification in order to improve each other. Chan, Chen, and King (2020) predict the sentiment label from the review text representation as well as the summary representation with a consistent constraint between them. Ye, Dai, Dong, and Wang (2021) fuse the information contained in different features by proposing a novel multi-view ensemble learning method. Fei, Ren, Wu, Li, and Ji (2021) capture the latent target-opinion distribution behind the documents and incorporate the prior knowledge into the classification process.

One key limitation of the above-mentioned approaches is that they ignore the fine-grained classification signals (i.e., matching clues between text words and class description). More precisely, they merely treat the categories as indexes in the label vocabulary while lack of modeling category description to explicitly mention what to classify. Several recent efforts (Du et al., 2019; Huang, Chen, Xiao, & Jing, 2019; Xiao, Huang, Chen, & Jing, 2019) have attempted to leverage the fine-grained classification signals into the learning process of text representation. By explicitly modeling the fine-grained classification signals (i.e., the category description), they can force the model to attend to the most salient texts with respect to the label (Chai, Wu, Han, Wu, & Li, 2020). A motivating example is shown in Fig. 1. For example, Du et al. (2019) introduce the interaction mechanism to calculate the matching scores between text words and class labels. Since the class labels are not explicitly given and are modeled in an implicit manner, it would result in an inferior performance. To address this

* Corresponding author.
E-mail addresses: zxf@cqut.edu.cn (X. Zhu), 51190324101@2019.cqut.edu.cn (Z. Peng), guojiafeng@ict.ac.cn (J. Guo), stefan.dietze@gesis.org (S. Dietze).

https://doi.org/10.1016/j.eswa.2022.119194
Received 8 September 2021; Received in revised form 17 October 2022; Accepted 30 October 2022
Available online 8 November 2022

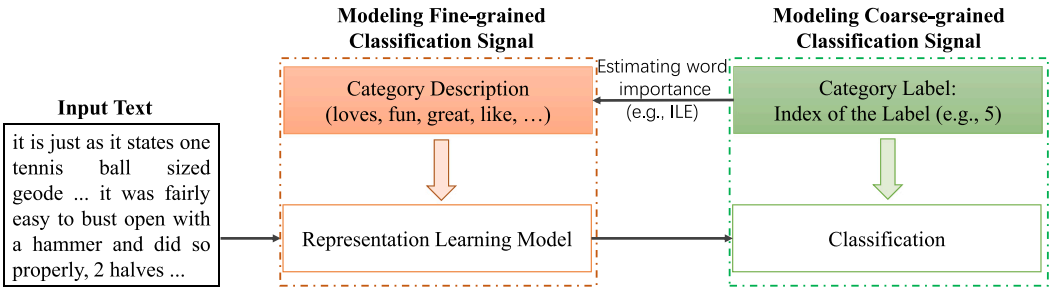


Fig. 1. A motivating example of leveraging label description for sentiment classification. We propose a novel strategy for generating label description by introducing a well-designed measurement, i.e., Inverse Label Entropy (ILE), for estimating the word importance score. The label description will serve as fine-grained classification signals for matching text words and class labels.

issue, there are several works attempting to incorporate label content in an explicit manner. Xiao et al. (2019) exploit the text content and label content to learn text representations. Specifically, they embed each label and explicitly capture the semantic relations between text words and labels. This work only focuses on learning the interaction between words and labels while ignoring the correlation among labels. Huang et al. (2019) establish an explicit label-aware representation for each text with a hybrid attention deep neural network. Different from Xiao et al. (2019), they further propose to model a label co-occurrence graph to learn a better label embedding which can maintain the label structure. This approach works well when there is a large-scale label set for exploiting the label structure. In the presence of a small number of class labels, which is typical for sentiment classification, the learned embeddings may not benefit from modeling the label structure.

Although the incorporation of the label description has been proven beneficial in many NLP tasks, a deficiency is that they heavily rely on the availability of label content and become inapplicable when there is no label content available. Very recently, some effort attempts to generate label description automatically, e.g., Chai et al. (2020) propose to generate a summary through a reinforcement learning module (Kumar, Ramakrishnan, & Li, 2019; Yuan et al., 2017) and then use the summary as label description to guide the model to attend to the most salient words. As the generated label description varies from text to text, this approach is still far from generating a satisfied description for each distinct sentiment label. Therefore, how to generate effective label description for text classification remains a challenging research question.

In this work, we propose a novel label-guided dual-view sentiment classifier (LGDSC) based on an automatic strategy for generating effective label description. In particular, we generate label description by introducing a novel discrimination capabilities-based word importance measurement, i.e., the inverse label entropy based word importance score. The label-guided dual-view sentiment classifier mainly consists of four components, i.e., a text encoder to learn a contextual representation, a summary decoder to generate a summary which will serve for the summary-view representation learning, a dual-channel label-guided attention network which is designed to learn a label-guided text representation from two different channels, as well as a dual-view sentiment classifier which aims to learn a text representation from both the source view and the summary view. After that, we leverage a two-layer feed-forward neural network to predict the sentiment label based on each learned representation. Fig. 2 illustrates the overall architecture of LGDSC.

We carry out extensive experiments on four widely used public review datasets, including the domains Sports, Toys, Home and Movies. The results show that our proposed approach LGDSC outperforms state-of-the-art sentiment classification baselines on all four datasets in terms of both macro F1 score and the balanced accuracy. We conduct further experiments to explore how label-guided attention influences the performance of sentiment classification. The main contributions of this work are summarized as follows:

- We propose a novel strategy for generating effective label descriptions by introducing a well-designed measurement into the estimation process of the word importance score.
- We design a novel Dual-Channel Label-guided Attention Network to learn text representation via two different channels.
- Extensive experiments are conducted on four widely used datasets (i.e., Sports, Toys, Home, and Movies), and experimental results demonstrate that our proposed approach shows superior performance compared with state-of-the-art baseline methods.

The rest of the paper is organized as follows. In Section 2, we give a brief review of the related work. Section 3 describes the generation of label descriptions. We introduce our proposed label-guided dual-view sentiment classifier in Section 4 and discuss the experimental results of our empirical studies in Section 5. In Section 6, we conclude the paper.

2. Related work

In this section, we briefly review related works including *sentiment classification* and *label-indicative document classification*.

Sentiment Classification. Sentiment classification is an important task for understanding customer needs and has been widely studied in recent years. Existing approaches usually fall into two categories: single classification models and joint classification and summarization models. Single classification models are usually regarded as a text classification task (Pang & Lee, 2005). Early research works focus on extracting effective features and then apply supervised machine learning methods (Pang & Lee, 2005; Pang, Lee, & Vaithyanathan, 2002) to conduct sentiment classification. Pang et al. (2002) employ machine learning methods (i.e., SVM, Maximum Entropy, and Naive Bayes) to perform sentiment classification based on textual features extracted by standard natural language processing techniques. Since the textual features would be not reliable enough to estimate the sentiment polarity, Gao, Yoshinaga, Kaji, and Kitsuregawa (2013) further exploit user leniency and product popularity to improve sentiment classification. Along with the success of deep learning in many applications, deep learning has also been explored for sentiment classification recently (Zhang, Wang, & Liu, 2018). Tang et al. (2014) incorporate the supervision from sentiment polarity of text to learn sentiment-specific word embedding for sentiment classification. Tang et al. (2015) propose a hierarchical framework for document-level sentiment classification. Specifically, they first learn sentence representation with convolutional neural network or long short-term memory. Then, they adaptively encode the semantics of sentences and their relations with the gated recurrent neural network.

Joint classification and summarization models are proposed to improve the capability of sentiment classification in recent works. It explores to incorporate a review summarization component to jointly improve the performance of review summarization and sentiment classification. Ma et al. (2018) propose an end-to-end framework to improve both text summarization and sentiment classification. They first

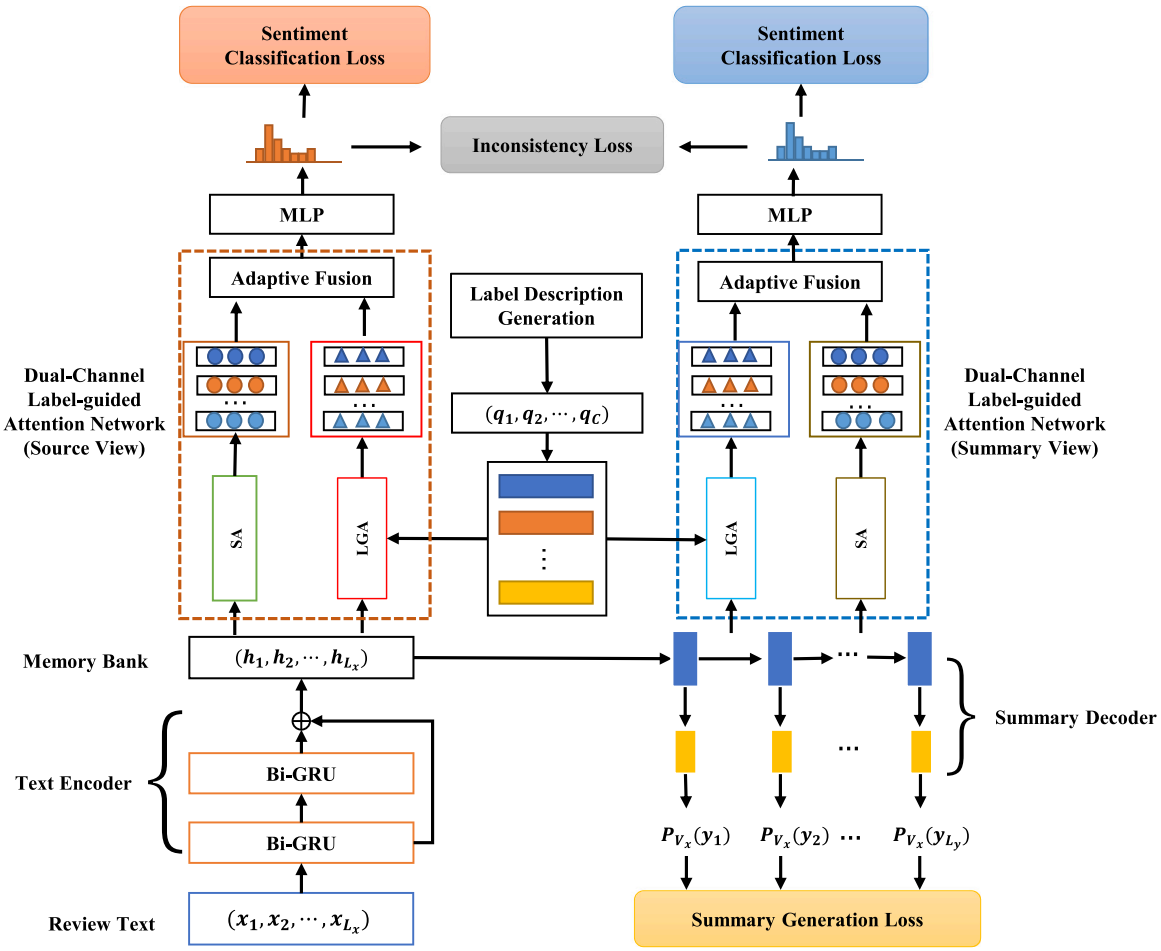


Fig. 2. The overall architecture of the proposed model LGDSC. It mainly consists of four components: (1) text encoder, which converts the input text into a memory bank; (2) summary decoder, which generates the summary based on the memory bank output by the text encoder; (3) dual-channel label-guided attention network (DLAN), which learns label-guided text representations via two different channels; (4) dual-view sentiment classifier, which learns a dual-view representation by applying the DLAN module on both source-view (input review text) and summary-view (generated summary by the summary decoder).

leverage a summarization layer to compress the input text into short sentences and then employ a sentiment classification layer to further "summarize" the text into a sentiment class. To effectively utilize the shared sentiment information in both sentiment classification task and review summarization task, Chan et al. (2020) further propose a dual-view approach which jointly improves the performance of these two tasks. Our work falls into the second category where we employ a joint framework for sentiment classification. Different from existing state-of-the-art works, we introduce label content information into the learning process of text representation in order to obtain a label-indicative text presentation. In addition, we design a novel dual-channel label-guided attention network to learn a text representation via two different channels.

Label-Indicative Text Classification. Conventional deep learning based text classification approaches usually follow a encoding-based framework, and the probability of a text belonging to a class is mainly determined by their overall matching score regardless of word-level matching signals (Du et al., 2019). Recently, some research works have attempted to incorporate the label content into the learning process of text representation to force the model attend to the most salient texts with respect to the class label. Du et al. (2019) employ the interaction mechanism (Wang & Jiang, 2016) to incorporate word-level matching clues into the process of text classification. The main limitation of their work is that they ignore the class context and model the class in an implicit manner, i.e., they project classes into real-valued latent representations. Xiao et al. (2019) propose to learn text representation by explicitly exploiting the text content and label content. More precisely, they incorporate the label description to embed each label into a vector like embedding, and explicitly calculate the semantic relations between the input text and the labels. They further design an adaptive fusion strategy to extract a proper semantic information to construct label-specific document representation. Huang et al. (2019) attempt to address the extreme multi-label text classification task by modeling the label structure among extreme labels. The label embedding is then determined by exploring a label co-exist graph. The main limitation of previous research efforts is that they heavily rely on the availability of label content and become inapplicable when there is no label content available. Our work can generate effective label description by introducing a novel discrimination capabilities-based word importance measurements, i.e., Inverse Label Entropy (ILE) based word importance score.

3. Label description generation

In this section, we explore how to generate an effective description for each sentiment label in order to learn a label-indicative text representation. Specifically, we first calculate the relevance of each word with respect to a sentiment label $c \in \{c_1, c_2, \ldots, c_C\}$ where C denotes the number of distinct sentiment labels, and then select the most effective words as the description of c. Intuitively, an effective label description should satisfy two requirements:

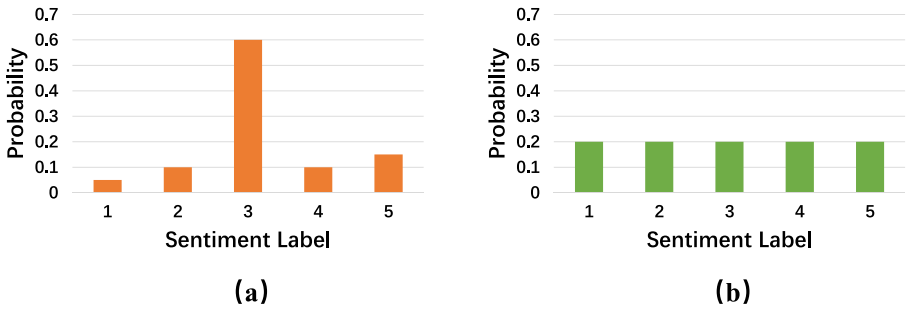


Fig. 3. Two words' corresponding sentiment label frequency distribution (e.g., 5 labels in this case). Both words have same Inverse Label Frequency (ILF) score while their corresponding sentiment label frequency distributions are significantly different. (a) a word with high skewness of sentiment label frequency distribution and (b) a word with high balance of sentiment label frequency distribution.

Requirement 1 (*Relevance*). *All selected words in the description of a label should be semantically relevant to the label, i.e., each word should have a strong relevance score to the label.*

Requirement 2 (*Discrimination*). *Each selected word should have a high discriminative capability among different labels, i.e., the words in the description of a label should be strongly correlated to that label and less correlated to other labels.*

To address the first requirement *Relevance*, we measure the relevance score of a word for a given label. Formally, given the text corpus D, a sentiment label c, a word w, and a review text $d \in D$, we measure the word relevance score of w with respect to c as follows:

$$r'_{w,c} = \sum_{d \in D^c} s_{w,d} \tag{1}$$

where D^c denotes all texts in D with the sentiment label c, $s_{w,d}$ is the importance score of word w in d.

It is worth noting that there are several ways to calculate the word importance score $s_{w,d}$, and in this work we leverage TFIDF (Ramos et al., 2003) to compute $s_{w,d}$ since it has achieved promising performance in many applications (Chen, 2017; Qin, Xu, & Guo, 2016). The TFIDF value of a word w in a review text d is defined as follows:

$$\begin{aligned} s_{w,d} &= TFIDF(w,d) \\ &= f_{w,d} \times log(|D|/f_{w,D}) \end{aligned} \tag{2}$$

where $f_{w,d}$ denotes the number of times w appears in d, $|D|$ is the size of the corpus D, and $f_{w,D}$ is the number of review texts in which w appears in the corpus D.

To deal with the second requirement *Discrimination*, we need to evaluate the discrimination of a word with respect to all sentiment labels. It is worth noting that we can design a measurement, i.e., Inverse Label Frequency (ILF), which measures the discrimination capability of a word w based on a similar strategy of the Inverse Document Frequency (IDF). For example, we have:

$$LF(w) = |\cup_{d \in D} \{l_d | w \in d\}| \tag{3}$$

where l_d is the corresponding sentiment label of d. After that, we have the ILF-based importance score of a word w with respect a sentiment label c as follows:

$$\begin{aligned} r^{ILF}_{w,c} &= \frac{r'_{w,c}}{LF(w)} \\ &= \frac{\sum_{d \in D^c} f_{w,d} \times log(|D|/f_{w,D})}{|\cup_{d \in D}\{l_d | w \in d\}|} \end{aligned} \tag{4}$$

However, one major limitation of the above ILF model is that it does not take into account the skewness of sentiment label frequency distribution of a word. For example, in Fig. 3, there are two words and both of them appear in all sentiment labels (e.g., 5 labels in this case). According to the definition of the inverse label frequency, they will have the same ILF score. However, a word with high skewness of sentiment label frequency distribution usually demonstrates a stronger discriminative capability as compared to a word with high balance of sentiment label frequency distribution. Based on this observation, we propose to introduce the information entropy (Núñez, Cincotta, & Wachlin, 1996) to measure the skewness of sentiment label frequency of a word, and define the label entropy as follows:

$$LE(w) = -\sum_{c=1}^{C} p_c(w) \log(p_c(w)), \tag{5}$$

where C is the number of classes and $p_c(w)$ is the probability that w appears in a text with label c which is defined as:

$$p_c(w) = \frac{|\cup_{d \in D^c} \{d | w \in d\}|}{|D^c|} \tag{6}$$

It should be noted that a higher value of $LE(w)$ indicates that the word w has smaller discriminative power among different labels. As each word in the sentiment label description should have a high discrimination power, we need to penalize words which have small discriminative capacity. To the end, we introduce the inverse label entropy into the process of estimating the importance score of each word where

$$ILE(w) = \frac{1}{LE(w)} \tag{7}$$

Therefore, we have the inverse label entropy based importance score of a word w with respect a label c as follows:

$$\begin{aligned} r^{ILE}_{w,c} &= r'_{w,c} \cdot ILE(w) \\ &= \frac{\sum_{d \in D^c} f_{w,d} \cdot log(|D|/f_{w,D})}{-\sum_{c=1}^{C} p_c(w) \cdot \log(p_c(w))}, \end{aligned} \tag{8}$$

where $f_{w,d}$ is the number of times w appears in d, $f_{w,D}$ is the number of review texts in which w appears in the corpus D, and $|D|$ is the number of review texts in the corpus D.

For a sentiment label c, we measure the relevance score $r^{ILE}_{w,c}$ for each distinct word w which appears in a text with label c, and then select the top-K words $(w^c_1, w^c_2, \dots, w^c_K)$ with the highest scores as the description of the sentiment label c.

Finally, we obtain the representation of a sentiment label c as follows:

$$\mathbf{q}_c = \frac{1}{K} \sum_{i=1}^{K} emd\left(w^c_i\right), \tag{9}$$

where $emd(w^c_i)$ is an operation that maps word w^c_i to a low-dimensional embedding via a lookup table. Similarly, we can get representations of all sentiment labels: $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_C) \in \mathbb{R}^{C \times d_e}$ where d_e is the dimension of the label embedding.

4. Label-guided dual-view sentiment classifier

In this section, we introduce our proposed model Label-guided Dual-view Sentiment Classifier (LGDSC), which mainly consists of four components, including a text encoder, a summary decoder, a dual-channel label-guided attention network, and a dual-view sentiment classifier.

4.1. Text encoder

The text encoder aims to encode the input text sequence $d = (w_1, \ldots, w_{L_x})$ with L_x words into a contextual representation $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_{L_x})$, which forms the memory bank for other components. Specifically, for the input text sequence d, we first use a lookup table $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d_e}$ to convert each word w_i to a word embedding vector $\mathbf{x}_i \in \mathbb{R}^{d_e}$, where $|\mathcal{V}|$ is the size of the vocabulary and d_e denotes the embedding size. Then we employ two consecutive Bi-directional Gated-Recurrent Unit (BiGRU) with a residual connection. Formally, a BiGRU is used to convert $\mathbf{x}_i$ to a representation $\mathbf{u}_i \in \mathbb{R}^{d_u}$ as follows:

$$\vec{\mathbf{u}}_i = GRU^{(1)}(\mathbf{x}_i, \vec{\mathbf{u}}_{i-1}) \tag{10}$$

$$\overleftarrow{\mathbf{u}}_i = GRU^{(1)}(\mathbf{x}_i, \overleftarrow{\mathbf{u}}_{i+1}), \tag{11}$$

where $\mathbf{u}_i = [\vec{\mathbf{u}}_i; \overleftarrow{\mathbf{u}}_i]$. $\vec{\mathbf{u}}_i \in \mathbb{R}^{d_u/2}$ and $\overleftarrow{\mathbf{u}}_i \in \mathbb{R}^{d_u/2}$ are the forward and backward representations, respectively. $[;]$ denotes the concatenation operation. After that, another BiGRU is leveraged to convert $\mathbf{u}_i$ to $\tilde{\mathbf{h}}_i \in \mathbb{R}^{d_u}$:

$$\vec{\tilde{\mathbf{h}}}_i = GRU^{(2)}(\mathbf{u}_i, \vec{\tilde{\mathbf{h}}}_{i-1}) \tag{12}$$

$$\overleftarrow{\tilde{\mathbf{h}}}_i = GRU^{(2)}(\mathbf{u}_i, \overleftarrow{\tilde{\mathbf{h}}}_{i+1}), \tag{13}$$

where $\tilde{\mathbf{h}}_i = [\vec{\tilde{\mathbf{h}}}_i; \overleftarrow{\tilde{\mathbf{h}}}_i]$. In order to alleviate the gradient vanishing issue (He, Zhang, Ren, & Sun, 2016), a residual connection is also incorporated to fuse the outputs of the two consecutive BiGRUs, i.e., $\mathbf{u}_i$ and $\tilde{\mathbf{h}}_i$, as follows:

$$\mathbf{h}_i = \delta \tilde{\mathbf{h}}_i + (1 - \delta)\mathbf{u}_i, \tag{14}$$

where $\delta \in [0, 1]$ is a hyperparameter. $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_{L_x}) \in \mathbb{R}^{L_x \times d_u}$ is the output of the text encoder which will be used as the memory bank.

4.2. Summary decoder

The summary decoder uses a undirectional GRU to generate an output summary $y = (y_1, y_2, \ldots, y_{L_y})$ step by step. We employ the decoder of the pointer generator network (See, Liu, & Manning, 2017) as our summary decoder. Specifically, on each step t, an unidirectional GRU receives the word embedding of the previous prediction $\mathbf{y}_{t-1} \in \mathbb{R}^{d_e}$ and the previous decoder hidden state $\mathbf{s}_{t-1} \in \mathbb{R}^{d_u}$ and produces the current decoder state $\mathbf{s}_t$:

$$\mathbf{s}_t = GRU^{(3)}\left(\mathbf{y}_{t-1}, \mathbf{s}_{t-1}\right), \tag{15}$$

and note that $\mathbf{y}_0$ is the embedding of the start token. In order to aggregate the most important information from the review text, we introduce the attention mechanism to calculate an attention score $a_{t,i}$:

$$\alpha_{t,i} = \mathbf{v}^T \tanh\left(\mathbf{W}_h \mathbf{h}_i + \mathbf{W}_s \mathbf{s}_t + \mathbf{b}_{attn}\right) \tag{16}$$

$$a_{t,i} = \frac{\exp\left(\alpha_{t,i}\right)}{\sum_{j=1}^{L_x} \exp\left(\alpha_{t,j}\right)}, \tag{17}$$

where $\mathbf{W}_h \in \mathbb{R}^{d_u \times d_u}$, $\mathbf{v} \in \mathbb{R}^{d_u}$, $\mathbf{W}_s \in \mathbb{R}^{d_u \times d_u}$ and $\mathbf{b}_{attn} \in \mathbb{R}^{d_u}$ are model parameters. Next, the attention score is used to calculate a weighted sum of the memory bank $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_{L_x})$ and produce a vector $\mathbf{h}_t^* = \sum_{i=1}^{L_x} a_{t,i}\mathbf{h}_i$, which serves as the representation of the input sequence d at step t.

Then we use the decoder hidden state $\mathbf{s}_t \in \mathbb{R}^{d_u}$ and $\mathbf{h}_t^* \in \mathbb{R}^{d_u}$ to calculate the probability distribution over the words in the predefined vocabulary $\mathcal{V}$:

$$P_{\mathcal{V}}(y_t|y_{1:t-1}, d) = softmax(\mathbf{W}_{v_2}(\mathbf{W}_{v_1}[\mathbf{s}_t; \mathbf{h}_t^*] + \mathbf{b}_{v_1}) + \mathbf{b}_{v_2}), \tag{18}$$

where $\mathbf{W}_{v_1} \in \mathbb{R}^{d_u \times 2d_u}$, $\mathbf{W}_{v_2} \in \mathbb{R}^{|\mathcal{V}| \times d_u}$, $\mathbf{b}_{v_1} \in \mathbb{R}^{d_u}$ and $\mathbf{b}_{v_2} \in \mathbb{R}^{|\mathcal{V}|}$ are trainable parameters, $y_{1:t-1}$ denotes the partial sequence of previous generated words (i.e., $(y_1, \ldots, y_{t-1})$). In order to solve the problem that the decoder cannot generate out-of-vocabulary (OOV) words, we incorporate the copy mechanism of See et al. (2017) to predict OOV words by copying words from the input text. Specifically, we introduce a generating-copying switch $p_{gen} \in [0, 1]$ between generating a word from the predefined vocabulary $\mathcal{V}$ according to $P_{\mathcal{V}}$ and copying a word from the source text d according to the attention distribution. We leverage the encoder–decoder attention weight a_t as the copy distribution (See et al., 2017) which determines where to attention in step t. The final probability distribution of the ground-truth target word y_t is:

$$P(y_t) = p_{gen}P_{\mathcal{V}}(y_t) + (1 - p_{gen})P_{copy}(y_t) \tag{19}$$

$$p_{gen} = sigmoid(\mathbf{v}_g^T[\mathbf{h}_t^*; \mathbf{s}_t; \mathbf{y}_{t-1}] + b_{gen}) \tag{20}$$

$$P_{copy}(y_t) = \sum_{i: w_i = y_t} a_{t,i}, \tag{21}$$

where we use $P_{\mathcal{V}}(y_t)$ to denote $P_{\mathcal{V}}(y_t|y_{1:t-1}, d)$, $\mathbf{v}_g \in \mathbb{R}^{2d_u + d_e}$ and $b_{gen} \in \mathbb{R}$ are trainable parameters.

The loss function of the summarization $\mathcal{L}_g$ is defined as the negative log-likelihood of the ground-truth target word y_t for each step t, which is formulated as follows:

$$\mathcal{L}_g = -\sum_{t=1}^{L_{y^*}} log P(y_t^*), \tag{22}$$

where L_{y^*} denotes the number of words in the ground-truth summary y^*.

4.3. Dual-channel label-guided attention network (DLAN)

The DLAN module is designed to learn a label-guided text representation, consists of three sub-modules, i.e., *Label-Guided Attention*, *Self-Attention*, and *Adaptive Fusion*. Label-guided attention aims to explicitly introduce the label description information into the words' contribution estimation process. In contrast, self-attention attempts to implicitly employ a learnable label representation to measure each word's contribution based on their representations. Adaptive fusion is utilized to fuse the outputted representations from both label-guided attention and self-attention. It is worth noting that the DLAN is leveraged to learn label-guided text representations from two different views, i.e., the input text view (source view) and the generated summary-view (summary view). For simplicity, here we only focus on discuss DLAN based on the source view, and the counterpart based on the summary view will share a similar process.

4.3.1. Label-guided attention (LGA)

The LGA module explicitly introduces the sentiment label representation $\mathbf{Q} \in \mathbb{R}^{C \times d_e}$ into the text representation learning process. It first computes a label-guided attention matrix $\mathbf{A}^{(C)} \in \mathbb{R}^{C \times L_x}$ between the text sequence $\mathbf{H} \in \mathbb{R}^{L_x \times d_u}$ and the label representation $\mathbf{Q} \in \mathbb{R}^{C \times d_e}$, which is formulated as follows:

$$\mathbf{A}^{(C)} = \tilde{\mathbf{Q}} \cdot \mathbf{H}^T \tag{23}$$

$$\tilde{\mathbf{Q}} = ReLU(\mathbf{Q}\mathbf{W}_q), \tag{24}$$

where $\mathbf{W}_q \in \mathbb{R}^{d_e \times d_u}$ is trainable weight. The jth row of $\mathbf{A}^{(C)}$ (i.e., $\mathbf{A}_{j\cdot}^{(C)} \in \mathbb{R}^{L_x}$) indicates the attention scores of the jth sentiment label pay attention to all words in the input text sequence.

After that, we can obtain the explicit label-guided text representation $\mathbf{M}^{(C)} \in \mathbb{R}^{C \times d_u}$ as follows:

$$\mathbf{M}^{(C)} = \mathbf{A}^{(C)} \cdot \mathbf{H} \tag{25}$$

It is worth noting that $\mathbf{M}^{(C)}$ can be considered as the text representation along all sentiment labels.

4.3.2. Self-attention (SA)

The above-mentioned LGA module learns a label-guided text representation $\mathbf{M}^{(C)}$ in an explicitly way, i.e., directly introducing the label description information into the text representation learning process. Inspired by Xiao et al. (2019), we also introduce an implicit way to capture the contribution of all words to each sentiment label, and obtain the corresponding label-guided text representation $\mathbf{M}^{(T)} \in \mathbb{R}^{C\times d_u}$. Different from LGA, we compute the implicit label-guided attention matrix $\mathbf{A}^{(T)} \in \mathbb{R}^{C\times L_x}$ only based on text sequence $\mathbf{H} \in \mathbb{R}^{L_x\times d_u}$, which is formulated as follows:

$$\mathbf{A}^{(T)} = softmax(\mathbf{W}_{a_2} \cdot tanh(\mathbf{W}_{a_1} \cdot \mathbf{H}^T)), \tag{26}$$

where $\mathbf{W}_{a_1} \in \mathbb{R}^{d_u\times d_u}$, $\mathbf{W}_{a_2} \in \mathbb{R}^{C\times d_u}$ are trainable parameters. Similarly, the jth row of $\mathbf{A}^{(T)}$ (i.e., $\mathbf{A}^{(T)}_{j\cdot} \in \mathbb{R}^{L_x}$) can also be considered as the attention scores of the jth sentiment label assigning to all words in the input text sequence. Therefore, we can obtain the implicit label-guided text representation $\mathbf{M}^{(T)} \in \mathbb{R}^{C\times d_u}$ as follows:

$$\mathbf{M}^{(T)} = \mathbf{A}^{(T)} \cdot \mathbf{H} \tag{27}$$

4.3.3. Adaptive fusion

After we obtain both $\mathbf{M}^{(C)}$ and $\mathbf{M}^{(T)}$, the gating mechanism is leveraged to adaptively fuse them. In particular, we introduce two fuse weights (i.e., $u^{(C)}$ and $u^{(T)}$) to determine the importance of $\mathbf{M}^{(C)}$ and $\mathbf{M}^{(T)}$, respectively. The weights can be obtained as follows:

$$u^{(C)} = \sigma(\mathbf{M}^{(C)} \cdot \mathbf{v}_c) \tag{28}$$

$$u^{(T)} = \sigma(\mathbf{M}^{(T)} \cdot \mathbf{v}_T), \tag{29}$$

where σ is the sigmoid function, $\mathbf{v}_c \in \mathbb{R}^{d_u}$ and $\mathbf{v}_T \in \mathbb{R}^{d_u}$ are trainable parameters. $u^{(C)}_i \in \mathbb{R}$ and $u^{(T)}_i \in \mathbb{R}$ indicate the importance of the explicit label-guided text representation $\mathbf{M}^{(C)}$ and the implicit label-guided text representation $\mathbf{M}^{(T)}$ along the ith sentiment label, respectively. We further normalize $u^{(C)}_i$ and $u^{(T)}_i$ as follows:

$$u^{(C)}_i = \frac{u^{(C)}_i}{u^{(C)}_i + u^{(T)}_i} \tag{30}$$

$$u^{(T)}_i = 1 - u^{(C)}_i \tag{31}$$

Then we obtain the adaptively fused label-guided text representation $\tilde{\mathbf{M}}_{i\cdot} \in \mathbb{R}^{1\times d_u}$ along the ith label:

$$\tilde{\mathbf{M}}_{i\cdot} = u^{(C)}_i \mathbf{M}^{(C)}_{i\cdot} + u^{(T)}_i \mathbf{M}^{(T)}_{i\cdot} \tag{32}$$

At last, the fused label-guided text representation along all labels is $\tilde{\mathbf{M}} = (\tilde{\mathbf{M}}_{i\cdot}, \dots, \tilde{\mathbf{M}}_{C\cdot}) \in \mathbb{R}^{C\times d_u}$, and we use the average-pooling mechanism to get the final source-view representation $\mathbf{m}_s \in \mathbb{R}^{d_u}$:

$$\mathbf{m}_s = AveragePooling(\tilde{\mathbf{M}}) \tag{33}$$

Similarly, we can obtain the summary-view representation $\mathbf{m}_t \in \mathbb{R}^{d_u}$ in a similar way by applying DLAN on top of the hidden states $\mathbf{S} = (\mathbf{s}_1, \dots, \mathbf{s}_{L_y}) \in \mathbb{R}^{L_y\times d_u}$ of the summary decoder.

4.4. Dual-view sentiment classifier

After we get the text representations from both source-view and summary-view, we employ a two-layer feedforward neural network with ReLU as the activation function (Chan et al., 2020; Ma et al., 2018) to produce the probability distribution over the sentiment label on each representation, respectively.

Here we first formulate the classification process from the source-view representation $\mathbf{m}_s$. In particular, a softmax function is leveraged to produce the probability distribution on all sentiment labels. Formally, we have:

$$P_s(c|\mathbf{m}_s) = softmax(\mathbf{W}_{s_2}(ReLU(\mathbf{W}_{s_1}\mathbf{m}_s + \mathbf{b}_{s_1}) + \mathbf{b}_{s_2})), \tag{34}$$

where $\mathbf{W}_{s_1} \in \mathbb{R}^{d_u\times d_u}$, $\mathbf{W}_{s_2} \in \mathbb{R}^{C\times d_u}$, $\mathbf{b}_{s_1} \in \mathbb{R}^{d_u}$, $\mathbf{b}_{s_2} \in \mathbb{R}^{C}$ are trainable parameters. The sentiment label with the highest probability will be used as the predicted sentiment label of the source view. We use the negative log-likelihood as the classification loss function:

$$\mathcal{L}_s = -log(P_s(c^*|d)), \tag{35}$$

where c^* is the ground-truth sentiment label. Similarly, we have the probability distribution on all sentiment labels based on the summary-view representation $\mathbf{m}_t$, which are formulated as follows:

$$P_t(c|\mathbf{m}_t) = softmax(\mathbf{W}_{t_2}(ReLU(\mathbf{W}_{t_1}\mathbf{m}_t + \mathbf{b}_{t_1}) + \mathbf{b}_{t_2})), \tag{36}$$

where $\mathbf{W}_{t_1} \in \mathbb{R}^{d_u\times d_u}$, $\mathbf{W}_{t_2} \in \mathbb{R}^{C\times d_u}$, $\mathbf{b}_{t_1} \in \mathbb{R}^{d_u}$, $\mathbf{b}_{t_2} \in \mathbb{R}^{C}$ are trainable parameters. And its corresponding classification loss is:

$$\mathcal{L}_t = -log(P_t(c^*|d, y)) \tag{37}$$

4.4.1. Inconsistency loss function

The sentiment attitudes of the source-view and the summary-view should be consistent, therefore we introduce an inconsistency loss function to maintain a consistent between their predicted sentiment labels. Here, we define the inconsistent loss function as the Kullback–Leibler (KL) divergence between P_s and P_t:

$$\begin{aligned}\mathcal{L}_c &= KL(P_s \parallel P_t) \\ &= \sum_{c=1}^{C} P_s(c|d) \log \frac{P_s(c|d)}{P_t(c|d, y)}\end{aligned} \tag{38}$$

It is worth noting that the inconsistency loss will force the two classifiers learn from each other to improve the performance of sentiment classification. In addition, since the summary-view sentiment classifier adopts the decoder's hidden states to predict the sentiment label of the generated summary, the inconsistency loss will also encourage the sentiment information in the decoder states to be close to the sentiment information in the encoder memory bank.

4.5. Objective function

The overall loss consists of four parts, i.e., the summarization loss $\mathcal{L}_g$, the source-view sentiment classification loss $\mathcal{L}_s$, the summary-view sentiment classification loss $\mathcal{L}_t$, and the inconsistent loss $\mathcal{L}_c$. We jointly optimize the four losses as follows:

$$\mathcal{L} = \beta_g\mathcal{L}_g + \beta_s\mathcal{L}_s + \beta_t\mathcal{L}_t + \beta_c\mathcal{L}_c, \tag{39}$$

where β_g, β_s, β_t and β_c are hyper-parameters to balance the weights of the four losses.

5. Experiments

In this section, we first introduce the datasets, baseline methods, and the evaluation metrics used in our experiments. Then we compare the proposed approach LGDSC with seven competitive baseline methods on all datasets.

5.1. Datasets

To conduct a fair and comprehensive evaluation, we adopt four widely used datasets from different domains in our experiments. These datasets are collected from the Amazon 5-core review repository (Chan et al., 2020; McAuley, Targett, Shi, & van den Hengel, 2015), and product reviews from four domains including Sports & Outdoors, Home & Kitchen, Movies & TV, and Toys & Games, are leveraged.

Each data sample includes a review text, a summary, and a label. It is worth noting that we treat the rating, which is an integer in the range of [1,5], as the sentiment label. All datasets are preprocessed by lowercasing all letters and tokenizing the text using Stanford CoreNLP (Manning, Surdeanu, Bauer, Finkel, Bethard, & McClosky, 2014). If a summary sentence is not ended properly, a period

Table 1
Statistics of the datasets. "Max.RL", "Min.RL" and "Ave.RL" respectively represent the maximum length, minimum length and average length of review in the training set. "Max.SL", "Min.SL" and "Ave.SL" respectively represent the maximum length, minimum length and average length of summary in the training set. "Lk" means the ratio of the data samples with sentiment kth label in the training set.

Dataset		Number	Max.RL	Min.RL	Ave.RL	Max.SL	Min.SL	Ave.SL	L1	L2	L3	L4	L5
Sports	Train	183,714	800	11	108.3	74	4	6.7	3.30%	3.90%	9.10%	22.90%	60.80%
	Valid	9000	769	11	106	30	4	6.8	2.80%	3.70%	9.40%	23.00%	61.10%
	Test	9000	796	11	108	53	4	6.7	3.20%	3.90%	9.30%	23.50%	60.10%
Toys	Train	104,296	800	16	125.9	55	4	6.8	2.90%	4.10%	11%	24.00%	58.00%
	Valid	8000	800	16	126.3	30	4	6.9	3.10%	4.30%	10.20%	23.70%	58.70%
	Test	8000	790	17	124.6	29	4	6.8	2.90%	4.10%	10.60%	24.00%	58.40%
Home	Train	367,395	800	16	120.9	64	4	6.8	5.30%	4.90%	9.10%	20.10%	60.60%
	Valid	10,000	800	16	120.6	30	4	6.7	5.70%	4.70%	9.10%	20.50%	60.00%
	Test	10,000	795	16	120.3	34	4	6.8	6.30%	6.30%	12.60%	23.30%	51.50%
Movies	Train	1,200,601	800	16	183.1	53	4	6.6	5.30%	4.90%	9.10%	20.10%	60.60%
	Valid	20,000	800	16	184.3	35	4	7.3	6.20%	6.50%	13.00%	23.20%	51.10%
	Test	20,000	795	16	184.3	32	4	7.3	6.20%	6.70%	12.80%	23.30%	51.00%

will be appended. In order to reduce the noise in these datasets, data samples with review length less than 16 or longer than 800, or the summary length is less than 4, will be discarded. At last, each dataset is split randomly into training, validation, and testing sets. Table 1 shows the statistics of the datasets.

5.2. Baselines

We compare our proposed approach LGDSC with seven methods which can be roughly categorized into two groups, including single sentiment model (i.e., BiGRU+Attention, DARLM) and joint sentiment and summarization model (i.e., HSSC, MAX, HSSC+copy, Max+copy, Dual-view):

Single sentiment model only considers the input review text for sentiment classification, which includes:

- BiGRU-Attention: This method first leverages a bi-directional GRU layer (Cho et al., 2014) to encode the input review into a hidden state. Then it incorporates the attention mechanism (Bahdanau, Cho, & Bengio, 2015) with glimpse operation (Vinyals, Bengio, & Kudlur, 2016) to aggregate information from the hidden state to generate a vector, which will be further fed into a two-layer feedforward neural network to predict the sentiment label.
- DARLM (Zhou, Wang, & Dong, 2018): This is the state-of-the-art model for sentence classification. It attempts to alleviate the attention bias problem on sentence classification. The model has two branches of attention subnets and an example discriminator. The two branches are jointly trained where one branch tries its best to classify all sentences and the other is enabled for sentences that cannot be handled well by the former. An example discriminator is designed to select the suitable attention.

Joint sentiment and summarization model simultaneously models both review text and summary for handling the task of sentiment classification, which includes:

- MAX (Ma et al., 2018): This method first encodes the input review with a bi-directional GRU layer, and the output of the encoder will be shared by a summary decoder and a sentiment classifier. The sentiment classifier utilizes a max pooling to aggregate the hidden state of the encoder into a vector. At last, a two-layer feedforward neural network predicts the sentiment label based on the vector.
- Max-copy: It is another strong baseline, which is a variant of MAX (Ma et al., 2018) via expanding the MAX model with the copy mechanism.
- HSSC (Ma et al., 2018): It explores a hierarchical end-to-end model, which consists of a summarization layer and a sentiment classification layer for improving both text summarization and sentiment classification.
- HSSC-copy: It is a strong baseline. This baseline is a variant of HSSC (Ma et al., 2018), where the copy mechanism (See et al., 2017) is incorporated into HSSC model.
- Dual-view (Chan et al., 2020): This method aims to effectively leverage the sentiment information in the review and the summary, and proposes a novel dual-view model for jointly improving the performance of review summarization and sentiment classification. It encourages the sentiment information in the decoder states to be close to that in the review context representation, and the sentiment classifiers from two distinct views can learn from each other in order to improve the performance of the sentiment classification.

5.3. Evaluation metrics

From Table 1, we can observe that the class distribution of the sentiment labels is imbalanced. Therefore, we employ the macro-averaged F1 score (Peng et al., 2018) and the balanced accuracy (Brodersen, Ong, Stephan, & Buhmann, 2010) as the evaluation metrics. We denote the macro-averaged F1 score and the balanced accuracy as "M.F1" and "B.Acc", respectively.

- Macro-averaged F1 Score (M.F1): In this work, we use the macro-averaged F1 score which evaluates averaged F1 score of all distinct sentiment labels. It gives equal weight to each label. Formally, the macro-averaged F1 score is defined as:

$$M.F1 = \frac{1}{C}\sum_{c=1}^{C}\frac{2P_cR_c}{P_c + R_c}, \tag{40}$$

where $P_c = \frac{TP_c}{TP_c+FP_c}$, $R_c = \frac{TP_c}{TP_c+FN_c}$ and TP_c, FP_c, FN_c denote the true-positives, false-positives, and false-negatives for the cth label in the label set $\{1, 2, \ldots, C\}$, respectively.
- Balanced Accuracy (B.Acc): The balanced accuracy is a variant of the accuracy metric for imbalanced datasets, which is defined as the macro-average of the recall obtained on each class. Formally, the balanced accuracy is defined as:

$$B.Acc = \frac{1}{C}\sum_{c=1}^{C}R_c \tag{41}$$

, where $R_c = \frac{TP_c}{TP_c+FN_c}$ and TP_c, FN_c denote the true-positives and false-negatives for the cth label in the label set $\{1, 2, \ldots, C\}$, respectively.

5.4. Implementation details

We train a 128-dimensional word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) on the training set of each dataset to initialize the word embeddings of all models including the baseline models. The

Table 2
The performance comparison of all approaches in terms of Macro-averaged F1 Score (M.F1) and Balanced Accuracy (B.Acc) on all four datasets. The best performing approach is shown in bold. Note that we use Sports, Toys, Home, Movies to indicate the dataset Sports & Outdoors, Toys & Games, Home & Kitchen, and Movies & TV, respectively.

Method	Sports		Toys		Home		Movies	
	M.F1	B.Acc	M.F1	B.Acc	M.F1	B.Acc	M.F1	B.Acc
BiGRU-Attention	54.21	53.03	53.54	52.82	59.32	58.03	61.14	59.80
DARLM	49.60	47.95	50.58	48.67	54.49	53.43	57.75	53.96
Max	53.27	52.64	55.02	53.64	58.31	57.36	60.66	59.34
Max-copy	53.95	52.53	53.52	52.01	58.85	58.05	60.60	59.25
HSSC	53.49	51.99	54.24	53.66	58.51	57.42	60.67	59.23
HSSC-copy	53.14	52.63	54.38	53.32	58.78	58.02	60.68	59.32
Dual-view	56.31	54.28	55.70	54.06	60.73	59.63	62.00	60.52
LGDSC	**57.51****	**56.17****	**58.02****	**58.42****	**61.26****	**60.63****	**62.26***	**60.85****

*Indicates statistical significance at p-value < 0.05 using the paired t-test with regard to the strongest baseline Dual-view.
**Indicates statistical significance at p-value < 0.01 using the paired t-test with regard to the strongest baseline Dual-view.

vocabulary is defined as the 50,000 words that appear most frequently in the training set. In the experiment, d_e is set to 128, d_u is set to 512, δ is set to 0.5, β_g, β_s, β_t and β_c are set to 0.8, 0.2, 0.2 and 0.2, respectively. We use the Adam optimization algorithm (Kingma & Ba, 2014) with an initial learning rate of 0.001 and a batch size of 32. If the validation set loss stops decreasing, the learning rate will be reduced by half. In the testing phase, we use the sentiment labels predicted by the source-view classifier as the final classification prediction. The reason is that the decoder has the problem of exposure bias during the test (Ranzato, Chopra, Auli, & Zaremba, 2015), which affects the performance of the summary-view classifier during classification.

5.5. Overall performance

To demonstrate the overall performance of our approach LGDSC, we compare it with seven strong baselines. The overall performance in terms of both M.F1 and B.Acc on four datasets is shown in Table 2. From Table 2, we can observe that among the two single sentiment models, BiGRU-Attention achieves a better performance than DARLM. This may be because DARLM is mainly designed to address the sentence classification, and it would be not effective to deal with the review texts which usually are comprised of multiple sentences. Comparing with the single sentiment model (e.g., BiGRU-Attention and DARLM), these joint sentiment and summarization models usually demonstrate a superior performance. The best performing baseline method is Dual-view which significantly outperforms all baselines in terms of both M.F1 and B.Acc.

Our approach LGDSC outperforms all baseline methods on all four datasets in terms of both metrics. More precisely, the relative performance improvement of LGDSC over the best performing baseline (i.e., Dual-view) are 2.13% (3.48%) on Sports, 4.17% (8.07%) on Toys, 0.87% (1.68%) on Home and 0.42% (0.55%) on Movies in terms of the metric M.F1 (B.Acc). We have conducted significant tests based on t-test, and the results suggest that LGDSC has a significant improvement over the best performing baseline. The result shows the effectiveness of incorporating the matching clues between text words and class labels into the learning process of text representation as it can force the model to attend to the most salient texts with respect to the class label. Moreover, it also demonstrates the effectiveness of the generated label description by introducing a novel discrimination capabilities based word importance measurements, i.e., Inverse Label Entropy (ILE) based word importance score.

5.6. Ablation study

In this section, we perform ablation study to analyze the role of each component in our model LGDSC. In particular, we have the following variants:

- **-SA:** Instead of utilizing the dual-channel label-guided attention network (DLAN), we leverage a single-channel label-guided attention network by removing the self-attention channel from the dual-channel label-guided attention network. Note that removing the self-attention channel will affect the representation learning from both source-view and summary-view since our approach LGDSC learns a dual-view representations by applying the dual-channel label-guided attention network on both source-view and summary-view.
- **-LGA:** Similar to "-SA", we employ a single-channel label-guided attention network by removing the label-guided attention channel from the dual-channel label-guided attention network. Note that removing the label-guided attention channel will affect the representation learning from both source-view and summary-view as well.
- **-DLAN:** We modify the representation learning of both the source-view and the summary-view at the same time by replacing the dual-channel label-guided attention network with a simple attention mechanism.
- **-DLANSource:** We update the source-view representation learning by replacing the dual-channel label-guided attention network with the attention mechanism with the glimpse operation (Vinyals et al., 2016).
- **-DLANSummary:** We revise the summary-view representation learning by replacing the dual-channel label-guided attention network with the attention mechanism with the glimpse operation.
- **-Full:** This is our propose approach LGDSC, which learns a dual-view representation by applying the dual-channel label-guided attention network on both source-view (input review text) and summary-view.

The results of the ablation study for all datasets are shown in Table 3. First, we observe that removing one channel, e.g., "-SA" or "-LGA", results in a significant performance degradation on all datasets in terms of both M.F1 and B.Acc. For example, removing the SA channel will cause a degradation of 1.12% and 2.62% on the Toys dataset in terms of M.F1 and B.Acc, respectively. Similar trends can be observed on other three datasets. Second, removing the Label-guided Attention (LGA) channel mostly cause a higher performance degradation as compared with removing the SA channel. For example, on the Toys dataset, removing the Label-guided Attention (LGA) channel will lead to a performance degradation of 2.20% and 4.71% in terms of M.F1 and B.Acc, respectively. Third, removing the dual-channel label-guided attention network from both views (i.e., -DLAN) or one of the two views (i.e., -DLANSource and -DLANSummary) will result in a considerable performance degradation as compared to the proposed model (i.e., Full). For example, on the Toys dataset, "-DLANSource" is inferior than "Full" with a performance degradation of 3.18% and 6.49% in terms of M.F1 and B.Acc, respectively. Similar results can also be observed for the variant "-DLANSummary". In addition, the

Table 3
Results of ablation study on four datasets (i.e., Sports, Movies, Home and Toys).

Method	Sports		Toys		Home		Movies	
	M.F1	B.Acc	M.F1	B.Acc	M.F1	B.Acc	M.F1	B.Acc
-SA	57.19	55.26	57.38	56.93	60.31	59.58	61.89	60.56
-LGA	57.18	55.57	56.77	55.79	58.80	59.64	61.95	60.29
-DLAN	57.09	55.52	55.74	56.61	60.57	59.32	61.95	60.50
-DLANSource	57.20	55.75	56.83	55.99	61.23	60.25	61.88	60.62
-DLANSummary	57.36	55.56	56.23	54.86	60.63	59.94	62.11	60.38
Full	**57.51**	**56.17**	**58.02**	**58.42**	**61.26**	**60.63**	**62.26**	**60.85**

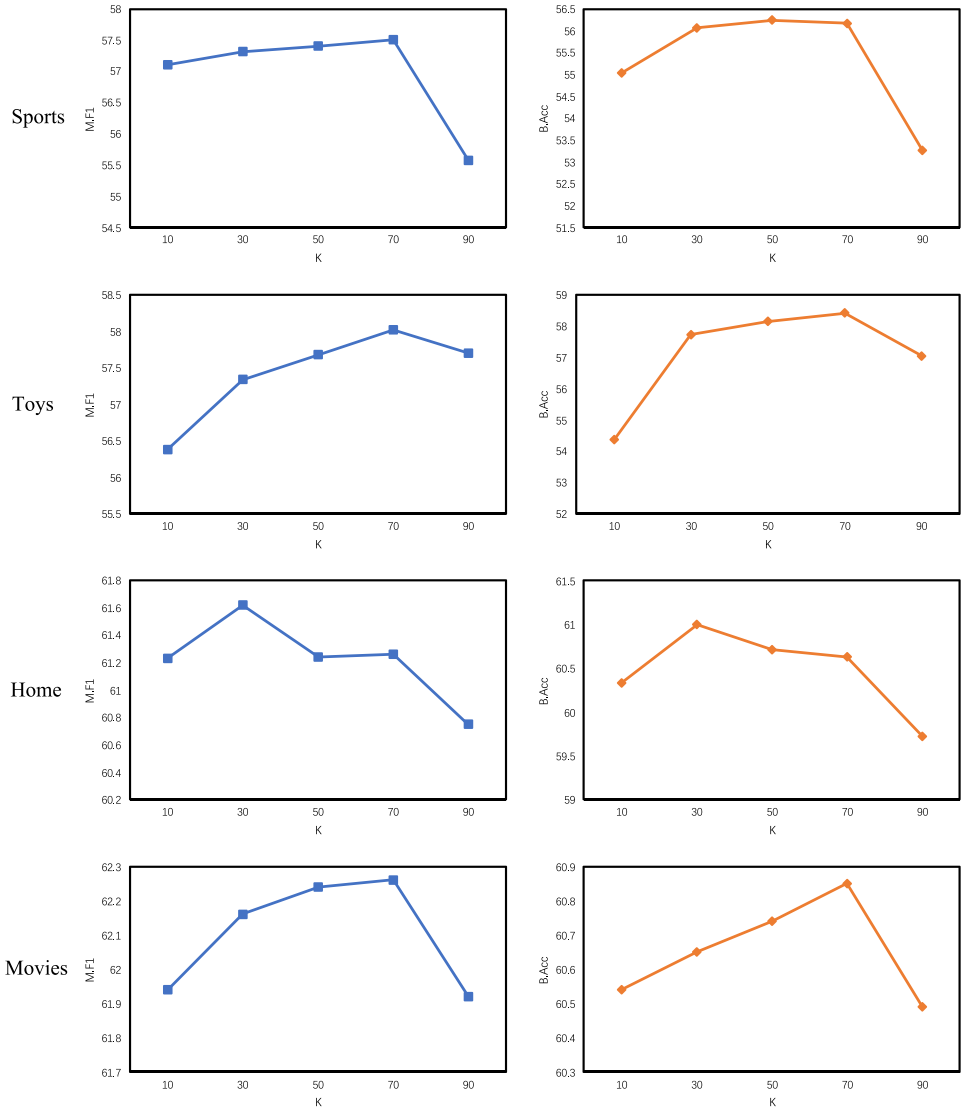


Fig. 4. The performance of LGDSC on four datasets with different number of words K which is selected as the description of a sentiment label.

performance of "-DLAN" is inferior to that of both "-DLANSource" and "-DLANSummary", which shows that the dual-channel label-guided attention network plays an important role in the representation learning of both the source-view and the summary-view.

5.7. Model sensitivity

In this section, we study the sensitivity of the proposed approach LGDSC to the parameter K, and also explore the performance of LGDSC with respect to different proportion of training data.

Parameter K. We first look into the parameter K, which is the number of words selected for generating the description of sentiment labels. Fig. 4 shows the performance of LGDSC on four datasets with the K value varying as $\{10, 30, 50, 70, 90\}$. From the figure, we can observe that the value for K affects the performance of LGDSC in both M.F1 and B.Acc. On the Sports dataset, the performance of LGDSC first keeps rising and achieves the highest M.F1 and B.Acc when K equals to 70. After that, it starts to drop quickly. The performance of LGDSC on the Toys and Movies datasets shares a similar trend as that on the Sports dataset. For the Home dataset, both the M.F1 and B.Acc increase quickly and reach the peak when K equals to 30, and then start to decline. The changing trend is reasonable as some important words would be overlooked when K is small, the performance keeps raising when more useful words are selected as the indicator of the

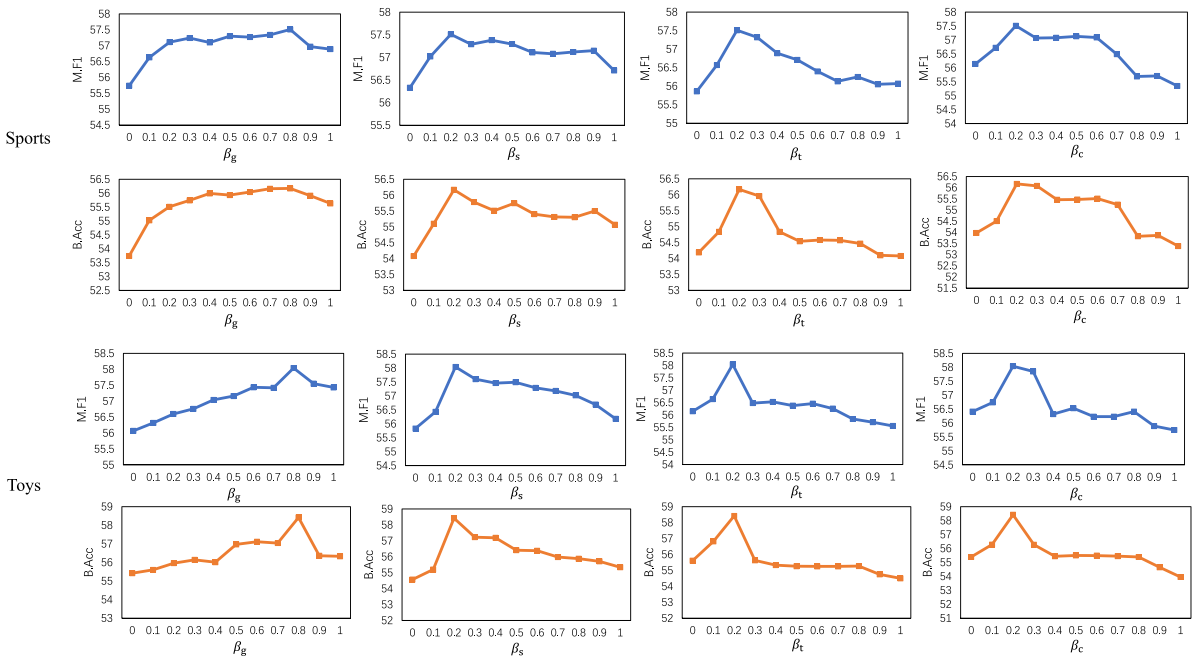


Fig. 5. Impact of β_g, β_s, β_t and β_c to the performance of the proposed model LGDSC on the two datasets (i.e., Sports and Toys).

label. However, when K is too large, more noisy words would be introduced which will inevitably affect the performance of the proposed method. The results verify that the proposed strategy of generating label description is effective as it can select semantically relevant words as the description of a label, as well as maintain a high discriminative capability among different labels.

Parameter β_g, β_s, β_t and β_c. We analyze the impact of the four hyper-parameters β_g, β_s, β_t and β_c to our proposed model, where β_g, β_s, β_t, β_c weight the summarization loss, the source-view sentiment classification loss, the summary-view sentiment classification loss and the inconsistent loss in the objective function, respectively. To study the influence of the individual parameter on the classification results, we vary the target parameter from 0 to 1.0 with a step size 0.1, while keeping the other three parameters fixed. Fig. 5 shows the performance of the proposed model on the Sports and Toys datasets with respect to the metrics M.F1 and B.Acc. On both datasets, the performance of LGDSC continues to raise when we increase β_g and reaches the peak when $\beta_g = 0.8$. If we further increase β_g, it starts to decrease. This indicates that the summarization quality plays a substantial role in the proposed model. For the weight of the source-view sentiment classification loss β_s, we can see that on both datasets the performance of LGDSC rises quickly with the increase of β_s and reaches the peak when $\beta_s = 0.2$. After that, it starts to drop gradually. Similar trends are observed for β_t and β_c.

Impact of the Size of Training Data. To evaluate the performance of LGDSC, we compare it with two state-of-the-art baseline approaches, i.e., HSSC (Ma et al., 2018) and Dual-view (Chan et al., 2020), with respect to different proportion of training data $\{20\%, 40\%, 60\%, 80\%, 100\%\}$. From Fig. 6, we can observe that with the growth of training data, the performance of all approaches raises gradually in terms of both M.F1 and B.Acc. Moreover, our proposed approach LGDSC consistently outperforms the two most competitive baselines at all different proportions of training data.

5.8. Comparison of different generation strategies of label description

In this section, we compare our proposed model LGDSC, i.e., LGDSC(TFIDF-ILE), with six different generation strategies of label description. LGDSC(TF) is a variant which only leverages the Term-Frequency (TF) as the measurement to select the top-K words for describing sentiment labels. LGDSC(TF-ILF) and LGDSC(TF-ILE) further incorporate the discrimination capability, i.e., Inverse Label Frequency (ILF) and Inverse Label Entropy (ILE), into the LGDSC(TF) model, respectively. Similarly, LGDSC(TFIDF) is the variant which utilizes the TFIDF (Ramos et al., 2003) as the measurement to select the top-K words. And LGDSC(TFIDF-ILF) and LGDSC(TFIDF-ILE) are the two variants of introducing ILF and ILE into the LGDSC(TFIDF) model, respectively.

As shown in Table 4, we can observe that: (1) Considering the TF based variants, incorporating the discrimination capability, such as ILF and ILE, usually improve the performance effectively. For example, both LGDSC(TF-ILF) and LGDSC(TF-ILE) have demonstrated superior performance to LGDSC(TF); (2) The TFIDF based variants usually show a better performance when comparing with their corresponding TF based variants; (3) The ILE based variants, such as LGDSC(TF-ILE) and LGDSC(TFIDF-ILE), demonstrate superior performance to their corresponding variants; (4) Our proposed model LGDSC, i.e., LGDSC(TFIDF-ILE), consistently outperforms all other variants.

5.9. Case study

In Table 5, we present a randomly sampled example to illustrate the attention distribution of words generated by our LGDSC and two most competitive baselines, i.e., HSSC (Ma et al., 2018) and Dual-view (Chan et al., 2020), which demonstrates how the rationale behind of proposed approach. We use the red color of the background to indicate the attention scores of words. The darker the color of a word, the higher attention score of that word. From Table 5, we can see that HSSC pays more attentions to words, such as “9”, “old”, “one”, and “big”, which are less important for identifying the sentiment of the review. Dual-view performs better than HSSC as it pays attentions to words, e.g., “fun” and “loves”, which carry useful signals for capturing the sentiment of the review. However, Dual-view casts attentions to some less sentiment relevant words, such as “tennis ball” and “geode”, while attentions paid to “fun” and “loves” are relatively small. Moreover, some important words, like “surprise”, are even overlooks. In contrast, in our approach LGDSC, these important words, such as “properly”, “surprise”, and “fun”, are assigned with larger attention weights. Meanwhile, less important words are assigned with relatively very small

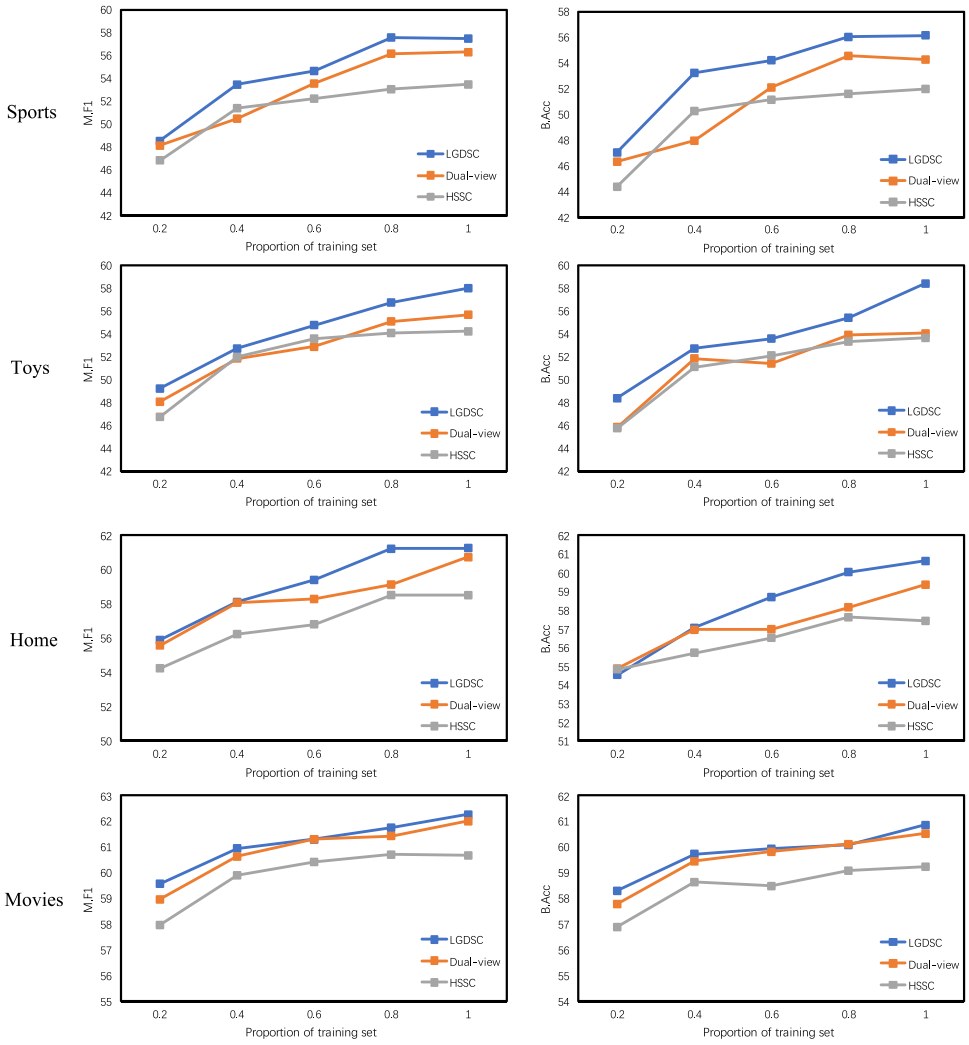


Fig. 6. The performance of LGDSC with respect to different proportion of training data (20%, 40%, 60%, 80%, 100%).

Table 4
Comparison of different generation strategies of label description. The best results are in bold. Note that we use Sports, Toys, Home, Movies to indicate the dataset Sports & Outdoors, Toys & Games, Home & Kitchen, and Movies & TV, respectively.

Method	Sports		Toys		Home		Movies	
	M.F1	B.Acc	M.F1	B.Acc	M.F1	B.Acc	M.F1	B.Acc
LGDSC (TF)	56.71	55.13	57.42	56.43	60.62	59.25	61.56	59.87
LGDSC (TF-ILF)	57.08	55.27	57.63	56.30	60.66	59.49	62.02	60.40
LGDSC (TF-ILE)	57.25	55.45	57.77	56.72	60.78	59.85	62.05	60.57
LGDSC (TF-IDF)	57.14	56.14	57.99	56.85	61.03	60.33	61.89	60.56
LGDSC (TFIDF-ILF)	57.36	55.67	57.73	57.98	60.43	60.10	62.10	60.77
LGDSC (TFIDF-ILE)	**57.51**	**56.17**	**58.02**	**58.42**	**61.26**	**60.63**	**62.26**	**60.85**

attention weights. Through this comparison, we can observe that our proposed approach can make the attention mechanism more effective as compared with state-of-the-art baseline methods, and has a better capacity to capture the sentiment characteristics within a review text.

5.10. Discussion

In this subsection, we discuss where the proposed approach succeeds and where it fails. As mentioned before, a major deficiency of existing methods is that they heavily rely on the availability of label content, and become impracticable when label content is unavailable. The proposed approach aims at automatically generating informative label descriptions by developing a novel inverse label entropy based word importance measurement. It is applicable in scenarios where label content is unavailable, and achieves the state-of-the-art performance.

Despite the successes the proposed approach in the task of sentiment analysis, it still suffers from several limitations. First, the label description is generated with a pre-extraction strategy, i.e., we estimate the word importance by the inverse label entropy based word importance measurement. It is better to design an adaptive way to generate effective label description and train the model in an end-to-end way. Second, the proposed approach employs a summary decoder to generate a summary and introduces a summarization loss to guide the summary-view representation learning process. It will becomes impracticable when the corresponding summary information of each input text is unavailable.

Table 5
Attention distribution of words generated by our LGDSC and two most competitive baselines, i.e., HSSC (Ma et al., 2018) and Dual-view (Chan et al., 2020). The darker the color of a word, the higher attention score of that word.

Model	Review	Predict label	True label
HSSC	it is just as it states one tennis ball sized geode ... it was fairly easy to bust open with a hammer and did so properly , 2 halves ... my 9 yr old son loves gems , rocks , etc and this was fun for him to crack open and see the surprise inside ... i opted for this one over others that had several small ones since the real fun i think ends after the first one ... so go big	3	5
Dual-view	it is just as it states one tennis ball sized geode ... it was fairly easy to bust open with a hammer and did so properly , 2 halves ... my 9 yr old son loves gems , rocks , etc and this was fun for him to crack open and see the surprise inside ... i opted for this one over others that had several small ones since the real fun i think ends after the first one ... so go big	4	5
LGDSC	it is just as it states one tennis ball sized geode ... it was fairly easy to bust open with a hammer and did so properly , 2 halves ... my 9 yr old son loves gems , rocks , etc and this was fun for him to crack open and see the surprise inside ... i opted for this one over others that had several small ones since the real fun i think ends after the first one ... so go big	5	5

6. Conclusion

In this paper, we propose a novel label-guided dual-view sentiment classifier LGDSC. Our model generates effective label description by introducing a well-designed measurement, i.e., inverse label entropy based word importance measurement. Moreover, we design a novel DLAN module to learn text representation via two different channels. The DLAN will further serve for learning label-guided text representations from two different views, i.e., the source view and the summary view. At last, on top of each learnt representation, a two-layer feed-forward neural network will be utilized to predict the sentiment label. We evaluate the performance of our propose model on four widely used public datasets, and compare it with seven competitive baseline methods. Experimental results show that our model is consistently superior to all baseline methods in terms of both M.F1 and B.Acc.

CRediT authorship contribution statement

Xiaofei Zhu: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition. **Zhanwang Peng:** Investigation, Methodology, Software, Writing – original draft. **Jiafeng Guo:** Writing – review & editing, Supervision. **Stefan Dietze:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant number 62141201]; the Major Project of Science and Technology Research Program of Chongqing Education Commission of China [grant number KJZD-M202201102]; the Federal Ministry of Education and Research, Germany [grant number 01LE1806A].

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd international conference on learning representations*.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *Proceeding of the 20th international conference on pattern recognition* (pp. 3121–3124). IEEE.

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, *31*(2), 102–107.

Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok, K. (2020). SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *Proceedings of the 29th ACM international conference on information and knowledge management* (pp. 105–114).

Chai, D., Wu, W., Han, Q., Wu, F., & Li, J. (2020). Description based text classification with reinforcement learning. *Vol. 119*, In *Proceedings of the 37th international conference on machine learning* (pp. 1371–1382).

Chan, H. P., Chen, W., & King, I. (2020). A unified dual-view model for review summarization and sentiment classification with inconsistency loss. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 1191–1200).

Chen, C. (2017). Improved TFIDF in big news retrieval: An empirical study. *Pattern Recognition Letter*, *93*, 113–122.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Du, C., Chen, Z., Feng, F., Zhu, L., Gan, T., & Nie, L. (2019). Explicit interaction model towards text classification. *Vol. 33*, In *Proceedings of the AAAI conference on artificial intelligence* (01), (pp. 6359–6366).

Fei, H., Ren, Y., Wu, S., Li, B., & Ji, D. (2021). Latent target-opinion as prior for document-level sentiment classification: A variational approach from fine-grained perspective. In *Proceedings of the web conference 2021* (pp. 553–564).

Gao, W., Yoshinaga, N., Kaji, N., & Kitsuregawa, M. (2013). Modeling user leniency and product popularity for sentiment classification. In *Proceedings of the sixth international joint conference on natural language processing* (pp. 1107–1111).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huang, X., Chen, B., Xiao, L., & Jing, L. (2019). Label-aware document representation via hybrid attention for extreme multi-label text classification. CoRR arXiv:1905.10070.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *Computer Science*.

Kumar, V., Ramakrishnan, G., & Li, Y. (2019). Putting the horse before the cart: A generator-evaluator framework for question generation from text. In *Proceedings of the 23rd conference on computational natural language learning* (pp. 812–821).

Lin, Y., Fu, Y., Li, Y., Cai, G., & Zhou, A. (2021). Aspect-based sentiment analysis for online reviews with hybrid attention networks. *World Wide Web*, *24*(4), 1215–1233.

Ma, S., Sun, X., Lin, J., & Ren, X. (2018). A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence* (pp. 4251–4257).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 55–60).

McAuley, J. J., Targett, C., Shi, Q., & van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 43–52).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceeding of the 27th annual conference on neural information processing systems* (pp. 3111–3119).

Núñez, J., Cincotta, P., & Wachlin, F. (1996). Information entropy. In *Chaos in gravitational N-body systems* (pp. 43–53). Springer.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. arXiv preprint cs/0506075.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070.

Peng, H., Li, J., He, Y., Liu, Y., Bao, M., Wang, L., ... Yang, Q. (2018). Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *Proceedings of the 2018 world wide web conference* (pp. 1063–1072).

Qin, P., Xu, W., & Guo, J. (2016). A novel negative sampling based on TFIDF for learning word representation. *Neurocomputing*, *177*, 257–265.

Ramos, J., et al. (2003). Using tf-idf to determine word relevance in document queries. *Vol. 242*, In *Proceedings of the first instructional conference on machine learning* (1), (pp. 29–48). Citeseer.

Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2015). Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732.

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointergenerator networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (pp. 1073–1083).

Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1422–1432).

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (pp. 1555–1565).

Vinyals, O., Bengio, S., & Kudlur, M. (2016). Order Matters: Sequence to sequence for sets. In *Proceeding of the 4th international conference on learning representations*.

Wang, S., & Jiang, J. (2016). Learning natural language inference with LSTM. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *The 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 1442–1451).

Xiao, L., Huang, X., Chen, B., & Jing, L. (2019). Label-specific document representation for multi-label text classification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 466–475).

Ye, X., Dai, H., Dong, L., & Wang, X. (2021). Multi-view ensemble learning method for microblog sentiment classification. *Expert Systems with Applications*, *166*, Article 113987.

Yuan, X., Wang, T., Gülçehre, Ç., Sordoni, A., Bachman, P., Zhang, S., ... Trischler, A. (2017). Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd workshop on representation learning for NLP* (pp. 15–25).

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), Article e1253.

Zhou, Q., Wang, X., & Dong, X. (2018). Differentiated attentive representation learning for sentence classification. In *IJCAI* (pp. 4630–4636).

Zhu, L., Zhu, X., Guo, J., & Dietze, S. (2022). Exploring rich structure information for aspect-based sentiment classification. *Journal of Intelligent Information Systems*.

Zhu, X., Zhu, L., Guo, J., Liang, S., & Dietze, S. (2021). GL-GCN: global and local dependency guided graph convolutional networks for aspect-based sentiment classification. *Expert Systems with Applications*, *186*, Article 115712.