IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 2 January 2024, accepted 18 February 2024, date of publication 26 February 2024, date of current version 15 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3370858

RESEARCH ARTICLE

Detect Sarcasm and Humor Jointly by Neural Multi-Task Learning

YUFENG DIAO[1], LIANG YANG[2], SHIQI LI[1], ZHANG HAO[1], XIAOCHAO FAN[3], AND HONGFEI LIN[2]

[1]School of Computer Science and Technology, Inner Mongolia Minzu University, Tongliao 028000, China
[2]Department of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
[3]School of Computer Science and Technology, Xinjiang Normal University, Xinjiang 830010, China

Corresponding author: Yufeng Diao (diaoyufeng@imun.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 62366040, Grant 62006130, and Grant 62066044; in part by the Inner Mongolia Science Foundation under Grant 2022MS06028; in part by the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region; in part by the National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian; in part by the Inner Mongolia Direct College and University Scientific Basic, in 2022 and 2023, under Grant 202203016; in part by the Youth Elite Talent Program of Xinjiang Normal University, in 2022, under Grant XJNUQB2022-23; and in part by Inner Mongolia University Doctoral Initiation Fund under Grant 662.

ABSTRACT Sarcasm is a sophisticated speech act that is intended to express contempt or ridicule on social communities such as Twitter. In recent years, the prevalence of sarcasm on the social media has become highly disruptive to sentiment analysis systems due to not only its tendency of polarity flipping but also usage of figurative language. It is observed that sarcastic texts often convey a humorous effect. Thus, determining the humor in texts can be pertinent to the successful detection of sarcasm, and vice versa. However, current works always regard sarcasm detection and humor identification as separate tasks. In this paper, we argue that these tasks should be treated as a joint, collaborative, effort, considering the semantic connections between sarcasm and humor expressed in texts. Enlightened by the multi-task learning strategy, we present a joint architecture that settles two highly pertinent tasks, sarcasm detection and humor identification. As the basic of deep neural networks, we learn both tasks jointly exploring weight sharing to capture the task-specific features for each task and task-cross features between the two tasks. Extensive experiments on real-world datasets demonstrate that our presented model consistently enhances both sarcasm detection and humor identification tasks consistently with the help of the strong semantic relationships, achieving much better performance than state-of-the-art baselines.

INDEX TERMS Sarcasm detection, humor identification, multi-task learning, sentiment analysis.

I. INTRODUCTION

Sarcasm, commonly defined as an ironical taunt to express contempt, is an important element of everyday human communication. With a rapid development of conversational systems and natural language processing applications on social medias, the task of automatic detection of sarcasm and other types of figurative language has gained a lot of attention [1], [3], [4], [5]. Meanwhile, it is both imperative and intuitive that effective sarcasm detectors are able to bring many benefits for sentiment analysis and opinion mining [2].

Sarcasm is mainly associated to several linguistic features [4], such as an explicit contrast between sentiments or disparity between conveyed emotion and context, which refers to the phenomena as incongruity. Meanwhile, sarcasm is a sophisticated speech act that is intended to express contempt or ridicule [6] in order to convey the humorous effect. Min et al. [42] built a sarcasm detector based on augmentation of potential result and reaction including the result and human reaction caused by its observable content. Therefore, determining the humor of texts should be pertinent to the successful detection of sarcasm, and vice versa.

Humor, a highly intelligent activity in personal communication, provokes laughter or provides amusement [7]. The

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Gaggero.

task of humor identification has captured a lot of attention due to the urge to process large amounts of user generated texts and rise of conversational agents [8]. Many studies put forward interpretable features to model humor [9], [10], [48], [49], such as incongruity, ambiguity and phonetic style.

Inspired by the success of multi-task learning (Caruanna, 1998), [12], we attempt to reinforce sarcasm detection and humor identification together via mutual feedback in a unified joint architecture. Different from existing approaches that regard the two tasks independent, in this study we present a unified joint multi-task model that understand a set of common, bilaterally task-cross features relevant to both of the tasks to facilitate their mutual interaction while each task also has ability to strengthen their task-specific features based on a mutual neural learning process. For task-specific features of each task, this is achieved by using word representation layer, contextual understanding layer and attentive understanding layer to capture different types of semantic information. For task-cross features, we employ a shared layer by a fusion mechanism to accommodate their mutual corresponding parameters. Benefited from not only having more data for training, the usage of multi-task learning architecture also reduces overfitting to each individual task. Hence, the learned representation can result in more compact methods than those built from surface-form features on a single task. Experimental results demonstrate that the joint multi-task learning on the two sarcastic-related tasks together can enhance the performance of each task importantly relative to learning them in separate.

Briefly, the prime contributions of this work can be summarized as follows:

- To the best of our knowledge, this is the first work that aims to tackle sarcasm detection and humor identification together in a unified joint method via multi-task learning, which successfully learns the incongruity and classifies data for two core tasks.
- We present a multi-task frame work for capturing specific features based on RNNs, attention and contextual embeddings due to each task, and mutual features by a gated shared mechanism between tasks to improve the performance of two tasks.
- We empirically verify our presented approach via extensive experiments on real-world datasets, showing that our multi-task method significantly achieves the state-of-the-art baselines on both tasks.

The rest of this paper is structured in the following. Section II mainly reviews the related work on sarcasm detection and humor identification. Section III introduces our proposed model in details. Section IV describes the experimentation details and result analysis. At last, Section V draws the conclusion and offers the future work.

II. RELATED WORK

In this section, we introduce a brief review of the research related to ours in three main areas: sarcasm detection, humor identification and multi-task learning.

A. SARCASM DETECTION

Sarcasm is a complex linguistic phenomenon that has long fascinated both linguists and NLP researchers. There are many theories to research the sarcasm, such as Theory [14] and the Negation Theory [15]. Naturally, many works in this area have regarded the sarcasm detection task as a standard classification problem and attempted to discover the lexical and pragmatic indicators to detect the sarcasm. Manual feature engineering-based approaches often designed a wide effective range of features, such as ngram [19], syntactic patterns [16], sentiment lexicons [17], word frequency [18], word shape and pointedness features [20] and readability [21]. These above features always have proposed based on incongruity, whether in sentiment, number or situation of context.

Deep learning based methods have attracted considerable interest in many areas of NLP in recent years. Ghosh and Veale [22] proposed CNN-LSTM-DNN model to get the excellent results for sarcasm detection. According to the social media Twitter, Zhang et al. [23] leveraged a recurrent neural network and gated pooling scheme for detecting sarcasm. Meanwhile, a comprehensive survey on automatic sarcasm detection had been explored by Joshi et al. [24]. Based the contextualized representation ELMo, [13] applied a character-level model to extract the complex morpho-syntactic features and serve as indicators for sarcasm detection. Reference [44] employed a bert-based SD method in order to derive dynamic commonsense knowledge and fused the knowledge to enrich the contexts with attention model. Reference [45] used a novel approach with the contradictory nature of sarcasm. Min et al. [42] built a sarcasm detector based on augmentation of potential result and reaction including the result and human reaction caused by its observable content. Liu et al. [43] designed several prompt templates in order to mimic the actual intention behind the sarcastic literal content and proposed a simple prompt tuning method SarcPrompt for recognizing sarcasm in texts.

B. HUMOR IDENTIFICATION

Humor, as a human-specific attribute, plays a significant role in human communication. With the encouragement of exploiting the essence of humor, great progress has been made in the research of humor theories. There are many well recognized theories, including superiority theory [25], relef theory [26] and incongruity theory [27], which explain the origin and essence of humorous feelings. Yang et al. [10] put forward interpretable features to model humor, such as incongruity, ambiguity, interpersonal effect and phonetic style. Liu et al. [9] exploited the syntactic structures to reveal stylistic characteristics of humor by using constituent parsing and dependency parsing.

Further improvements both in terms of classic and deep learning approaches came as a result of humor identification task. Chen and Soo [7] implemented a convolutional neural network with extensive filter size, number and highway networks to increase the depth of networks. Weller and Seppi [28] (transformer) employed a transformer architecture in learning from sentence context to detect the humor. Zhao et al. [29] used a novel tensor embedding method that could effectively extract lexical features for humor recognition and achieved the best results. Fan et al. [49] proposed a phonetics and ambiguity comprehension gated attention network for humor recognition. Ren et al. [50] presented an attention network via pronunciation, lexicon and syntax model which contained the pronunciation understanding unit, lexicon understanding unit, syntax analysis unit and context understanding unit for humor recognition. Inácio et al. [46] presented a re-implementation of the previous state-of-the-art method for Humor Recognition in the Portuguese language, alongside a novel fine-tuned BERT model for the same task, reaching a nearly-perfect F1 score of 99.64%. Najafi-Lapavandani et al. [47] introduced a new Persian dataset for humor detection and presented a pre-trained language model trained on the dataset. Zhao et al. [48] proposed a new implicit sentiment analysis framework (KIG) which consisted of Higher-quality graph structures and node representations for the joint iterative learning of graph structures and multi-view knowledge fusion.

C. MULTI-TASK LEARNING

The general idea of multi-task learning framework should date back to (Caruanna, 1998) that is mainly to enhance the performance of one task leveraging other associated tasks. Most of multi-task learning or joint learning methods can be considered as parameter sharing ways. These models can be jointly trained and their parameters are shared across multiple tasks. Multi-task learning has received great traction within the NLP research community in recent years and are able to settle with many NLP applications, such as word segmentation, POS tagging and dependency parsing [30], [31], [32], and more on text classification [12], (Ma et al., 2018).

In the context of neural models, multi-task learning has been proven effective in many related problems. Collobert and Weston [33] investigated an unified architecture by using a shared lookup table as inputs, and then jointly learned a lot of NLP tasks based on convolutional neural networks, such as part-of-speech tagging, named-entity recognition and semantic role labeling. In order to settle with query classification and ranking for web search, Liu et al. [12] explored a multi-task deep neural network to learn the shared representations for arbitrary text among multiple tasks. Luong et al. [34] applied multi-task sequence-to-sequence models to learn the ensemble of a wide range of tasks, such as syntactic parsing, machine translation, image caption. Liu et al. [12] developed three RNN-based frameworks to build text sequence that imported different information sharing mechanisms to deal with multiple text classification tasks. Enlightened by the multi-task learning strategy, Ma et al. (2018) proposed a joint architecture that unified the two highly pertinent tasks, rumor detection and stance classification.

For most of these methods, multi-task frameworks basically share some layers across all tasks to determine task-cross features, while the remaining layers can be learned task-specific features. Our model is inspired from the general sharing structure as the basis of RNN-based multi-task learning [12], (Ma et al., 2018). Our main challenge lies in designing an effective shared weighting approach to extract the task-specific features of each task and task-cross features for two tasks and improve the performance of sarcasm detection and humor identification.

III. METHOD

In this section, we will introduce our tasks, and describe our proposed joint model in details.

A. PROBLEM FORMULATION

Our goal is to formulate a multi-task model that jointly learns the sarcasm detection and humor identification models, where one task does not employ data from the same source as the other. For instance, we can typically use online debate dataset in sarcasm detection but use the different sources in humor identification, considering the availability of training data and specific setting.

Sarcasm Detection

We consider this task as a supervised sequence classification problem, which learns a classifier f to identify this sentence $\{x_{1i}, x_{2i}, \ldots, x_{ni}\}$ sarcastic or not. That is, f: x_{1i}, x_{2i},..., $x_{ni} \rightarrow Y_i$.

Humor Identification

This task refers to determine the humorous type of orientation that each sentence expresses in the social media. We formulate it as a sequence labeling or sequence classification problem depending on a classifier g. That is, g: x_{1j}, x_{2j},..., $x_{mj} \rightarrow Y_j$

B. MULTI-TASK LEARNING NETWORK

In this section, our model is composed of task-specific unit and task-cross unit to jointly detect the sarcasm and identify the humor by the multi-task learning framework. The task-specific unit consists of embedding layer, contextual understanding layer and attentive understanding layer, which imports the contextual representation and attentive contextual information. While task-cross unit includes fusion share layer and task classification layer to accommodate their mutual corresponding parameters of two tasks. With this design, our proposed model is capable of paying attention to the task-specific features and task-cross features for jointly sarcasm detection and humor identification. Figure 1 demonstrates the architecture of our module with the overall sarcasm detection and humor identification system. In the following sections, we would introduce the details of these parts.

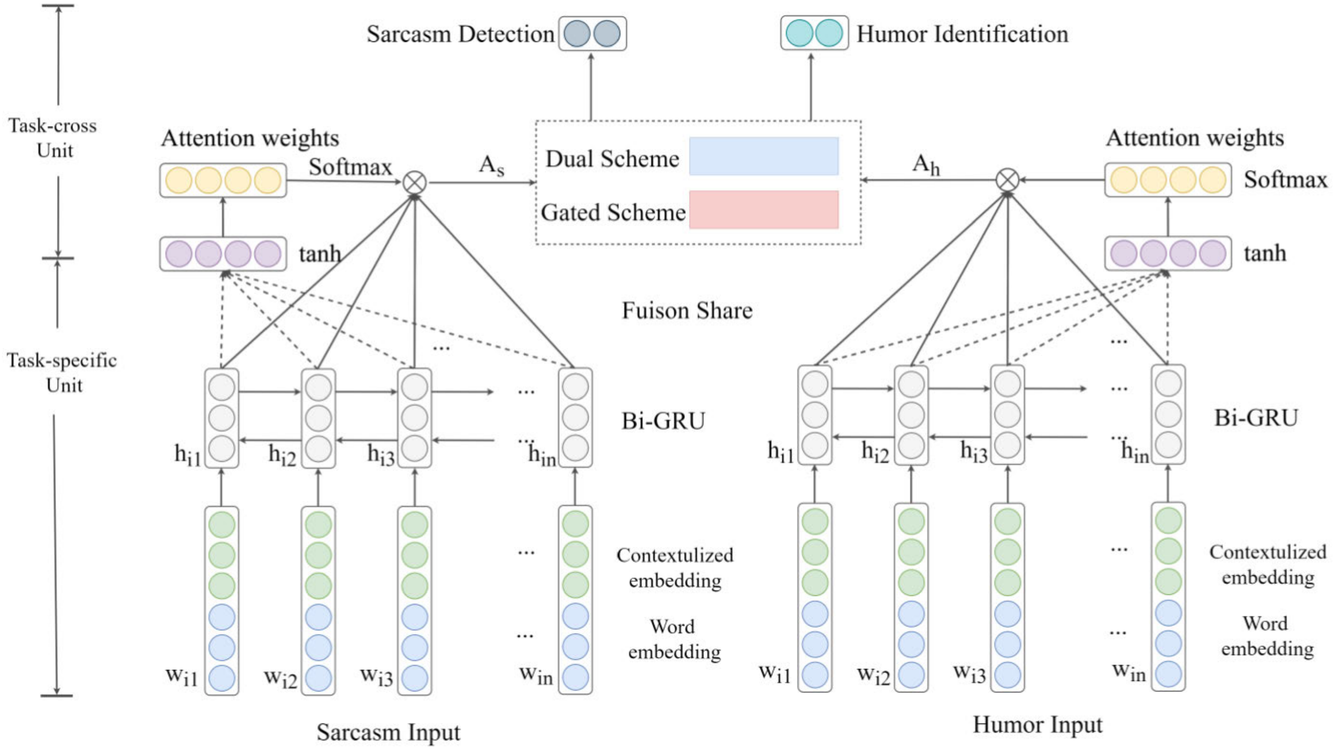

FIGURE 1. The architecture of multi-task learning network for sarcasm detection and humor identification (MTSH).

C. TASK-SPECIFIC UNIT NETWORK

Compared to standalone learning models, multi-task learning approach can take advantage of the related tasks to learn complex signals indicative of sarcasm and humor information. Here, we explore the task-specific unit to extract the own effective features to express the incongruity of each task. In the following subsections, we will describe the details of these parts.

1) EMBEDDING LAYER

In order to make use of incongruity of sarcasm or humor, contextualized representation and word representation could be imported into our framework. Our system receives a tokenized sentence as inputs and maps it into embedding layer, by concatenating representations from pre-trained contextualized representation ELMo model and word representation GloVe model as following.

2) CONTEXTUALIZED REPRESENTATION LAYER

A model requires understanding a high-quality semantic representation, which provides rich syntax and semantic relationship from a linguistic context to identify the incongruity of sarcasm or humor. Here, ELMo [37] is applied as the contextualized representation that is a function based on the input sequence. This model concludes two-layer Bi-LSTM by combining with convolutional features to pre-train the embeddings, where the source is come from a large scale dataset.

Firstly, the probability of the sequence (t_1,t_2, ..., t_N) is calculated for a forward language model as the basic of the probability of the given history (t_1,t_2, ..., t_{k-1}) with token t_k as below:

$$p(t_1, t_2, \ldots, t_N) = \prod_{K=1}^{N} p(t_k | t_1, t_2, \ldots, t_{k-1}) \tag{1}$$

Generally, neural language models are capable of indicating an independent contextual semantic relationship and then transpose it into L layers based on a forward LSTM. Therefore, a contextual dependent embedding is the corresponding output for each LSTM layer, and its position is k and j is range from 1,...,L. The output of the top layer $\overrightarrow{h_{k,L}}$ of a forward LSTM layer is mainly used to predict the next token t_{k-1} by a softmax layer.

Moreover, a backward LM model is similar with the above forward language model (LM). It also is able to identify the previous token according to the given following context except reversing the sequence as follows:

$$p(t_1, t_2, \ldots, t_N) = \prod_{K=1}^{N} p(t_k | t_K, t_{K+1}, t_{k+2} \ldots, t_N) \tag{2}$$

For the prediction, we can get the backward LSTM $\overleftarrow{h_{k,L}}$in an L layer deep model with the similar to forward LM, where the token is t_k and its given context is $(t_{k+1}, t_{k+2}, \ldots, t_N)$. So a biLM, the combination of forward LM and backward LM, would attempt to maximize the log likelihood of the final biLM.

Finally, ELMo representation and intermediated layer representations are combined as the contextualized representation in a biLM. For each token t_k, an L layer biLM can generate a set of 2 L+1 representations as follows.

$$E_c = \{x_k, \overrightarrow{h_{k,j}}, \overleftarrow{h_{k,j}} | j = 1, \ldots, L\} = \{h_{K,j}^{LM}\} \tag{3}$$

where $\overrightarrow{h_{k,L}}$ is a forward biLSTM layer and $\overleftarrow{h_{k,j}}$ is a backward biLSTM layer.

Therefore, the contextualized representation Rc is leveraged to provide the rich complex syntax and semantic features to better understand the linguistic contexts for detecting the incongruity of sarcasm or humor.

3) WORD REPRESENTATION LAYER

Word representation is able to capture the latent regularity vector between words that is suitable to settle with NLP tasks. Hence, we also explore word embeddings as the basic features to detect the sarcasm and humor. GloVe [35], as a pre-trained embedding approach, usually initializes and fine-tunes the parameters in the training procedure. Here, the word representation R_w is obtained and its dimension is d_w.

Our multi-task system explores a tokenized sentence as input and maps it into embedding E_x, by concatenating representations from ELMo R_E and GloVe R_w, which could better learn the latent contextual connection and detect the sarcasm or humor.

4) CONTEXT UNDERSTANDING LAYER

As we know, recurrent neural based networks (RNNs) have been widely attracted the attention in the NLP areas in order to largely enhance their performance in the sequence modeling. Based on the embedding input Ex, a particularly effective implementation of sequence RNN model, called a Gated Recurrent Unit (GRU) [36], is applied to build the word sequences and then capture the semantic contextual information.

With this design, there is a reset gate, z_t, and an update gate, r_t, at each step t in this GRU. It mainly exploits a current input, x_t, and previous hidden state, h_{t-1}, to generate the immediate hidden state, $h_t = GRU(x_t, h_{t-1})$, in the follows.

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\ \widetilde{h_t} &= F(W_h x_t + r_t \cdot U_r h_{t-1} + b_h) \\ h_t &= z_t \cdot h_{t-1} + (1 - z_t) \cdot h_t \end{aligned} \tag{4}$$

where parameters $\{W_z, W_r, W_h, U_z, U_r, U_h\}$ are the training weights to update in the learning procedure,$\{b_z, b_r, b_h\}$ are the bias terms that randomly initializing the vectors, "·" is regarded as a element-wise multiplication operation. F denotes the activation function that can be set to the tanh function in the following experiments.

It is clearly that an excellent model is the bidirectional gated recurrent unit (Bi-GRU) for many NLP tasks, which is composed of the left-to-right direction GRU $\overrightarrow{h_t}$ and right-to-left direction GRU $\overleftarrow{h_t}$. Finally, h_t is the combination of two GRU states to capture the contextual semantic information from two directions.

Generally, the Bi-GRU is explored to encode a sentence of the combination of word representation and contextualized representation as inputs to obtain the semantic contextual understanding for better detecting sarcasm and humor.

5) CONTEXT UNDERSTANDING LAYER

To capture the significant semantic relationship in the linguistic context for sarcasm detection and humor identification, an attention mechanism is suitable to enable our model to concentrate on the salient significant contextualized information. Here, we leverage the contextual understanding encoded information as inputs to extract the contextual-specific attention signal. The formulas are as follows:

$$\begin{aligned} u_t &= \tanh(w \cdot h_a + b) \\ \alpha_{ti} &= \frac{exp(u_{hi}^T h_{vi})}{\sum_i exp(u_{hi}^T h_{vi})} \\ A_t &= \sum_i \alpha_{hi} h_{ti} \end{aligned} \tag{5}$$

where h_t is the contextual embedding by two-directions Bi-GRU, h_a and h_v are the weighted matrices to update in the training procedure, A_t is the attentive vector after encoding as the input of attentive embedding layer.

After attentive understanding layer, we obtain the task-specific features that are associated with each task, where there are A_s as task-specific features for the task of sarcasm detection and A_h as task-specific for the task of humor detection.

D. TASK-CROSS UNIT NETWORK

As an alternative to task-specific unit network, task-cross unit provides parallel models with dedicated parameters for each task, while also connecting them together to allow for information exchange. This module is composed of fusion share layer and task classification layer, which would be introduced in the following.

1) FUSION SHARE LAYER

This layer is used to learn the mutual semantic information between the tasks of sarcasm detection and humor identification, which could better understand the incongruity and enhance the performance. Here, there are two schemes, dual scheme and gated scheme, to ensemble the connection among different tasks.

Dual Scheme

Inspired by cross-stitch unit [41], it provides the equivalent weighted parameters for each task. This scheme gives α-parameters to regulate the information flow in each direction, which should be optimized in the training process. We apply dual scheme sharing with task-specific features A_s for the sarcasm detection task and A_h for the humor detection task as:

$$F_s = \alpha_{SS}A_S + \alpha_{sh}A_h \tag{6}$$

$$F_h = \alpha_{hh}A_h + \alpha_{hs}A_s$$

where A_s and A_h are the attentive features by task-specific unit network from parallel task sarcasm detection and humor identification. The α-parameters are used to control the directions of information flow, which are initialized with a bias towards favoring the information in the same framework, with $\alpha_{ss} = \alpha_{hh} = 0.9$ and $\alpha_{sh} = \alpha_{hs} = 0.1$. These above values are optimized in the training but remain static during the testing process.

Gated Scheme

The dual scheme learns a single set of shared values for α-parameters in optimization. However, we should construct a network that computes these values dynamically for each input sentence, even at testing time. This allows our model more flexibility and modulates the information flow relying on the particular input sentence. Inspired by Dankers et al. [38], our gated scheme is exploited into each pair of parallel attentive layers, where one gate controls the information flow from the main to the auxiliary task and the other gate regulated from the opposite direction. The following equations detail the gating mechanism:

$$\begin{aligned} g_s &= \sigma(W_s[A_s; A_h] + b_s) \\ F_s &= g_s \cdot A_s + (1 - g_s) \cdot A_h \\ g_h &= \sigma(W_h[A_s; A_h] + b_n) \\ F_h &= g_h \cdot A_h + (1 - g_h) \cdot A_s \end{aligned} \tag{7}$$

where g_s and g_h are the gated mechanisms for sarcasm detection task and humor identification task, W_s and W_h are weight matrices, b_s and b_h are bias vectors, and the bias parameters of gates are initialized with a bias towards one task.

2) TASK CLASSIFICATION LAYER

Follow previous work, we formulate sarcasm detection and humor identification as both classification problems in the following.

In the output layer of sarcasm detection task, we feed the gated share features into a fully connected layer with a softmax activation function to generate the prediction for each sentence.

$$z_s = Softmax(W_{zs} \cdot F_s + b_{zs}) \tag{8}$$

where $z_s \in R^C$ is the vector of predicted probability for sarcasm. Here, C represents the number of classes of sarcasm labels, W_{zs} and b_{zs} are parameters of the sarcasm classification layer. This gives a softmax-probability over whether a sentence is sarcastic or not.

In the output layer of humor identification task, we feed the gated fusion characteristics into a fully connected layer by a softmax activation function to obtain the prediction for each sentence.

$$z_h = Softmax(W_{hs} \cdot F_h + b_{hs}) \tag{9}$$

where $z_h \in R^M$ is the vector of predicted probability for humor. Here, M represents the number of classes of humor labels, W_{zh} and b_{zh} are parameters of the humor classification layer. It mainly computes a softmax-probability over whether a sentence is humorous or not.

E. MODEL TRAINING

In this subsection, we describe the learning and optimization details of our multi-task learning architecture. The parameters of the proposed multi-task learning model for each task are trained to minimize the cross-entropy of the predicted and ground truth distributions as following:

$$L = -\sum_i \sum_j y_i^j \log \widehat{y}_i^j + \lambda ||\theta||^2 \tag{10}$$

where y denotes the ground truth distribution and $\widehat{y}$ denotes the predicted distribution, i is the index of sentence, j is the index of class, L2 regularizer trades off the error and the scale of the model, λ is the trade-off coefficient and θ indicates all parameters.

For optimization method, we train our model using stochastic gradient decent (SGD) by looping over the tasks to minimize the objective function. Meanwhile, we also add dropout strategy to prevent the co-adaptation of the parameters to settle the overfitting problem.

IV. EXPERIMENTS

In this section, we first describe the datasets and evaluation metrics. Then we examine the performance of our multi-task learning model comparison with the current approaches on the tasks of sarcasm detection and humor identification, respectively. At last, we give the detailed of our proposed model.

A. DATASETS AND EVALUATION METRICS

For sarcasm detection task, we make expansion based on a public online debate datasets,It can be viewed and downloaded the data from https://nlds.soe.ucsc.edu/sarcasm1. We explore two open datasets Sarcasm Corpus V1 (IAC-V1) and Sarcasm Corpus V2 (IAC-V2) that are subsets of the Internet Argument Corpus(IAC), It can be viewed and downloaded the data from https://nlds.soe.ucsc.edu/sarcasm1 This corpus is mainly concerned with long text for research on political debates on online forums. Below we describe each dataset in our experiments, please see Table 1 below for a summary.

For humor identification task, we leverage Pun of the Day dataset. In the beginning, this dataset only contains pun content. Then it collects the negative samples to balance the distribution of positive and negative examples for reducing the domain discrepancy. The negative source is from Yahoo!Answer, AP News, Proverb and New York

TABLE 1. Statistics of datasets on sarcasm detection.

Datasets	Train	Valid	Test	Total
IAC-V1	1396	199	400	1995
IAC-V2	3284	469	939	4692

TABLE 2. Statistics of datasets on humor identification.

Datasets	Positive	Negative	Average Length
Pun of the Day	2423	2403	13.5

Times. Table 2 describes a complete statistical information of our dataset.

We apply the following standard criteria precision, recall, accuracy and F1-score for evaluation that also adopted as evaluated metrics in previous work for sarcasm detection and humor identification.

Training Details

We perform 5-fold cross-validation throughout all the experiments. We leverage the GloVe embedding and its dimension is set to 300. Moreover, we employ the ELMo embedding and its dimension is set to 1024. The size of the GRU hidden layer is fixed to 150. We experiment with dropouts ranging from 0.1 to 0.5 and select 0.5 as dropout rate to avoid the overfitting problem. All models are trained by mini-batch of 64 instances.

B. SARCASM DETECTION

Table 3 and Table 4 compare sarcasm detection results on IAC-V1 dataset and IAC-V2 dataset of the following state-of-the-art systems:

- **NBOW** is a simple basic neural baseline with bag-of-words that computes the whole word representations and passes the vectors by a logistic regression layer.
- **CNN** is a Convolutional Neural Network based on a max-pooling operator. CNNs are considered as compositional encoders to extract the n-gram features where the filter width is set to 3 and number of filters f is set to 100.
- **GRNN** is a Bidirectional Gated Recurrent Unit model for detecting sarcasm [23]. The gated pooling scheme is mainly applied to integrate the hidden representation.
- **CNN-LSTM-DNN** is a combination of a CNN, LSTM, and deep neural network by stacking operator to detect the sarcasm [22].
- **MIARN** is an attention-based neural model that captures in-between instead of across, enabling it to explicitly build with a contrastive theme [39].
- **ELMo-BiLSTM** is a deep learning model based on character-level word representations provided from ELMo to achieve the state-of-the-art performance in sarcasm [13].
- **ADGCN** is a GCN-based method with sentic graph and dependency graph. The initial input of GCN is the hidden state of Bi-LSTM (Lou et al., 2021).

TABLE 3. Summary of our obtained results in IAC-V1 dataset. The results with superscript* are reported in Tay et al. [39] and Liu et al. [45].

IAC-V1	P(%)	R(%)	F1(%)
NBOW	57.17*	57.03*	57.00*
CNN	58.21*	58.00*	57.95*
GRNN	56.21*	56.21*	55.96*
CNN-LSTM-DNN	55.50*	54.60*	53.31*
MIARN	63.88*	63.71*	63.18*
ELMo-BiLSTM	65.0*	64.6*	64.4*
ADGCN	64.6*	64.3*	64.3*
DC-Net	66.6*	66.5*	66.4*
MTSH-single	65.21	64.83	65.02
MTSH	**67.73**	**67.79**	**67.76**

TABLE 4. Summary of our obtained results in IAC-V2 dataset. The results with superscript* are reported in Tay et al. [39] and Min et al. [42].

IAC-V1	P(%)	R(%)	F1(%)
NBOW	66.01*	66.03*	66.02*
CNN	68.45*	68.18*	68.21*
GRNN	62.26*	61.87*	61.21*
CNN-LSTM-DNN	64.31*	64.33*	64.31*
MIARN	72.92*	72.93*	72.75*
ELMo-BiLSTM	76.0*	76.0*	76.0*
ADGCN	78.0*	–	78.0*
DC-Net	78.0*	–	77.9*
SD-APRR	78.8*	–	77.8*
MTSH-single	77.33	77.59	77.46
MTSH	**79.34**	**79.88**	**79.61**

TABLE 5. Summary of our obtained results in Pun of the Day dataset. The results with superscript* are reported in [7], [28], and [48].

Pun of the Day	P(%)	R(%)	F1(%)
Word2Vec+HCF	77.6*	83.6*	70.5*
CNN	88.0*	85.9*	86.9*
CNN+F+HN	86.6*	94.0*	90.1*
PACGA	88.69*	92.76*	90.81*
KIG	89.21*	**93.72***	91.15*
MTSH-single	89.29	90.47	89.88
MTSH	**91.34**	91.88	**91.61**

- **SarDeCK8** (Li et al., 2021b) is a BERT-basedSD method that uses the COMET to derivedynamic commonsense knowledge and fusesthe knowledge to enrich the contexts with attention.
- **DC-Net** is a dual-channel sarcasm detection model that reconstructs the text into literal semantics and implied semantics separately and models them independently [45].
- **SD-APRR** is a [42] novel Sarcasm Detector with Augmentation of Potential Result and Reaction inspired by the direct access view.
- **MSTH-single**: Our MTSH model that ignores the humor identification component.
- **MSTH**: Our multi-task model with task-specific and task-cross features to jointly detect the sarcasm and humor.

Table 3 and Table4 report a performance comparison of all benchmarked models on the IAC-V1 and IAC-V2 datasets respectively. The benefit of using multi-task learning is

TABLE 6. Example sentences.

Example	Task	True	Predict
I used to be a banker but I lost interest.	H	1	1
It's a fact taller people sleep longer in bed.	H	1	0
Well you wouldn't have to 5hit yourself everytime you walked down the streets if you had a gun yourself.	S	1	1
When do you advocate breeding blond haired, blue eyed citizens to purify the US?	S	1	0

obvious among all the strong baselines. We observe that our proposed MTSH model obtains the best results across on both datasets, which relative improvement differs across domain and datasets. For IAC-V1 and IAC-V2 datasets, our model enhances over the best baselines with an average of 1.1% to incorporate humor information as additional features. This is because the proposed multi-task architecture cannot just understand the representation of sarcasm detection task itself effectively due to a neural model, but also is able to strengthen the latent features by transferring some useful information from the task of humor identification. Our MTSH model is also higher than MTSH-single, showing that the learned representation is more effective due to the task-cross unit in addition to the shared layer.

Overall, the performance of MTSH is often marginally better than SD-APRR, DC-Net, ADGCN, ELMo-BiLSTM and MIARN. It is clearly that task-specific and task-cross features could be learned the semantic connection from sarcasm detection and humor identification. The performance brought by our additional multi-task learning framework and fusion scheme could be further observed by comparing against CNN-LSTM-DNN, GRNN and CNN. We believe that this is attributed to the fact that more specific and mutual information can be learned by our multi-task learning and fusion share strategy.

At last, the relative performances of competitor approaches are as expected. NBOW performs the worse, because of bag-of-words without any compositional or sequential information. Meanwhile, our proposed MTSH model is effective than other baselines, showing that our multi-task learning can better learn the core of incongruity to capture and transfer the effective connection from sarcasm and humor.

C. HUMOR IDENTIFICATION

Table 5 shows the results on humor identification by comparing the following systems:

- **Word2Vec+HCF**: a computational approach to recognize humor based on semantic structures behind humor and sets of features for each structure [10].
- **CNN**: a convolutional neural network (CNN) to detect the humor, where window sizes would be (5, 6, 7) and filter number is 100 [40].
- **CNN+F+HN**: a convolutional neural network (CNN) with extensive filter size, number and Highway networks to increase the depth of networks [7].
- **PACGA**: a model with speech information and semantic information for implicit sentiment recognition [49].
- **KIG**: a knowledge-fusion-based iterative graph structure learning framework based on co-occurrence statistics, cosine similarity and syntactic dependency trees [48].
- **MTSH-single**: Our MTSH model that ignores the sarcasm detection component to identify the humor.
- **MTSH**: Our multi-task model with task-specific and task-cross features to mutually detect the sarcasm and humor.

Table 5 shows the experiments on Pun of the Day. (1) It is obvious that all the neural network based methods outperform Word2Vec+HCF, showing the effectiveness of deep learning models and also save the human cost. (2) CNN+F+HN is higher than CNN to identify the humor. The reason is that fully connect operation and highway network are suitable to settle with several NLP tasks. (3) PACGA has the better performance because of the abundant semantic information. (4) KIG is better than PACGA by the iterative graph structure which imports the knowledge information and occurrence statistics. (5) Our proposed model MTSH achieves the comparable results which shows that our multi-task learning architecture can both learn the representation of each task and also enhance the connection by transferring useful information between two joint tasks. (6) In our multi-task models, MTSH performs better than MTSH-single, suggesting the improved effectiveness by adding the task-cross unit to each task upon the fusion shared layer.

With these comprehensive experiments on both sarcasm detection and humor identification tasks, we confirm the advantages of our multi-task architecture over a few strong state-of-the-art baselines.

D. DISCUSSION

In this subsection, we show some instances to get a sense of what kinds of sentences are predicted correctly and incorrectly. The examples are shown in table 6.

From the table6, we can see that the predicted correctly sentence "I used to be a banker but I lost interest" shows that our model seems to be able to extract the literal meaning between "banker" and "lost interest" to better identify the humor. Besides, the correct sentence "Well you wouldn't have to 5hit yourself everytime you walked down the streets if you had a gun yourself." suggests that "5hit yourself everytime you walked down the streets" and "had a gun yourself" could result in the inconsistent phenomenon of sarcasm. Model misclassifies certain instances such as "When do you advocate breeding blond haired, blue eyed citizens to purify the US?", which need some common sense derived from a knowledge base to jointly detect the humor and sarcasm. To deal with more subtle cases, our model has room to be improved.

V. CONCLUSION AND FUTURE WORK

As we know, previous research works tackle sarcasm detection and humor identification separately. In this study, we attempt to jointly optimize the two tasks as the basic of unified neural multi-task learning architecture. Generally, we explore a combination of task-specific features and task-cross features, based on contextualized embedding, RNNs, attention model and fusion shared mechanism, which is used to build the sharing information and representation reinforcement between both tasks. The experimental results demonstrate that the multi-task method consistently outperforms many strong baselines for both two tasks, indicating that training these sarcasm-related tasks jointly with multi-task framework seems a better strategy to detect the incongruity of nature sarcasm.

Beyond sarcasm detection and humor identification tasks, we believe that there are other associated tasks to incorporate into such unified framework, such as pun, metaphor together with current tasks.

REFERENCES

[1] H. P. Grice, P. Cole, and J. L. Morgan, "Syntax and semantics," *Logic Convers*, vol. 3, pp. 41–58, Sep. 1975.

[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 1, no. 2, pp. 91–231, 2006.

[3] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process. (Short Papers)*, vol. 2, 2015, pp. 757–762.

[4] Y. Tay, L. A. Tuan, S. C. Hui, and J. Su, "Reasoning with sarcasm by reading in-between," 2018, *arXiv:1805.02856*.

[5] A. Dubey, L. Kumar, A. Somani, A. Joshi, and P. Bhattacharyya, "'When numbers matter!': Detecting sarcasm in numerical portions of text," in *Proc. 10th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2019, pp. 72–80.

[6] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic sarcasm detection: A survey," *ACM Computing Surv.*, vol. 50, no. 5, p. 73, 2017.

[7] P.-Y. Chen and V.-W. Soo, "Humor recognition using deep learning," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., (Short Papers)*, vol. 2, 2018, pp. 113–117.

[8] V. Blinov, V. Bolotova-Baranova, and P. Braslavski, "Large dataset and language model fun-tuning for humor recognition," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4027–4032.

[9] L. Liu, D. Zhang, and W. Song, "Exploiting syntactic structures for humor recognition," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1875–1883.

[10] D. Yang, A. Lavie, C. Dyer, and E. Hovy, "Humor recognition and humor anchor extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2367–2376.

[11] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[12] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," 2016, *arXiv:1605.05101*.

[13] S. Ilić, E. Marrese-Taylor, J. A. Balazs, and Y. Matsuo, "Deep contextualized word representations for detecting sarcasm and irony," 2018, *arXiv:1809.09795*.

[14] D. Wilson, "The pragmatics of verbal irony: Echo or pretence?" *Lingua*, vol. 116, no. 10, pp. 1722–1743, Oct. 2006.

[15] R. Giora, "On irony and negation," *Discourse Process.*, vol. 19, no. 2, pp. 239–264, Mar. 1995.

[16] O. Tsur, D. Davidov, and A. Rappoport, "ICWSM—A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 4, no. 1, May 2010, pp. 162–169.

[17] R. G.-I. Anez, S. Muresan, and N. Wacholder, "Identifying sarcasm in Twitter: A closer look," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Hum. Language Technol., Short Papers*, 2011, pp. 581–586.

[18] F. Barbieri, H. Saggion, and F. Ronzano, "Modelling sarcasm in Twitter, a novel approach," in *Proc. 5th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, France, 2014, pp. 50–58.

[19] A. Reyes, P. Rosso, and T. Veale, "A multidimensional approach for detecting irony in Twitter," *Lang. Resour. Eval.*, vol. 47, no. 1, pp. 239–268, Mar. 2013.

[20] T. Cek, I. Habernal, and J. Hong, "Sarcasm detection on Czech and english Twitter," in *Proc. COLING, 25th Int. Conf. Comput. Linguistics, Tech. Papers*, 2014, pp. 213–223.

[21] A. Rajadesingan, R. Zafarani, and H. Liu, "Sarcasm detection on Twitter: A behavioral modeling approach," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, Feb. 2015, pp. 97–106.

[22] A. Ghosh and D. T. Veale, "Fracking sarcasm using neural network," in *Proc. 7th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2016, pp. 161–169.

[23] M. Zhang, Y. Zhang, and G. Fu, "Tweet sarcasm detection using deep neural network," in *Proc. COLING 26th Int. Conf. Comput. Linguistics, Proc. Conf., Tech. Papers*, Osaka, Japan, 2016, pp. 2449–2460.

[24] A. Joshi, V. Tripathi, K. Patel, P. Bhattacharyya, and M. Carman, "Are word embedding-based features useful for sarcasm detection?" 2016, *arXiv:1610.00883*.

[25] C. R. Gruner, *The Game Humor: A Comprehensive Theory Why We Laugh*. Piscataway, NJ, USA: Transaction PUblishers, 1997.

[26] J. Rutter, *Stand-Up as interaction: Performance Audience Comedy Venues*. Salford, U.K.: Univ. Salford, 1997.

[27] J. M. Suls, "A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis," *Psychol. Humor, Theor. Perspect. Empirical*, vol. 1, pp. 81–100, Jan. 1972.

[28] O. Weller and K. Seppi, "Humor detection: A transformer gets the last laugh," 2019, *arXiv:1909.00252*.

[29] Z. Zhao, A. Cattle, E. Papalexakis, and X. Ma, "Embedding lexical features via tensor decomposition for small sample humor recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 6377–6382.

[30] B. Bohnet and J. Nivre, "A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learning*, 2012, pp. 1455–1465.

[31] J. Hatori, T. Matsuzaki, and Y. Miyao, "Incremental joint approach to word segmentation, POS tagging, and dependency parsing in Chinese," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, Long Papers*, 2012, pp. 1045–1053.

[32] Z. Li, M. Zhang, W. Che, T. Liu, and W. Chen, "Joint optimization for Chinese POS tagging and dependency parsing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 274–286, Jan. 2014.

[33] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 160–167.

[34] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," 2015, *arXiv:1511.06114*.

[35] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[36] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[37] M. E. Peters, M. Neumann, and M. Iyyer, "Deep contextualized word representations," Tech. Rep., 2018.

[38] V. Dankers, M. Rei, M. Lewis, and E. Shutova, "Modelling the interplay of metaphor and emotion through multitask learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 2218–2229.

[39] Y. Tay, A. T. Luu, and S. C. Hui, "Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis," in *Proc. AAAI*, 2018, pp. 5956–5963.

[40] L. Chen and C. M. Lee, "Predicting Audience's laughter using convolutional neural network," 2017, *arXiv:1702.02584*.

[41] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3994–4003.

[42] C. Min, X. Li, L. Yang, Z. Wang, B. Xu, and H. Lin, "Just like a human would, direct access to sarcasm augmented with potential result and reaction," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2023, pp. 10172–10183.

[43] Y. Liu, R. Zhang, Y. Fan, J. Guo, and X. Cheng, "Prompt tuning with contradictory intentions for sarcasm recognition," in *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2023, pp. 328–339.

[44] J. Li, H. Pan, Z. Lin, P. Fu, and W. Wang, "Sarcasm detection with commonsense knowledge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3192–3201, 2021.

[45] Y. Liu, Y. Wang, A. Sun, X. Meng, J. Li, and J. Guo, "A dual-channel framework for sarcasm recognition by detecting sentiment conflict," in *Proc. Findings Assoc. Comput. Linguistics NAACL*, 2022, pp. 1670–1680.

[46] M. Inácio, G. Wick-Pedro, and H. G. Oliveira, "What do humor classifiers learn? An attempt to explain humor recognition models," in *Proc. 7th Joint SIGHUM Workshop Comput. Linguistics Cultural Heritage, Social Sci., Humanities Literature*, 2023, pp. 88–98.

[47] F. Najafi-Lapavandani and M. H. S. Shahreza, "Humor detection in persian: A transformers-based approach," *Int. J. Inf. Commun. Technol. Res.*, vol. 15, no. 1, pp. 56–62, 2023.

[48] Y. Zhao, M. Mamat, A. Aysa, and K. Ubul, "Knowledge-fusion-based iterative graph structure learning framework for implicit sentiment identification," *Sensors*, vol. 23, no. 14, p. 6257, Jul. 2023.

[49] X. Fan, H. Lin, L. Yang, Y. Diao, C. Shen, Y. Chu, and T. Zhang, "Phonetics and ambiguity comprehension gated attention network for humor recognition," *Complexity*, vol. 2020, pp. 1–9, Apr. 2020.

[50] L. Ren, B. Xu, H. Lin, J. Zhang, and L. Yang, "An attention network via pronunciation, lexicon and syntax for humor recognition," *Int. J. Speech Technol.*, vol. 52, no. 3, pp. 2690–2702, Feb. 2022.

YUFENG DIAO was born in 1987. She received the B.S. degree in computer science and technology, in 2009, the M.S. degree in computer application, in 2012, and the Ph.D. degree in computer application technology from Dalian University of Technology, in 2020. She is currently an Associate Professor with the College of Computer Science and Technology, Inner Mongolia Minzu University. Her research interests include text mining and sentiment analysis.

LIANG YANG was born in 1986. He received the B.S. degree in computer science and technology, in 2009, and the Ph.D. degree in computer application, in 2016. He is currently a Lecturer with the School of Computer Science and Technology, Dalian University of Technology. His research interests include sentiment analysis and opinion mining.

SHIQI LI was born in 2000. She received the B.S. degree in data science and big data technology, in 2022. She is currently pursuing the master's degree in medical informatization with the School of Computer Science and Technology, Inner Mongolia Minzu University. Her research interests include text mining and sentiment analysis.

ZHANG HAO was born in 1999. He received the B.S. degree in software engineering, in 2022. He is currently pursuing the degree in electronic information with the School of Computer Science and Technology, Inner Mongolia Minzu University. His research interests include text mining and sentiment analysis.

XIAOCHAO FAN was born in 1982. He received the B.S. degree in computer science and technology, in 2005, and the M.S. degree in computer application, in 2014. He is currently a Lecturer with the School of Computer Science and Technology, Xinjiang Normal University. His research interests include sentiment analysis, text mining, and bioinformatics.

HONGFEI LIN received the B.S. degree in mathematics from Northeastern Normal University, Changchun, China, the M.S. degree in computer application from Dalian University, Dalian, China, and the Ph.D. degree in computer software and theory from Northeastern University, Shenyang, China, in 2000. From 2000 to 2005, he was an Associate Professor with the School of Electronic and Information Engineering, Dalian University of Technology, Dalian, where he has been a Professor with the School of Computer Science and Technology, since 2005. He founded a laboratory of information retrieval (DUTIR), in 2001. His publications run to over 150 articles and he holds seven patents. His research interests include natural language processing, text mining, sentiment analysis, social computing, information retrieval, and bioinformatics. He is a member of the Association for Computational Linguistics, the China Association of Artificial Intelligence, the China Computer Federation, and the Chinese Information Processing Society. He was a member of the IEEE International Conference on Bioinformatics & Biomedicine Program Committee, from 2009 to 2017. He serves several journals as an editor.