SICKNet: A Humor Detection Network Integrating Semantic Incongruity and Commonsense Knowledge

1st Penglong Huang
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China
243031504@qq.com

1st Xingwei Zeng
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China
2112106259@e.gzhu.edu.cn

2nd Jinta Weng*
School of Cyber Security
University of Chinese Academy of Sciences
Beijing, China
wengjinta@iie.ac.cn

3rd Ying Gao
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China
csgy@gzhu.edu.cn

4th Heyan Huang
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
hhy63@bit.edu.cn

Maobin Tang*
School of Computer Science and Cyber Engineering
Guangzhou University
Guangzhou, China
tmb178@gzhu.edu.cn

***Abstract*—Humor is a great linguistic tool to express feelings and enhance social bonding. Limited by the diversity of humor expressions and the differential understanding of listeners, automatic detection of humor text is still a difficult and important area in nature language processing. Current methods of humor detection mainly focus on fine-tuning of pre-trained language models, and rarely consider the degree of humor incongruity and knowledge distinction of contextual environments. To alleviate these challenges, we propose SICKNet, a novel multi-tasks learning network based on the incongruity theory of humor and commonsense knowledge. We first utilize the difference between *set-up* and *punchline* to detect the semantic incongruity of humor, and next use commonsense knowledge to detect the strength of humorous features. SICKNet achieves the start-of-the-art results on Reddit and TaivopJokes datasets, with accuracy rates of 76.27% and 73.64%, respectively. Our code is available at Github[1].**

***Keywords*—Humor Detection, Semantic Incongruity, Commonsense Knowledge, Multi-task learning, Natural Language Processing**

I. Introduction

Humor is an important means of resolving embarrassment, enlivening the atmosphere, and promoting communication, as well as a reflection of human wisdom and creativity [1]. Appropriate humor expression can result in positive effects in counseling, social negotiations and various presentations.

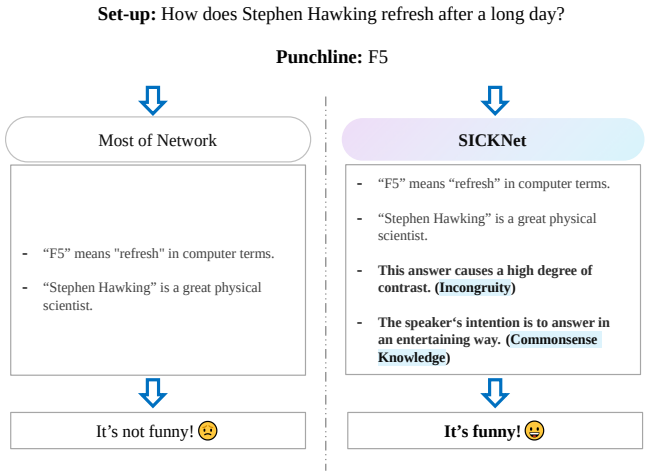

Fig. 1. The novelty we expect from SICKNet in the humor detection process: detection of incongruity and introduction of commonsense knowledge leading to deeper understanding

With the rapid development of natural language processing technology, the humor detection has become one of the hotspots of research. Early methods of humor detection are mainly based on the extraction of humor-specific features [2]. With the development of deep learning and pretrained models, Sun et al. [3] designed a RoBERTa-based [4] humor prediction model with three information, Xie et al. [5] used

*Corresponding author

[1]https://github.com/xing-wei-zeng/SICKNet

GPT-2 [6] language model and calculated the uncertainty and surprisal values of jokes to follow up this work. Amazon researchers [7] used LSTM [8] models for humor detection in their Community Question-Answering (CQA) platform to better understand how engaged users are with their products. Yang et. al [9] used RoBERTa to focus on humor detection work on Facebook about COVID-19. Most of the above work is modeling whole-sentence jokes using neural networks.

In fact, humor detection is complex, and existing models ignore the psychological principles of how humor works. Paulos [10] shows that humorous texts can often be divided into two parts: the *set-up* and the *punchline*. The *set-up* is the narration of the background, while the *punchline* is the continuation and reversal of the *set-up*. Humor maker often makes the joke funny through the contrast between the two part. The incongruity theory [11] also states that humor arises from the violation of the listener's expectations (the *set-up*) due to the incongruent outcome (the *punchline*). Moreover, the introduction of humor-related linguistics has been shown to make models for detecting humor more competitive, as demonstrated by Zhang et al. [12], which divides humorous texts into *set-up* and *punchline* and proposes Multi-Granularity Semantic Interaction Understanding Network to verified this view in the previous work [13]. Mihalcea et al. [14], Yang et al. [15], and Xie et al. [5] similarly established such an approach based on incongruity theory. However, these works do not fully explore degree of inconsistency in which *punchline* break the *set-up*.

In addition, people's understanding of *set-up* and *punchline* relies on an individual's commonsense knowledge such as cultural background and discourse intent. Commonsense knowledge is often used to perform content reasoning, decision making and many other reasoning tasks that controls aspects of the whole sentence, such as curiosity, interest, and sense of doubt, which helps us uncover the underlying information hidden in sentences. The General Theory of Verbal Humor (GTVH) [16] also suggests that language, narrative strategy and logical mechanism are constitutive humor knowledge resources. Based on the above ideas, we try to automatically introduce commonsense knowledge for humorous statements in the humor detection model to assist the model understanding.

Therefore, we proposes a humor detection network integrating Semantic Incongruity and Commonsense Knowledge——SICKNet. We design a vivid example to demonstrate the expected SICKNet process of understanding humor, shown in Fig. 1. This model needs to do two things: On the one hand, to explore the semantic differences between *set-up* and *punchline* to fit the humor incongruity theory. On the other hand, to fully understand the meaning of each part of the joke with the introduction of commonsense knowledge, and to model fine-grained features such as contextual features, syntactic features and inter-sentence relations to detect the strength of humorous features. Our major contributions are as follows:

- To the best of our knowledge, we are the first work to exploit the underlying commonsense knowledge in humorous sentences to improve humor detection performance.
- We propose a novel model SICKNet for humor level detection, based on a multi-task optimization objective approach.
- Our experiments show that SICKNet can achieve the best results. Specifically, Accuary are 2.09% and 1.58% higher than the current state-of-the-art model on the Reddit dataset and TaivopJokes dataset, respectively.

II. RELATED WORK

In this section, we describe related work in terms of humor theory and humor detection and its methods.

A. Humor Theory

Aristotle, who was the first known to reflect on the incongruity of humor more than 2000 years ago, believed that humor originated from a mixture of two different interpretation frameworks of the same statement (humor emerges from difference between immediate expectation and the actual result). Schopenhauer [11] also advocates incongruity theory. As the theory is widely studied, it lays the foundation for the study of humor. In addition, Raskin established the Semantic Script-based Theory of Humor [17] (SSTH) based on the incongruity theory, which holds that a sentence can only consistute a joke when satisfies the following two conditions: i) The text is compatible, fully or in part, with two different semantic scripts. ii) Two scripts compatible with this text are semantically opposite. Subsequently, the General Theory of Verbal Humor (GTVH) extended SSTH by adding other possible humorous statement references such as language, narrative strategy, goal, situation, and logical mechanisms. Based on the incongruity theory, Paulos et al. [10]divided humor into *set-up* and *punchline*, arguing that the two parts are the unity of opposites.

B. Humor Detection and Its Methods

Humor detection is to determine whether a sentence is humorous or to predict the humor level of a joke. There are many excellent approaches in the field of humor detection, which can be grouped into two categories: feature engineering and deep learning. Mihalcea and Strapparava [18], who proposed the One-Liner dataset, used a heuristic based method on humor-specific stylistic features (rhymes, antonyms and slang). Mihalcea et al. [14] focused on the incongruity theory to address a four-category-based humor detection problem by exploring knowledge-based and corpus-based semantic relations between *set-up* and *punchline*. Cattle and Ma [19] proposed a word association-based humor classification system. With the popularity of deep learning in recent years, many methods stand out in the field of humor detection. Yang et al. [15] proposed to represent a combination of four potential structures of humor and KNN features as a person-centered feature approach. Bertero and Fung [20] made the first attempt to use an LSTM-based framework to predict humor in conversations.

Chen and Soo [21] used CNN structure with highway network. In addition, there are some approaches based on Transformer architecture. Weller et al. [13] used the BERT model to detect humor level, and illustrates the effectiveness of applying pre-trained models to humor detection. Sun et al. [3] proposed to construct three information extractors based on RoBERTa to extract textual information, superiority and incongruity information, and humor anchor information respectively. Xie et al. [5] applied the GPT-2 model to evaluate Uncertainty and Surprisal based on the incongruity theory.

We are one of the few efforts to link humor theory and large-scale pre-trained models based on structural divisions of humor. The closest to our work is Xie et al. [5], but we have done a lot of different work. We use a multi-task model to optimize different objective functions separately, with the aim of exploring the incongruity of *set-up* and *punchline* in humorous statements, and introducing commonsense knowledge of humorous statements to assist the model in understanding interactions at a fine-grained feature level.

III. METHODOLOGY

This section will describe the detail of our proposed SICKNet.

A. Outline of SICKNet

The structure of our SICKNet, which consists of four main components: feature extractor, SID module (Semantic Incongruity Detection), CSK extractor (Commonsense Knowledge) and DSHF module (Detecting the Strength of Humorous Feature), as shown in Fig. 2.

First, a whole joke consists of the set-up and the punchline, which we call $u_{orginal}$ and $v_{orginal}$, respectively, and call the whole joke $w_{orginal}$. SICKNet use RoBERTa [4] as a feature extractor[2] to encode $u_{orginal}$, $v_{orginal}$, and $w_{orginal}$ sequences and then to obtain 768-dimensional sentence vectors u,v and w by mean pooling method. In order to learning better effective linguistic feature in pre-trained language models, inspired by the work of Jawahar et al. [22], we freeze the low-level and mid-level parameters of RoBERTa, and only allow the deep-level parameters to participate in the update. The process of feature extractor is defined as:

$$[u_t, v_t, w_t] = RoBERTa([u_{orginal}, v_{orginal}, w_{orginal}]) \quad (1)$$

$$[u, v, w] = MeanPooling([u_t, v_t, w_t]) \quad (2)$$

Next, we design two core modules of the model. First, we use the SID module to detect the degree of incongruity generated by *set-up* and *punchline* at the semantic level. Second, a DSHF module is introduced to quantificate the humorous feature strength of the sentence. What is more, we use the CSK extractor to assist the DSHF module to integrate more knowledge. Notably, considering the similarities between the SID and DSHF modules, we correlated these two modules by a multi-task learning design. In the following, we describe the details of the SID module, the CSK extractor, the DSHF module and the loss optimization strategy respectively.

[2]The reason for adopting RoBERTa is that it can achieve extremely advanced performance in the task of understanding contextual information.

B. Semantic Incongruity Detection (SID)

A text joke can generally be divided into *set-up* and *punchline* [17]. If the joke is funny, there might be a semantic incongruity between these two parts, and this contrast relationship is less strong while it is not funny. There are many types of humor, such as irony, wordplay, metaphor, and satire [23], and their textual representations are often semantically inconsistent. Yang et al. [15] revealed several underlying semantic structures behind humor, including semantic incongruity, ambiguity, semantic style, and personal emotion, and their experiments showed that semantic incongruity is more effective in overall humor detection. Kulka [2] research also shown that semantic incongruity is a necessary condition for humor generation.

Therefore, exploring the semantics of *set-up* and *punchline* separately and comprehensively considering the difference semantic information together, could helps us to predict the humor level.

Specifically, we calculate the vector of $|u - v|$ within u and v that obtained by RoBERTa, and then concatenate u, v, and $|u - v|$ into a vector of length 768×3. We then input this vector into a one-layer feed-forward network to extract a 768-dimensional vector Y that focuses on the semantic incongruity relationship. Finally, a linear layer and a softmax layer are introduced to get the probability distribution of the degree of humor incongruity(P_{SID}). The process of SID module could denoted as:

$$Y = FFN(concat(u, v, |u - v|)) \quad (3)$$

$$P_{SID} = softmax(W_s Y + b_s) \quad (4)$$

where FFN is a one-layer feed-forward network, W_s and b_s are the parameters that the linear layer needs to be trained.

In addition, we apply standard cross-entropy loss function and the loss L_{SID} is calculated as follows:

$$L_{SID} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log P_{SID,i} + (1 - y_i) \log(1 - P_{SID,i}) \quad (5)$$

where N is the number of samples, y_i is the expected class label of the i sample, $P_{SID,i}$ is the probability that sample i is predicted to be positive for the SID task.

C. The external Commonsense Knowledge extractor (CSK)

We use the commonsense transformer model COMET [24] and GRU model [25] to extract the commonsense features in the CSK module.

COMET is an encoder-decoder model that training in the pretrained autoregressive language model GPT. The COMET model obtains a triple from the knowledge graph and is training on ATOMIC [26]. Since ATOMIC is composed of nine different "if-then relationship" types to distinguish agents

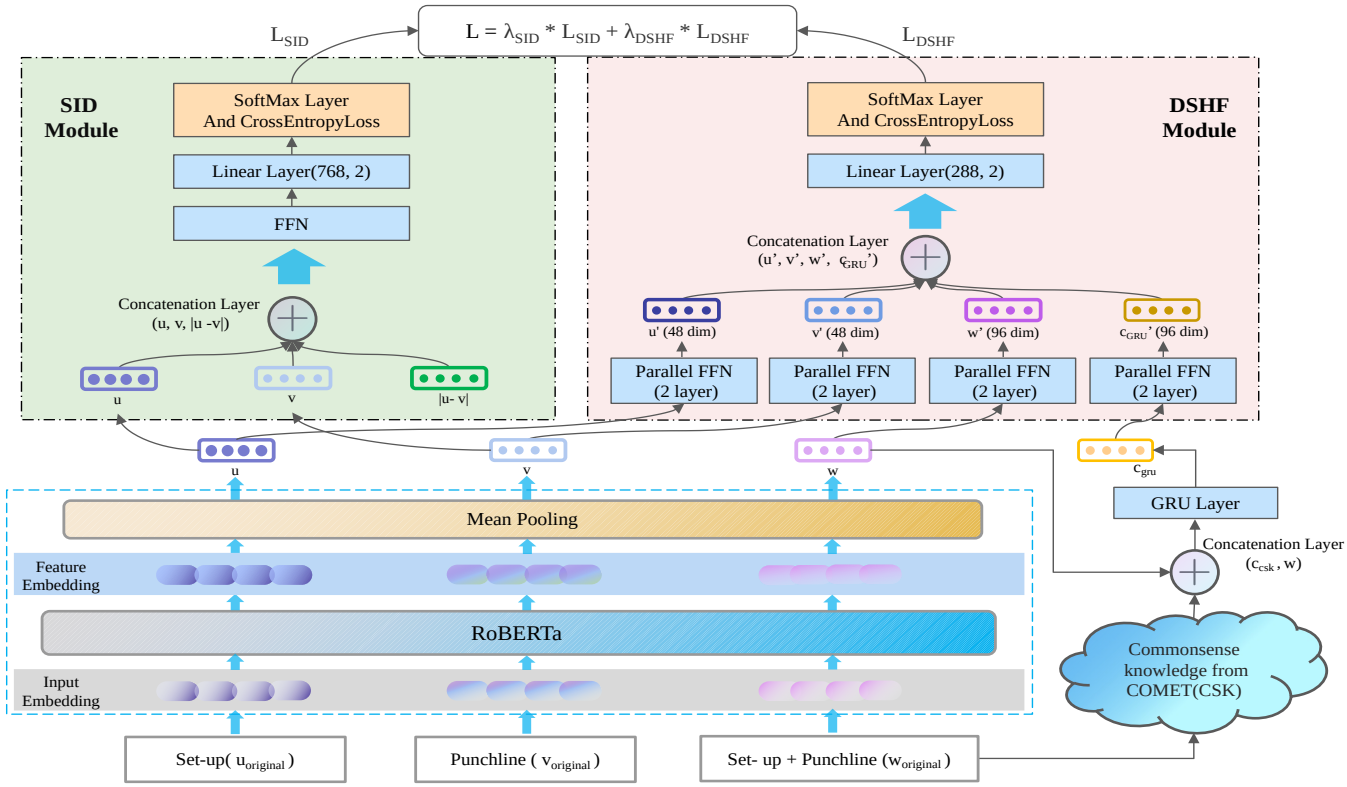

Fig. 2. IIustration of multi-task based SICKNet framework. SID: Semantic Incongruity Detection module. DSHF: Detecting the Strength of Humorous Features module.

/ themes, causes / effects, voluntary / non-voluntary events, and actions / mental states, COMET could learn more Commonsense Knowledge. Given an event in which X participates, the nine relation types (r) are inferred in Table. I.

In general, COMET is a generative model that produces a discrete sequence of knowledge conditional on s and r. However, our model needs to use a continuous vector of commonsense knowledge representations. To adapt our humor detection task, we use the COMET model pretrained on ATOMIC and then remove its decoding module.

we take the sentence that need to obtain commonsense knowledge as s and connect it with the relationship r to form $\{s + r\}$ to obtain the activation value of the the COMET encoder. However, we could not use all of the relationships in COMET, we then choose the most related relationship r : *(intent of X)*, since intent is not only a mental state, but also could use to determining the emotional dynamics of the sentence. The process of commonsense knowledge extraction is defined as:

$$c_{csk} = CSK(w_{original}) \tag{6}$$

$$c_{temp} = concat(c_{csk}, w) \tag{7}$$

where CSK [27] is Commonsense knowledge from COMET, $w_{original}$ is the whole joke, w is the feature vector encoded by RoBERTa.

In our framework, the final commonsense modeling is performed using GRU:

$$c_{gru} = GRU(c_{temp}) \tag{8}$$

where c_{temp} is a vector that concatenation commonsense knowledge vectors and w vectors, c_{gru} is the vector after the original sentence interacts with commonsense.

Finally, the commonsense vector c_{gru} will be introduced into the DSHF module to empower its external knowledge and improve its humor feature.

TABLE I
NINE RELATIONSHIP TYPES PREDICTED BY COMET

subject phrase (s)	relation phrase (r)	result
PersonX puts their arms around PersonY	*intent of X*	to comfort Y
	need of X	to be close to Y
	attribute of X	caring
	effect on X	get happy
	wanted by X	to show love
	reaction of X	friendly
	effect on others	be concerned about
	wanted by others	to hug X
	reaction of others	feels loved

D. Detecting the Strength of Humorous Features (DSHF)

The DSHF module is constructed based on the following setting: i) Assuming that we read the *set-up* and *punchline* in the joke alone, it may not easy to feel the existence of humor [13]. ii) The degree to which a person can perceive humor depends on the underlying commonsense knowledge, such as cultural background and intellectual scenario. Therefore, we not only consider modeling the *set-up* and *punchline* of jokes separately, but also introduce commonsense knowledge of humorous utterances to help the model fully understand the *set-up* and *punchline*.

We connect the feature vectors u,v,w and the vector c_{gru} combining commonsense knowledge in parallel to the two-layer feed-forward neural network. The network is used to extract the contextual features of each sentence. In order to balance four vectors contributions, we rearrange the four feature vectors, reducing their dimensionality to obtain 48-dimensional vectors u', v' and 96-dimensional vectors w', c'_{gru}. The four obtained vectors are concatenated together to obtain a vector Z that incorporates inter-sentence information and underlying features and it given as input to the linear layer with softmax to output the probability distributions of the strength of humorous features (P_{DSHF}), which has the form:

$$[u', v', w', c'_{gru}] = ParallelFNN([u, v, w, c_{gru}]) \quad (9)$$

$$Z = concat(u', v', w', c'_{gru}) \quad (10)$$

$$P_{DSHF} = softmax(W_d Z + b_d) \quad (11)$$

where $ParallelFNN$ are 4 two-layer feed-forward neural networks, W_d and b_d are the parameters that the linear layer needs to be trained.

The loss function of DSHF module is calculated as follows:

$$L_{DSHF} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log P_{DSHF,i} + (1-y_i)\log(1 - P_{DSHF,i}) \quad (12)$$

where $P_{DSHF,i}$ is the probability that sample i is predicted to be positive for the DSHF task.

E. Loss Optimization Strategy

We use the multi-objective loss optimization strategy mentioned by Liu et al. [28] for SICKNet in the training process. The loss $L(t)$ generated at time t is a weighted sum of the loss $L_{SID}(t)$ and loss $L_{DSHF}(t)$ generated by the two tasks, respectively. $L(t)$ is defined as:

$$L(t) = \lambda_1(t) \times L_{SID}(t) + \lambda_2(t) \times L_{DSHF}(t) \quad (13)$$

where $\lambda_k(t)$ is the weighting for task k, and is defined as follows:

$$\lambda_k(t) = \frac{K \times exp(r_k(t-1)/T)}{\sum_1^i exp(r_k(t-1)/T)} \quad (14)$$

$$r_k(t-1) = min(\frac{L_k(t-1)}{L_k(t-2)}, 10) \quad (15)$$

Here, $r_k(\cdot)$ calculates the relative descending rate in the range $(0, +\infty)$, t is an iteration index, and T represents a temperature which controls the softness of task weighting, similar to [29]. T has the effect of smoothing task weights. A smaller T results in a more discrete the distribution of weights of different tasks. Finally, the softmax operator, which is multiplied by K, ensures that $\sum_k \lambda_k(t) = K$.

IV. EXPERIMENTS

In this section, we describe the important parts of our experiments, including the datasets, baselines, experimental details, performance, and exploations.

A. Dataset

Our focus is on measuring the level of humor. To fairly assess performance of SICKNet, we need a dataset that includes both *set-up* and *punchline*, which can be divided into strong and weak humor. Therefore, we use the following two publicly available datasets: Reddit [13] and TaivopJokes [30]. The sample of the two datasets are in Table. II.

Reddit. The author have open sourced the code and data on github. The humorous statements in this dataset are text from the Reddit website with the "humor" tag, where a humorous statement includes the *set-up*, *punchline*, and likes. We chose this dataset to show how SICKNet performs compared to the previous state-of-the-art model on this dataset.

TaivopJokes. We find that the Reddit dataset had a wide range of sequence lengths with large variance. For example, the number of tokens for a sequence in Reddit is between 17 and 2870. This phenomenon leads to some problems, perhaps the model does not explore the meaning of humor but simply predicts based on the length of the sentence. Therefore, we also collected another dataset [30] consisting of two sentences with more quantity and more reasonable distribution of sequence lengths. We use r\Jokes part, in which each sample of data contains three important elements: *set-up*, *punchline* and score. Our processing steps for this dataset are as follows:

i) We analyze the percentile of the score in this dataset and categorize jokes with a score greater than 100 as strong humor and jokes with a score less than or equal to 10 as weak humor. We consider the score to represent a quantitative indicator of the level of humor in each joke as determined by user feedback (we recognize that everyone has a different opinion of the level of humor, but the score is somewhat reflective of how a particular community agrees with the humorous text).

ii) We delete the '\t' and '\n' characters in the *punchline* sequence and keep the word length greater than 0 and less than or equal to 200. We sample a total of 31,802 samples and balanced strong and weak humor types of jokes.

iii) The sampled 31,802 sample are equally distributed according to the humor level and split into three datasets: a training set containing 28622 sample and a validating and testing set containing 1590 sample.

TABLE II
EXAMPLE OF TWO DATASETS

datasets	Body	*punchline*	score
Reddit	Man, I was so tired last nigth; I was a muffler...	and I woke up exhausted	276
	I told my teenage niece to go get me a newspaper... She laughed at me, and said, "Oh uncle you're so old. Just use my phone."	So I slammed her phone against the wall to kill a spider.	28315
TaivopJokes	Why do a lot of math nerds wear glasses?	It helps with division.	473
	Math, please do me a favour...	I'm tired of solving you, solve yourself.	0

B. Baselines

To demonstrate the effectiveness of our model, we set up a number of baseline models as experimental comparisons for each of the two datasets.

Baselines for Reddit dataset. The experimental result for the following baselines are from Weller et al. [13] : Human and Transformer. The experimental result for the following baseline are from Zhang et al. [12]: TextCNN, LSTM, BiLSTM-Attention and MSIN(previously state-of-the-art model)

Baselines for TaivopJokes dataset. Experimental result for the following baselines are from our work: TextCNN [31], LSTM [8], BiLSTM-Attention [32], BERT [33], RoBERTa [4].

C. Experimental Details

We use the huggingface [34] and pytorch [35] tools to build the network structure of SICKNet. Specifically, we use hugingface's publicly available roberta-base model (with a 12-layer Encoder block) as the feature extraction model and freeze the low-level and mid-level features. In addition, we choose AdamW(weight_decay=0.01, eps=1e-8) [36] as our optimizer, with a linear update of the learning rate using warmup. We set different hyperparameters for each of the two datasets due to their different data distributions and quantities. For the Reddit dataset, the peak learning rate is 1e-5 and the maximum acceptable number of word tokens for the model is 512. For the TaivopJokes dataset, these two values are 3e-5 and 256, respectively. We train on a single Tesla P100 GPU with a batch size of 16 and train our models for a maximum of 10 epochs. Checkpoints are created along the way, saving the best model on the validating set for testing.

D. Experimental Performance

We report the results of the baselines and SICKNet experiments on the Reddit and TaivopJokes datasets in Table. III. For Reddit dataset, we observe that the SICKNet has an accuracy of 76.27%, which is 9.97% higher than the accuracy of humans on this dataset. In addition, Weller et al. [13], who presented the dataset, used the Transformer model with an accuracy of only 72.40%, which is 3.87% lower than our model. Most notably, SICKNet is 2.09% higher in accuracy than state-of-the-art model MSIN in this dataset.

Next, we observe the experimental data on TaivopJokes. Due to the window size constraints of TextCNN Convolution and the forgettable nature of implicit variables in LSTM, it is difficult to effectively capture semantic relationships over long distances, which is detrimental to humor level detection tasks. The BiLSTM-Attention model outperforms TextCNN and LSTM perhaps because of the bidirectional modeling of sequences and the focus on humor-friendly information, but still has limitations in detecting very long text sequences. In addition, the fine-tuned pre-trained BERT and RoBERTa have higher humor detection ability than the above baselines, but still cannot understand humor well. It is remarkable that SICKNet still shows the best results, with an Accuary of 73.64% and an F1 score of 76.26%, better than all other baselines.

SICKNet achieves this performance for the following reasons: i) Modeling a joke as the *set-up* and the *punchline* is more effective than modeling the whole joke to extract humor features. ii) We use a pre-trained model for feature extraction of long text sequences to avoid the disadvantage of not being able to capture long-range semantic relationships. iii) We introduce commonsense knowledge from COMET to complement the features of the original sentence well. iv) Using multi-task learning to share information and thus perform semantic incongruity extraction supported by commonsense.

TABLE III
THE RESULTS OF DIFFERENT METHODS ON TWO DATASETS

Method	Reddit		TaivopJokes	
	Accuary (%)	F1 (%)	Accuary (%)	F1 (%)
Human	66.30	-	-	-
TextCNN	69.42	69.82	70.08	69.87
LSTM	70.71	71.13	70.37	70.81
BiLSTM-Attention	70.97	71.85	71.49	70.89
Transformer(BERT)	72.40	-	71.57	73.35
RoBERTa	72.81	74.89	72.07	70.20
MSIN	74.18	74.22	-	-
SICKNet	**76.27**	**76.47**	**73.64**	**76.26**

E. Experimental Exploration

In order to explore the effectiveness of integrating the two modules in SICKNet and the effectiveness of designing RoBERTa as a feature extractor, we conduct two ablation experiments on the Reddit and TaivopJokes datasets, respectively. In addition, the stability of the model is confirmed from the statistical analysis of multiple sets of experiments.

The effectiveness of integrating SID and DSHF modules. The SICKNet as a whole is composed of the Semantic Incongruity Detection module and the Detecting the Strength of Humorous Features module, but we can still use one of these modules as a humor detection model. Therefore, we set up separate models for only the SID module and only the DSHF module. We conduct experiments on both datasets

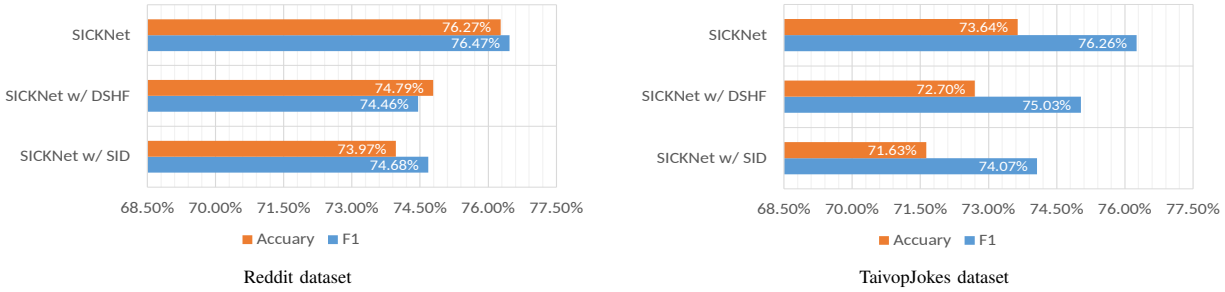

Fig. 3. Experimental results of SID and DSHF modules on two datasets

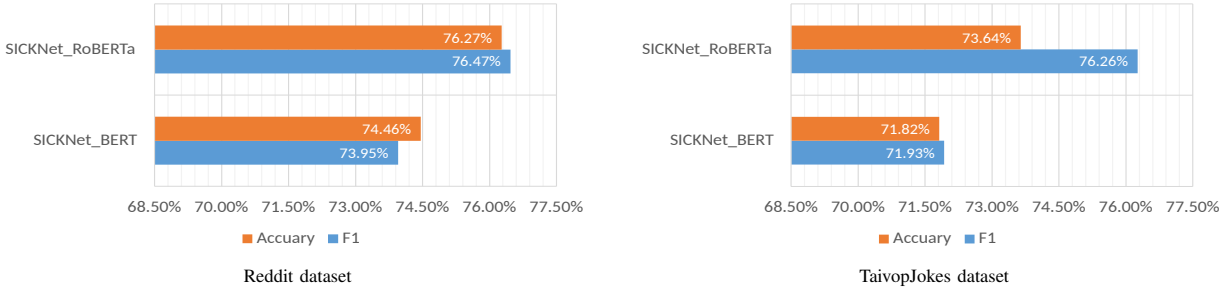

Fig. 4. Experimental results of SICKNet using different feature extractors on two datasets

separately and report in Fig. 3. It is worth noting that SICKNet only with DSHF module, which detects only the strength of humor features, is more effective on the Reddit dataset than the previous state-of-the-art model MSIN (Comparison with data in Table. III). This is because the commonsense knowledge added to this module complements the extensive contextual knowledge generated by the RoBERTa model and the module also fully take into account the humor structure, allowing for efficient extraction of humorous feature strengths. Of course, combining these two modules to construct the objective function respectively into a SICKNet based on multi-task learning has stronger stability and better effect in the humor level detection task.

The effectiveness of designing RoBERTa as a feature extractor. Experimental results for different feature extractors are shown in Fig. 4. We can find that the effect of using BERT as a feature extractor for SICKNet reaches an general level, while using RoBERTa can improve the performance of SICK-Net. The results show that in the relatively difficult humor detection task, the model needs to accept and understand more knowledge to extract higher quality sentence vectors, thereby improving the performance of humor detection(RoBERTa has 10 times more training data than BERT during pre-training). Obviously, the conclusion is that more knowledge can improve the performance of humor detection model. Notably, the BERT model is accuary of 72.4% [13] (data reported in Table. III) in the Reddit dataset, while SICKNet_BERT is accuary of 74.46%. The results further illustrate that the importance of our two modules for humor detection.

Experimental statistical analysis. In order to explore the stability and performance differences between the best baseline and SICKNet on the TaivopJokes dataset, we randomly select three random seeds to obtain the experimental results shown in Table. IV. We observe that the average of Accuary and F1 score produced by SICKNet in multiple experiments are higher than those of RoBERTa and the variance of SICKNet on these two metrics is much lower than those of RoBERTa. It illustrates that the SICKNet is not only effective in detecting humor levels, but also has better stability and is less affected by the initialized model weights.

TABLE IV
EXPERIMENTAL STATISTICAL ANALYSIS FOR TAIVOPJOKES

Method	Indicators	Accuary (%)	F1 (%)
RoBERTa	Avg	70.47	72.37
	s^2	1.3984	2.7777
SICKNet	Avg	**73.41**	**75.45**
	s^2	**0.0324**	**0.6228**

V. CONCLUSION

In this work, we proposed SICKNet, a framework based on multi-task learning that includes SID module and DSHF module. On the one hand, SID module calculates the extent to which *punchline* breaks expectations of *set-up* by considering the difference information between *set-up* and *punchline* at the semantic level based on incongruity theory. On the other hand, DSHF module introduces commonsense representation to enhance the ability to understand humorous texts to detect

the strength of humorous features. SICKNet alleviates the problem of the weak link between humor theory and humor detection networks and distinction of contextual environments. Significantly, it achieves new state-of-the-art results for humor detection across two benchmark datasets. The limitation of our work is that the introduced commonsense knowledge comes from everyday commonsense reasoning, which is not optimal in extracting commonsense in the humor domain. In the future, we try to apply humor-related commonsense atlas to train a commonsense transformer to extract more reasonable commonsense.

ACKNOWLEDGMENT

This work is partly supported by National Natural Science Foundation of China (Grant No.61977018 and Grant No.U21B2009) and Science and Technology Projects of Guangdong Province, China (Grant No. 2016B010127001).

REFERENCES

[1] S. Castro, M. Cubero, D. Garat, and G. Moncecchi, "Is this a joke? detecting humor in spanish tweets," *ibero-american conference on artificial intelligence*, 2016.

[2] T. Kulka, "The incongruity of incongruity theories of humor," *Organon F*, 2007.

[3] Y. Sun, Y. Li, and T. Zhao, "The improved neural network model in humor detection with traditional humor theory," *international conference on communications*, 2021.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv: Computation and Language*, 2019.

[5] Y. Xie, J. Li, and P. Pu, "Uncertainty and surprisal jointly deliver the punchline: Exploiting incongruity-based features for humor recognition," *meeting of the association for computational linguistics*, 2021.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[7] Y. Ziser, E. Kravi, and D. Carmel, "Humor detection in product question answering systems," *international acm sigir conference on research and development in information retrieval*, 2020.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, 1997.

[9] Z. Yang, S. Hooshmand, and J. Hirschberg, "Choral: Collecting humor reaction labels from millions of social media users," 2022.

[10] J. A. Paulos, "Mathematics and humor," 1980.

[11] t. Haldane, R.B, t. Kemp, J., and A. Schopenhauer, "The world as will and idea - vol.2," *STATE CENTRAL LIBRARY, KOLKATA*, 1883.

[12] J. Zhang, S. Zhang, X. Fan, L. Yang, and H. Lin, "A multi-granularity semantic interaction understanding network for humor level recognition (in Chinese)," *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, 2020.

[13] O. Weller and K. D. Seppi, "Humor detection: A transformer gets the last laugh," *empirical methods in natural language processing*, 2019.

[14] R. Mihalcea, C. Strapparava, and S. Pulman, "Computational models for incongruity detection in humour," *international conference on computational linguistics*, 2010.

[15] D. Yang, A. Lavie, C. Dyer, and E. Hovy, "Humor recognition and humor anchor extraction," *empirical methods in natural language processing*, 2015.

[16] S. Attardo and V. Raskin, "Script theory revis(it)ed: joke similarity and joke representation model," *Humor: International Journal of Humor Research*, 1991.

[17] V. Raskin, "Semantic mechanisms of humor," 1984.

[18] R. Mihalcea and C. Strapparava, "Making computers laugh: Investigations in automatic humor recognition," *empirical methods in natural language processing*, 2005.

[19] A. Cattle and X. Ma, "Recognizing humour using word associations and humour anchor extraction," *international conference on computational linguistics*, 2018.

[20] D. Bertero and P. Fung, "A long short-term memory framework for predicting humor in dialogues," *north american chapter of the association for computational linguistics*, 2016.

[21] P.-Y. Chen and V.-W. Soo, "Humor recognition using deep learning," *north american chapter of the association for computational linguistics*, 2018.

[22] G. Jawahar, B. Sagot, and D. Seddah, "What does bert learn about the structure of language," *meeting of the association for computational linguistics*, 2019.

[23] Y. Raz, "Automatic humor classification on twitter," *north american chapter of the association for computational linguistics*, 2012.

[24] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "Comet: Commonsense transformers for automatic knowledge graph construction," *meeting of the association for computational linguistics*, 2019.

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

[26] M. Sap, R. LeBras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "Atomic: An atlas of machine commonsense for if-then reasoning," *national conference on artificial intelligence*, 2018.

[27] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "Cosmic: Commonsense knowledge for emotion identification in conversations," *empirical methods in natural language processing*, 2020.

[28] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention." *computer vision and pattern recognition*, 2018.

[29] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv: Machine Learning*, 2015.

[30] T. Pungas, "A dataset of english plaintext jokes." 2017. [Online]. Available: https://github.com/taivop/joke-dataset

[31] Y. Kim, "Convolutional neural networks for sentence classification," *empirical methods in natural language processing*, 2014.

[32] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hongwei, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," *meeting of the association for computational linguistics*, 2016.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *north american chapter of the association for computational linguistics*, 2018.

[34] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," *neural information processing systems*, 2019.

[36] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," 2018.

Appendix

Table.V extends the experiments on the TaivopJokes dataset in Section IV-E of this paper, showing the experimental reports of multiple models with three random seeds. We observe that SICKNet performs well in different random seeds, demonstrating the validity and stability of the model design.

TABLE V
EXPERIMENTAL RESULTS FOR THREE RANDOM SEEDS

Seed	Method	Accuary (%)	F1 (%)
33	RoBERTa	69.24	72.67
	SICKNet w/ SID	70.44	74.18
	SICKNet w/ DSHF	71.94	73.04
	SICKNet_BERT	71.76	73.54
	SICKNet	**73.39**	**74.38**
222	RoBERTa	72.07	70.20
	SICKNet w/ SID	71.63	74.07
	SICKNet w/ DSHF	72.70	75.03
	SICKNet_BERT	71.82	71.96
	SICKNet	**73.64**	**76.26**
2021	RoBERTa	70.12	74.25
	SICKNet w/ SID	72.76	73.90
	SICKNet w/ DSHF	73.01	74.17
	SICKNet_BERT	72.20	73.44
	SICKNet	**73.20**	**75.71**
Average result of the above 3 random seeds	RoBERTa	70.47	72.37
	SICKNet w/ SID	71.61	73.15
	SICKNet w/ DSHF	72.55	74.08
	SICKNet_BERT	71.92	72.98
	SICKNet	**73.41**	**75.45**

Table VI shows the results of SICKNet with different hyperparameters on the two datasets. We observe that SICKNet still performs well under different hyperparameters.

TABLE VI
EXPERIMENTAL RESULTS WITH DIFFERENT BATCH SIZES AND DIFFERENT LEARNING RATES ON TWO DATASETS

datasets	Parameters		Accuary (%)	F1 (%)
Reddit	batch_size	**16**	**76.27**	**76.47**
		32	74.63	75.94
		64	75.45	76.46
	learning_rate	9e-6	75.62	76.13
		1e-5	**76.27**	**76.47**
		2e-5	75.95	76.60
TaivopJokes	batch_size	**16**	**73.41**	**75.45**
		32	70.57	73.26
		64	72.01	72.48
	learning_rate	1e-5	71.20	69.22
		2e-5	71.58	69.99
		3e-5	**73.41**	**75.45**

Table. VII shows the Reddit and TaivopJokes datasets sliced into three subsets according to the equal distribution of humor levels, respectively.

TABLE VII
THE DETAILS OF THE DATASETS

datasets		Strong humor	Weak humor	total
Reddit	train	9719	9719	19438
	validation	304	304	608
	test	304	304	608
TaivopJokes	train	14311	14311	28622
	validation	795	795	1590
	test	795	795	1590