Leveraging Social Context for Humor Recognition and Sense of Humor Evaluation in Social Media with a New Chinese Humor Corpus - HumorWB

Zeyuan Zeng, Zefeng Li, Liang Yang, Dongyu Zhang, Hongfei Lin[*]

School of Computer Science and Technology, Dalian University of Technology, Dalian, China

{zengzeyuan, chinese_lzf}@mail.dlut.edu.cn

{liang, zhangdongyu, hflin}@dlut.edu.cn

Abstract

With the development of the Internet, social media has produced a large amount of user-generated data, which brings new challenges for humor computing. Traditional humor computing research mainly focuses on the content, while neglecting the information of interaction relationships in social media. In addition, both content and users are important in social media, while existing humor computing research mainly focuses on content rather than people. To address these problems, we model the information transfer and entity interactions in social media as a heterogeneous graph, and create the first dataset which introduces the social context information - HumorWB[1], which is collected from Chinese social media - Weibo. Two humor-related tasks are designed in the dataset. One is a content-oriented humor recognition task, and the other is a novel humor evaluation task. For the above tasks, we purpose a graph-based model called SCOG, which uses heterogeneous graph neural networks to optimize node representation for downstream tasks. Experimental results demonstrate the effectiveness of feature extraction and graph representation learning methods in the model, as well as the necessity of introducing social context information.

Keywords: Sentiment Analysis, Humor Recognition, Sense of Humor Evaluation

1. Introduction

As a human-specified communication method, humor plays an important role in our daily lives. Humorous can be used to describe that someone or someone's creation makes people smile or laugh. It's more like fun or interesting, but it also can be more complicated and undescribable, and more importantly, it belongs to human being, it is created by wisdom.

Humor computing research is dedicated to enabling computers to recognize and understand humor to better serve human society. However, most humor expressions in real life are informal and accompanied by considerable noise, which limits the applicability of traditional humor computing research based on static linguistic resources such as jokes and puns. In today's rapid development of information technology, social media has become an important medium for people to interact and communicate. As the content in social media is contributed spontaneously by a large number of users, it is closely related to daily life, large in scale, and widely spread, making it an ideal source for humor computing research.

In addition, non-standard language expressions and emojis are common in social media, which increases the difficulty of research. Effectively studying these expressions can contribute to achieving more advanced AI systems. At present, there have

文本：失眠就是，想假装睡着，但又被自己识破。

Text: Insomnia is the act of pretending to be asleep, but being perceived by oneself.

reposts_count: 253   comments_count: 322   attitudes_count: 1455

Figure 1: An Example of Humor Tweet on Weibo.

been research on dataset construction and humor recognition based on social media. However, most of the research only treats social media as a data source and collect the corpus from it. For example, Zhang and Liu (Zhang and Liu, 2014), and Castro et al. (Castro et al., 2016) constructed dataset from Twitter, but they only used the text of tweets and neglected retweets and comments. Besides, non-textual features (e.g., number of comments), as well as the social relationship information of the users were discarded.

There's an example of humor tweet on Weibo (which is one of the largest social media platforms in China, having 600 million monthly active users), as shown in Fig. 1. Except for text, we can find number of reports, number of comments and number of thumbs. In the meantime, behind these numbers, there are tons of relations between users to users, users to articles, users to comments and so on. All of these can be useful in humor-related research, for example:

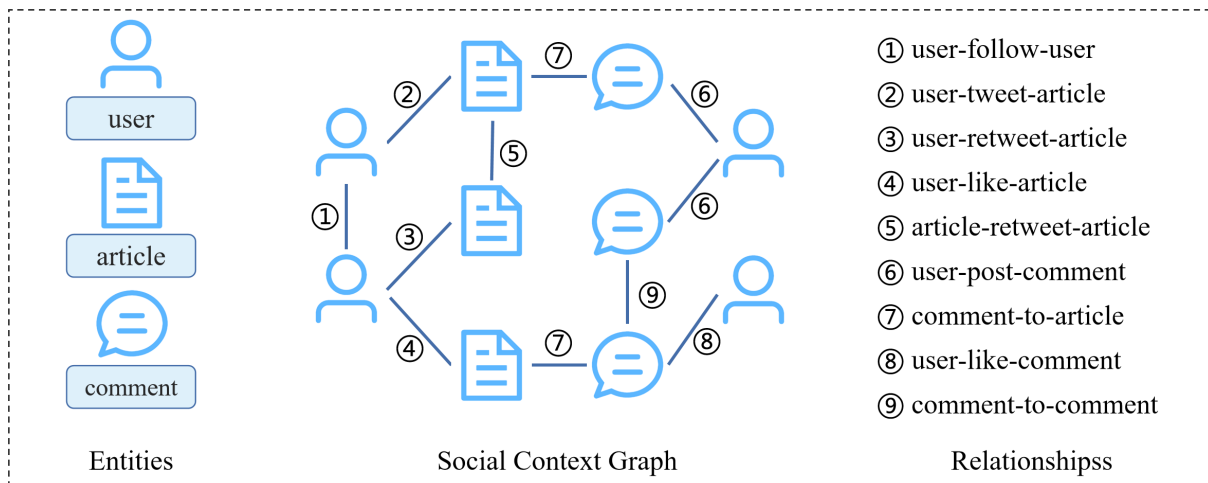- Funny content usually receives higher ratings

---

Figure 2: Entities and relationships in social media.

and likes. Weller and Seppi (Weller and Seppi, 2019) analyzed 16000 jokes collected from Reddit and found that only 6% of the jokes had likes between 200 and 20000. They then determined whether a joke was funny or not based on the cut-off of 200.

- There are certain differences between comments on humorous and non-humorous content, such as more positive and humorous. Chen et al. (Chen et al., 2021) collected a corpus of TED talks including speech and the audience's comments. By jointly modeling the content and the comments, they demonstrated that comments are essential in improving the recognition tasks.

- As content creators, users have tendencies towards the types of content they tweet. For example, entertainment bloggers tend to share news about celebrities, pet bloggers tend to share their or other people's pets, and humor bloggers tend to share jokes and memes.

- The dissemination property of information and the social property of social media lead to selective information dissemination, and humorous content is no exception. The theories of filter bubbles, echo chambers, and information cocoons are related to this phenomenon.

Therefore, it is highly feasible to introduce social context information from social media to assist in humor-related tasks. Although there is no research specifically on humor computing, social context has already been applied in various tasks such as fake news detection (Nguyen et al., 2020), sentiment polarity analysis and roll call prediction.

In addition, the social context contains information about users and their interaction relationships, which brings the possibility to study users' sense

of humor. Sense of humor is a personality trait that allows people to understand funny things, and is closely related to humor. Feingold and Mazzella (Feingold and Mazzella, 1993) considered humor as an ability to understand, transmit and create humor, which belongs to a cognitive ability. Humorous people are good at creating novel and interesting content through thinking, and control negative emotions with the help of humor. Therefore, research on individuals' sense of humor can be valuable in understanding the role and impact of humor in human society.

However, most existing research on humor computing focuses on objective things such as text and images. As previously mentioned, sense of humor plays an important role in human society, which deserves exploration. Unfortunately, the work of analyzing and evaluating sense of humor using artificial intelligence technology is still in a blank. And the primary method for evaluating sense of humor remains psychological scales, which are time-consuming, labor-intensive, and sample-limited. Social media provides abundant user-generated content and behavioral information, which can simulate the evaluation indicators in sense of humor scales to some extent. Therefore, it is feasible to conduct research on sense of humor evaluation through social media.

Therefore, we model the social context in social media as a heterogeneous graph, as shown in Fig. 2. The graph presents user, article and comment(nodes) with their interactions(edges), which contains a large amount of information. The experiments show the effectiveness.

In summary, the contributions of our work are as follows:

- We propose a graph representation called social context graph that models information and social entities, as well as their interactions.

10394

- We create the first dataset which introduces the social context information - HumorWB, which is collected from Chinese social media - Weibo. And we introduce two humor-related tasks, one of which is a novel sense of humor evaluation task and the other is a humor recognition task.

- We develop a model based on **S**ocial **CO**ntext **G**raph(**SCOG**) for the two tasks, and prove the effectiveness of our model through experiments

- We design experiments for three research questions to better understand our model for the humor recognition task.

2. Related Work

2.1. Humor Recognition

Humor recognition is an important part of humor computing and is often defined as a classification task. Artificial features were widely used in humor recognition research. Mihalcea and Strapparava (Mihalcea and Strapparava, 2005) proposed three humor-specific features including alliteration, antonymy, as well as adult slang. And they used Naive Bayes and SVM for classification. Barbieri and Saggion (Barbieri and Saggion, 2014) design several linguistic features to automatically detect irony and humor in twitter. Castro et al. (Castro et al., 2016) designed humor features based on Spanish tweets and compared the humor recognition performance of different classifiers. Liu et al. (Liu et al., 2018) combined semantic analysis and sentiment analysis for modeling sentiment-associated patterns, and use them for humor recognition.

With the development of deep learning, researchers started to apply deep learning methods to humor recognition tasks. Chen and Lee (Chen and Lee, 2017) designed a deep learning framework based on CNN for studying humor in speech and puns. They achieved better results than machine learning methods without using artificial humor features. Weller and Seppi (Weller and Seppi, 2019) applied Transformer to humor recognition tasks and found that Transformer performed better than other neural networks. Zhou et al. (Zhou et al., 2020) added pronunciation units to Transformer for better capturing implicit phonological properties, and experimental results proved that their method had significant advantages for the pun detection and localization tasks. In addition to using autoencoder language model, Xie et al. (Xie et al., 2021) used autoregressive language model for humor recognition. They captured humor by calculating the inconsistency scores between context and punchline using the GPT model.

2.2. Sense of Humor Evaluation

At present, research on sense of humor is primary in the psychology field. Feingold and Mazzella (Feingold and Mazzella, 1993) considered that the sense of humor is influenced by humor motivation and humor communication, while the production of humor is only influenced by humor cognition. Martin et al. (Martin et al., 2003) viewed humor as an individual's behavior pattern, worldview, and psychological characteristics in response to things. krish and Kuiper (Kirsh and Kuiper, 2003) considered humor in three dimensions: crude and obscene humor, social skills humor, and pretentious and demeaning humor, indicating that humor also may have negative effects.

The evaluation of sense of humor is mainly through psychological scales. The Multidimensional Sense of Humor Scale (MSHS) developed by Thorson and Powell (Thorson and Powell, 1993) is a representative work. The scale is a 5-point Likert-type scale which is composed of four factors: humor production, coping with humor, humor appreciation, and attitudes toward humor. Many researchers have utilized or adapted this scale in their research. For example, Dowling and Fain (Dowling and Fain, 1999) revised the MSHS to assess sense of humor in school-aged children.

Additionally, Ramsey and Meyer (Ramsey and Meyer, 2019) used the MSHS and three other existing scales to evaluate the criterion validity of their new scale that measures the humor purposes including identification, clarification, enforcement, and differentiation, sense of humor can even assist in psychological health detection.

Recently, Bielaniewicz et al.(Bielaniewicz et al., 2022) considered about the sense of humor, but they just took it for personalized annotations, they were not focusing on the sense of humor, it was just for the humor recognition task.

3. Methodology

In this section, we first introduce the modeling of social context graph, and then formally define two humor-related tasks: humor recognition and sense of humor evaluation. Finally, we describe our model structure in detail.

3.1. Graph Construction using Social Context

We model the social context in social media as a heterogeneous graph, as shown in Fig. 2. Among which nodes represent social and content entities, including articles, comments, and users; edges represent social relations and information flow, having nine types, like "user-follow-user", "user-tweet-article", "user-like-article" and so on.
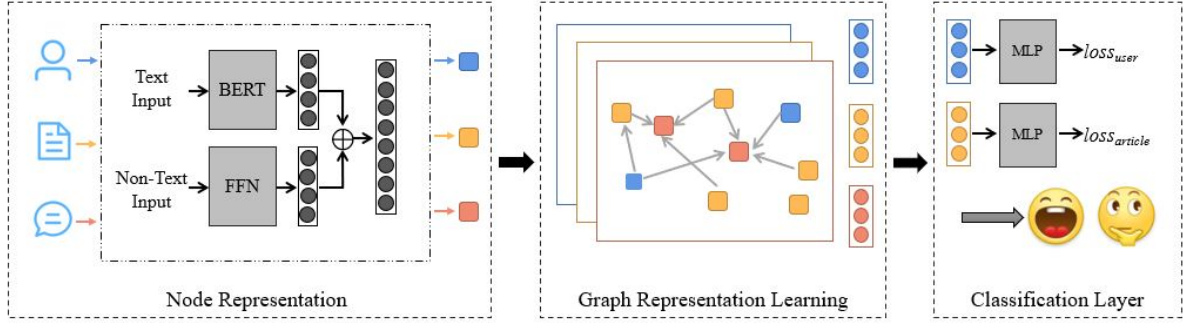
Figure 3: Overview of the model **SCOG**.

The social context graph G is a heterogeneous graph that includes entities that make up the basic elements of social media and their interaction relationships. It is defined as follows:

- $A = \{a_1, a_2, \cdots, a_m\}$ is a collection of articles, and each article $a_i \in A$ has its own initial information, including the text of article s_i^a and other non-textual information g_i^a.

- $C = \{c_1, c_2, \cdots, c_n\}$ is a collection of comments, and each comment $c_j \in C$ has its own initial information, including the text of comment s_j^c and other non-textual information g_j^c.

- $U = \{u_1, u_2, \cdots, u_p\}$ is a collection of users, and each user $u_k \in U$ has its own initial information, including the description of user s_k^u and other non-textual information g_k^u (the user name is not used).

- $E = \{e_1, e_2, \cdots, e_q\}$ is a collection of entity interaction relationships, and each entity interaction relationship $e_q \in E$ needs to be jointly represented by two entities $v_1, v_2 \in A \cap C \cap U$ and their relationship R_e shown in Fig. 2.

The non-textual information can be number of reports, number of comments, number of thumbs, number of followers and so on. And the text information for user is the description in the user's profile.

3.2. Task Definition

Given a social context graph $G = \{A, C, U, E\}$ constructed from articles A, comments C, users U, and entity interaction relationships E, humor recognition based on social context is defined as a binary classification task to predict whether an article $a \in A$ is humorous or not. And sense of humor evaluation based on social context is defined as a triple classification task to predict whether a user $u \in U$ has low, average or high sense of humor.

3.3. Model

Our model SCOG focuses on leveraging social context to optimize node representation. Fig. 3 shows the overview of our model SCOG, which consists of three components. The node representation layer embeds the original inputs of each node. The graph representation learning layer uses heterogeneous graph neural networks to enhance node representation for downstream tasks. And the classification layer provides outputs corresponding to the classes and computes the loss. The specific details of each layer are as follows:

3.3.1. Node Representation

Although articles, users, and comments have different initial information, they all contain textual and non-textual input. Therefore, they share similar feature extraction process. For the textual input $s_x = \{w_1, w_2, \cdots, w_n\}$ of each node $x \in \{a, c, u\}$, textual representation z_x is obtained by a text encoder like BERT (Kenton and Toutanova, 2019):

$$z_x = Encoder\left(s_x\right) \tag{1}$$

And non-textual input g_x is normalized and embedded by a single-layer feedforward neural network:

$$z_x^{'} = W_x \cdot \sigma\left(g_x\right) + b_x \tag{2}$$

where $\sigma\left(\cdot\right)$ denotes a normalized function, W_x and b_x denote trainable parameters.

Finally, textual feature z_x and non-textual feature $z_x^{'}$ are concatenated as the node representation d_x:

$$d_x = Concat\left(z_x, z_x^{'}\right) \tag{3}$$

3.3.2. Graph Representation Learning

The purpose of graph representation learning is to enhance node representation by utilizing graph structure. We implement this process using GNN,
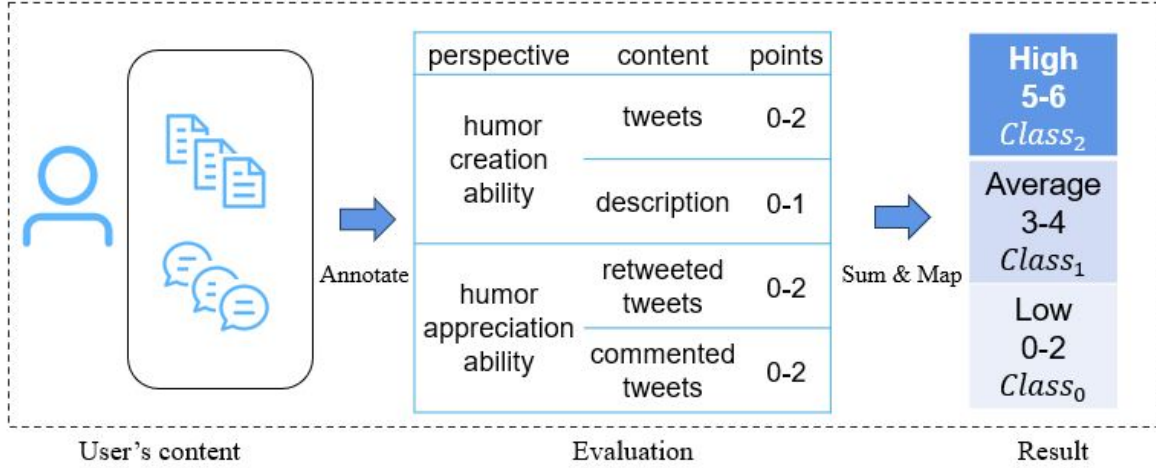
Figure 4: Annotation process of users' sense of humor.

the core of which is the message passing between nodes. With $\boldsymbol{h}_i^k$ denoting node features of node i in layer k, message passing process can be described as:

$$\boldsymbol{h}_i^k = \gamma^k \left(\boldsymbol{h}_i^{k-1}, \bigoplus_{r \in \mathcal{R}} \bigoplus_{j \in \mathcal{N}_i^r} \phi_r^k \left(\boldsymbol{h}_i^{k-1}, \boldsymbol{h}_j^{k-1} \right) \right) \quad (4)$$

where ϕ^k is a message passing function to calculate the message from neighbor $j \in \mathcal{N}_i$ of node i, such as MLPs. $\oplus$ denotes an aggregate function which is differentiable, permutation invariant such as mean, max and sum. γ^k is an update function to update self information, which can also be MLPs.

In heterogeneous graphs, node i can establish neighbor relationship $r \in \mathcal{R}$ with different types of nodes through direct connection or meta-path. These messages are considered by the outer aggregation function, which first aggregates messages from neighbors under the same type of relationship. And then re-aggregate the aggregated messages of different relationships.

3.3.3. Classification Layer

Both humor recognition task and sense of humor evaluation task are node classification tasks, so we connect the output $\boldsymbol{h}_x^l$ of node $x \in \{a, c, u\}$ with a single-layer feedforward neural network and the Softmax function to obtain the label probability distribution $\boldsymbol{p}_x = \{p_{x_0}, p_{x_1}, \cdots, p_{x_n}\}$:

$$\boldsymbol{p}_x = Softmax \left(\boldsymbol{W}_x \boldsymbol{h}_x^l + \boldsymbol{b}_x \right) \quad (5)$$

And we use cross entropy loss function to optimize the model:

$$\mathcal{L}_x = -\frac{1}{N_x} \sum_i^{N_x} \sum_c^C y_{i,c} \log \left(p_{i,c} \right) \quad (6)$$

where N_x denotes the number of training samples, C denotes the collection of labels and $y_{x,c}$ denotes the ground truth.

4. Dataset

4.1. Data Collection

During the dataset construction process, our focus is on retaining the structure of social context graph while balancing the amount of humorous and non-humorous content. Weibo is chosen as the data source, with the crawling process centered around users. A series of humor bloggers and control users such as actors are selected as seed users. And then (a) tweets of users, (b) comments and retweets of tweets, and (c) following relationships of users are crawled iteratively.

The crawled tweets are then coarsely filtered to determine whether the text length is appropriate, and the comments, retweets, and likes of these tweets are crawled. Then we count the top users in the comments, retweets and likes, and repeat the process.

4.2. Data Annotation

There are two annotation tasks. One is to annotate binary labels of humor for the articles, which is relatively straightforward, as annotators only need to judge whether an article is humorous or not. The other involves annotating users' sense of humor, for which we refer to relevant psychological research.

In brief, users' sense of humor are evaluated from two perspectives: humor creation ability and humor appreciation ability, according to the MSHS scale. Specifically, for humor creation ability, we rate users' sense of humor in terms of the humor

10397

Model	Number of articles			Average text length			Number of users			
	All	$Class_0$	$Class_1$	All	$Class_0$	$Class_1$	All	$Class_0$	$Class_1$	$Class_2$
All	5291	3281	2010	42.34	41.48	43.75	2151	895	594	662
Train	3703	2287	1416	42.56	41.65	44.04	1504	626	415	463
Valid	794	499	295	41.99	41.31	43.12	322	134	89	99
Test	794	495	299	41.68	40.86	43.03	325	135	90	100

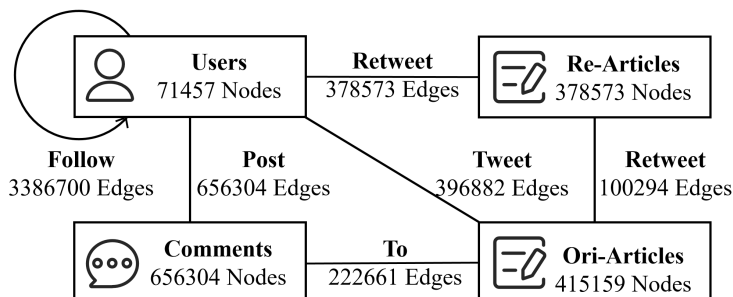Table 1: Labeled Data distribution of the dataset.

Figure 5: Graph structure of the whole dataset with statistics(Labeled and Unlabeled).

level of tweets (0-2 points) and whether their personal descriptions are humorous (0-1 points). For humor appreciation ability, we rate the humor level of retweeted tweets (0-2 points) and commented tweets (0-2 points). Afterwards, we sum all the scores and map the total scores to the corresponding labels: low (0-2 points), average (3-4 points) and high (5-6 points) sense of humor. The process is shown in Fig. 4.

The whole annotation process was done by three authors. Each label was decided by all three persons, and we followed the principle of majority in the whole process. The raw data is large, and we labled a few(5291 articles and 2151 users). The rest can be used to construct the graph, and for further study.

4.3. Data Analysis

We analyzed the annotated data and graph information of the dataset. Table 1 shows the distribution of classes, and the average character length of articles. It reveals that classes are slightly unbalanced and the length of humorous text is slightly greater. Fig. 5 illustrates the graph structure and the number of nodes and edges based on the whole dataset, including labeled data and unlabeled data. In addition, articles are divided into original and retweeted ones for a more intuitive visualization.

5. Experiment

5.1. Experimental Setup

In this section, we evaluate our model on two challenging tasks: humor recognition and sense of humor evaluation. Our model use BERT-Base for text

Model	Acc	P	R	F1
LR	76.83	71.86	63.21	67.26
SVM	77.71	72.93	64.88	68.67
TextCNN	78.97	71.15	74.25	72.67
BiLSTM	76.45	65.73	78.26	71.45
BERT	82.87	85.28	65.89	74.34
RoBERTa	82.75	**85.53**	65.22	74.00
SCOG(Ours)	**83.00**	76.11	**79.93**	**77.98**

Table 2: Performance of our model and baselines on the humor recognition task.

representation and GraphSage (Hamilton et al., 2017) for graph representation learning unless otherwise stated. The baseline models include machine learning models (SVM and LR) and deep learning models (TextCNN (Chen, 2015), BiLSTM (Zhou et al., 2016), BERT and RoBERTa (Liu et al., 2019)). Except for BERT and RoBERTa, baseline models use GloVe (Pennington et al., 2014) as the word embedding methods.

5.2. Humor Recognition Results

For better discussion, our humor recognition experiment is conducted on a subgraph. Specifically, articles need have labels and comments need orient to these articles. The experimental results are shown in Table 2. When using the same text feature input, TextCNN (Chen, 2015) and BiLSTM perform better than machine SVM and LR, indicating that deep neural networks can handle features more efficiently. Fine-tuning BERT and RoBERTa result in higher accuracy and F1 than TextCNN and BiLSTM, indicating that contextualized word embedding can capture more effective semantic features

10398

Model	Performance under different data completeness (Acc/Macro-F1)			
	100%	90%	70%	50%
$\text{SVM}_{\text{Non-Text}}$	44.00/27.57	-	-	-
SVM_{Text}	47.69/39.49	46.37/38.03	45.82/37.45	44.80/34.85
BERT	52.62/45.86	51.38/45.06	48.92/39.29	46.15/36.98
RoBERTa	53.85/48.90	52.62/47.15	50.46/40.39	45.85/35.89
SCOG(Ours)	**65.85/62.71**	**65.54/62.20**	**64.31/60.83**	**62.77/58.86**

Table 3: Performance of our model and baselines on the sense of humor evaluation task under different data completeness. Here, "Non-Text" denotes using non-textual features, while "Text" denotes using textual features.

Method	GraphSage				HAN			
	Acc	P	R	F1	Acc	P	R	F1
Random	69.43	57.99	51.85	54.75	70.19	60.54	47.09	52.98
GloVe	80.48	74.00	74.25	74.12	80.98	75.69	72.91	74.28
BERT	**83.00**	76.11	**79.93**	**77.98**	81.74	76.01	**75.25**	75.63
RoBERTa	82.49	76.32	77.59	76.95	81.86	77.00	73.91	75.43
ESimCSE	**83.00**	**77.89**	76.59	77.23	**82.24**	**77.24**	74.92	**76.06**

Table 4: Comparison between text representation methods.

of humor. Although our model uses BERT as the text encoder, the results are better than fine-tuning the BERT model, indicating that graph representation learning can effectively enhance the quality of node representation related to humor semantic.

5.3. Sense of Humor Evaluation Results

Since sense of humor evaluation requires all the information in the social context, we conduct experiments on the complete graph. However, due to the memory limitation, we train the BERT model using the articles to reduce the user's text representation to 100 dimensions, in the meantime, we set a parameter called data completeness, as the percentage of data we used in the graph-constructing. The experimental results are shown in Table 3, including the performance of our model and baselines on the sense of humor evaluation task under different data completeness. It can be seen that as the data increases, which means that the graph is denser, the performance of the model improves. SVM can barely classify non-textual features, suggesting that they can not reflect users' sense of humor. In the case of using the personal description, fine-tuned BERT and RoBERTa outperform SVM, but both of them are not satisfactory, showing the limitation of evaluating a user's sense of humor by personal description only. Our model SCOG has the best performance, indicating that it can effectively utilize the information of each node to evaluate the sense of humor of user nodes, also the importance of social context graph constructing.

6. Discussion

In this section, we answer the following research questions to better understand our model under the humor recognition task:

- **RQ1**: What is a more efficient way to represent text?

- **RQ2**: Which GNN models perform better in graph representation learning? What is the appropriate setting for the hidden dimension of GNNs?

- **RQ3**: What is the contribution of each module in our model SCOG?

6.1. Text Representation (RQ1)

We compare various text representation methods, including static word embedding model (GloVe), pre-trained language models (BERT and RoBERTa) and representation learning method (ESimCSE (Wu et al., 2022)). As the experimental results shown in Table 4, all methods perform better than using the random feature. BERT and RoBERTa outperform GloVe, indicating that Transformer has stronger semantic representation ability, while GloVe also achieves good results using less computing resource. RoBERTa is slightly weaker than BERT, suggesting that they are applicable to different tasks. In addition, ESimCSE has no performance advantage compared with BERT in our experiments.

6.2. Graph Neural Networks (RQ2)

Our model is compatible with different graph representation learning methods. To determine which

GNN Model	F1 at different hidden dimensions of GNN					
	50+	100	150+	200	250+	300
RGCN	75.09	76.11	76.10	**76.90**	76.87	76.62
GraphSage	76.17	77.30	77.32	**77.98**	77.18	77.31
HAN	74.96	75.67	75.50	75.63	75.68	**75.83**
HGT	75.43	75.77	76.14	75.90	75.73	**76.52**

Table 5: Comparison between GNN models with different hidden dimensions. "+" denotes "+2" for HAN and HGT as dimensions need to be divisible by heads.

Model	Acc	P	R	F1
w/o text features	67.13	58.72	42.81	49.52
w/o non-text features	82.37	75.73	78.26	76.97
w/o GNN	78.09	71.93	68.56	70.21
w/o user nodes	82.37	77.13	75.59	76.35
w/o comment nodes	81.61	75.76	75.25	75.50
SCOG(Ours)	**83.00**	**76.11**	**79.93**	**77.98**

Table 6: Results of ablation experiments.

id	text	label	prediction
1	请推荐一部你最近看了不错的电影给大家~ *Please recommend a movie that you have watched well recently to everyone~*	0	1
2	中国最高河流雅鲁藏布江！一种净化心灵的美 *The Yarlung Zangbo River, the highest river in China! A beauty that purifies the soul*	0	1
3	普通小狗看见警犬会觉得警察来了吗 *Do ordinary dogs feel like the police are coming when they see a police dog*	1	0
4	导演，没事你接着拍，牛顿的棺材板我帮你按住了 *Director, it's okay. You keep filming. I helped you hold down Newton's coffin board*	1	0

Table 7: Error Analysis.

method is better, we compare five GNN models: RGCN (Schlichtkrull et al., 2018), GraphSage, HAN (Wang et al., 2019), and HGT (Hu et al., 2020). The experimental results are shown in Table 5. Graph-Sage outperforms other GNN models, indicating that its message passing function can learn high-order neighborhood features and improve the semantic representation of humor more efficiently. In addition, dimensionality reduction is an important function of graph representation learning, so we compare the performance of GNNs with different hidden dimensions. It can be found that various GNNs perform well at low hidden dimensions. In addition, increasing the hidden dimension can enhance performance, while 200 dimensions strike a balance between speed and performance.

6.3. Ablation Study (RQ3)

The ablation study results are shown in Table 6. Removing text features from the node representation layer resulted in a substantial drop in performance, indicating that text features dominate in recognizing humor. Conversely, removing non-textual features has a limited impact on the results, suggesting that they have less difference between the data of different classes. Additionally, removing the graph representation learning layer leads to a significant drop in results, implying that the downstream tasks benefit from the process of representation learning. Furthermore, removing both comment nodes and user nodes from the graph degrades the performance, demonstrating that these nodes enhances the representation of article nodes in the humor recognition task. It also shows the effectiveness of our social context graph constructing method.

6.4. Error Analysis

Here are some prediction errors in Table 7. First two sentences are not humor text, but predicted true, while last two sentences are humorous, but predicted false.

The reason for the first two may be "~" and "!", the

"punctuation symbol" besides "," and ".". Humor text is usually full with emotion, and it can be easily expressed by symbols like "?", "!","..." and so on, the first two sentences have the symbol, this may cause the error.

The last two may due to the knowledge shortage. The third sentence is a classic identity misalignment type, model needs to know that police is not a concept for dogs; the last one is about an Internet slang in Chinese, people use "hold down Newton's coffin board" while hearing or seeing something is violating physical laws. If model doesn't know the slang, this sentence cannot be predicted correctly. More knowledge is needed in the model.

7. Conclusion

We propose a novel graph representation called social context graph, which models the information and social entities along with their interactions in social media. Based on the graph, we created the first dataset which introduces social context information - HumorWB, which was collected from Chinese social media - Weibo. And we define two humor-related tasks: humor recognition and sense of humor evaluation, with the latter being a pioneering exploration on humor computing research, and we annotated it basing on psychological theory. For these two tasks, we developed a model based on the social context graph(SCOG) and demonstrated its effectiveness through experiments. Additionally, we discuss the details of our model through three experiments on the humor recognition task.

8. Acknowledgements

9. Bibliographical References

Francesco Barbieri and Horacio Saggion. 2014. Automatic detection of irony and humour in twitter. In *ICCC*, pages 155–162.

Julita Bielaniewicz, Kamil Kanclerz, Piotr Miłkowski, Marcin Gruza, Konrad Karanowski, Przemysław Kazienko, and Jan Kocoń. 2022. Deep-sheep: Sense of humor extraction from embeddings in the personalized context. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 967–974.

Santiago Castro, Matías Cubero, Diego Garat, and Guillermo Moncecchi. 2016. Is this a joke? detecting humor in spanish tweets. In *IBERAMIA*, pages 139–150. Springer.

Huan-Yu Chen, Yun-Shao Lin, and Chi-Chun Lee. 2021. Through the words of viewers: Using comment-content entangled network for humor impression recognition. In *SLT*, pages 1058–1064. IEEE.

Lei Chen and Chungmin Lee. 2017. Predicting audience's laughter during presentations using convolutional neural network. In *BEA*, pages 86–90.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

Jacqueline S Dowling and James A Fain. 1999. A multidimensional sense of humor scale for school-aged children: Issues of reliability and validity. *Journal of Pediatric Nursisng*, 14(1):38–43.

Alan Feingold and Ronald Mazzella. 1993. Preliminary validation of a multidimensional model of wittiness. *Journal of Personality*, 61(3):439–456.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *WWW*, pages 2704–2710.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Gillian A Kirsh and Nicholas A Kuiper. 2003. Positive and negative aspects of sense of humor: Associations with the constructs of individualism and relatedness.

Lizhen Liu, Donghai Zhang, and Wei Song. 2018. Modeling sentiment association in discourse for humor recognition. In *ACL*, pages 586–591.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rod A Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. 2003. Individual differences in uses of humor and their relation

to psychological well-being: Development of the humor styles questionnaire. *Journal of research in personality*, 37(1):48–75.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *EMNLP*, pages 531–538.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *CIKM*, pages 1165–1174.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Matthew C Ramsey and John C Meyer. 2019. Exploring communicative functions of humor: the development and assessment of a new functions of humor scale. *Atlantic Journal of Communication*, 27(1):1–14.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *ESWC*, pages 593–607. Springer.

James A Thorson and Falvey C Powell. 1993. Development and validation of a multidimensional sense of humor scale. *Journal of clinical psychology*, 49(1):13–23.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *WWW*, pages 2022–2032.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *EMNLP-IJCNLP*, pages 3621–3625.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *COLING*, pages 3898–3907.

Yubo Xie, Junze Li, and Pearl Pu. 2021. Uncertainty and surprisal jointly deliver the punchline: Exploiting incongruity-based features for humor recognition. In *ACL-IJCNLP*, pages 33–39.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *CIKM*, pages 889–898.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *ACL*, pages 207–212.

Yichao Zhou, Jyun-Yu Jiang, Jieyu Zhao, Kai-Wei Chang, and Wei Wang. 2020. "the boating store had its best sail ever": Pronunciation-attentive contextualized pun recognition. In *ACL*, pages 813–822.