Computational Linguistics and Chinese Language Processing
Vol. 27, No. 2, December 2022, pp. 73-88

A Chinese Dimensional Valence-Arousal-Irony Detection on Sentence-level and Context-level Using Deep Learning Model

Jheng-Long Wu*, Sheng-Wei Huang*, Wei-Yi Chung*,

Yu-Hsuan Wu* and Chen-Chia Yu+

Abstract

Chinese multi-dimensional sentiment detection task is a big challenge with a great influence on semantic understanding. Irony is one of the sentiment analysis and the datasets established in the previous studies usually determine whether a sentence belongs to irony and its intensity. However, the lack of other sentimental features makes this kind of datasets very limited in many applications. Irony has a humorous effect in dialogues, useful sentimental features should be considered while constructing the dataset. Ironic sentences can be defined as sentences in which the true meaning is the opposite of the literal meaning. To understand the true meaning of a ironic sentence, the contextual information is needed. In summary, a dataset that includes dimensional sentiment intensities and context of ironic sentences allows researchers to better understand ironic sentences. The paper creates an extended NTU irony corpus, which includes valence, arousal and irony intensities on the sentence-level; and valence and arousal intensities on the context-level, which called the Chinese Dimensional Valence-Arousal-Irony (CDVAI) dataset. The paper analyzes the difference of CDVAI annotation results between annotators, and uses a lot of deep learning models to evaluate the prediction performances of CDVAI dataset.

Keywords: Irony Annotation, Dimensional Valence-Arousal-Irony, Sentiment Analysis, Deep Learning.

* Department of Data Science, Soochow University
E-mail: jlwu@gm.scu.edu.tw; {iwihwung11,hwork0511,ikaroskasane}@gmail.com
The author for correspondence is Jheng-Long Wu.
+ The University of Edinburgh School of PPLS
E-mail: alisonyu119@gmail.com

1. Introduction

There are billions of posts on various kinds of forums and social media every day, which shows the exchange of opinions online are high in action and frequency. Human conversations are complex behaviors, because opinions by the people may use direct or indirect presentation sentences. Therefore, the semantic understanding of online opinions is more complicated. In addition, metaphors, irony, sarcasm, etc. also widely appear on online social media. These kinds of expressions cause challenges for natural language understanding (NLU) and natural language processing (NLP). Joshi *et al*. (2018) has reviewed the irony detection problem. Although most of the literature lacks a clear and consistent definition of irony, they found that the most common feature of ironic sentences is the inversion of the literal meaning and true meaning. For example: "Great, it's raining, but I didn't bring an umbrella....", the literal meaning is that raining without an umbrella is a great situation. However, the context "it's raining, but I didn't bring an umbrella..." shows a negative emotion, contrasted with "Great" which is a positive emotion. This emotional contrast caused the semantic turn from negative to positive, which enables the expression of irony. In Chinese irony, the contrast between positive and negative emotions is often used to indicate the difference between sentences and contexts. This emotional contrast is often used to achieve ironic expressions (Veale & Hao, 2010). According to the grammatical structure mentioned above, this study argues that context must be considered to match the characteristics of ironic sentences to improve the performance on irony detection task. The work in sentiment analysis of irony has turned to the study of ironic language features (Colston, 2019). With the development of machine learning, some studies have begun to use machine learning methods to predict the intensity of irony (Chia *et al*., 2021; Dimovska *et al*., 2018). However, most of them still predict irony using whole sentences instead of considering context as mentioned above.

To improve machine learning performance of detecting ironic sentences, some studies proposed to annotate grammatical structural features or use feature selection to screen important irony spans in the English language (Kumar & Harish, 2019). Long *et al*. (2019) proposed the usage of capitalized words as a hint of irony in English. However, capitalization does not exist in Chinese so the capitalization is not suitable for use. In conclusion, while the grammatical structure of irony has been thoroughly studied in English, it is not appropriate to apply it directly to Chinese. Although some studies summarized Chinese irony grammatical structures (Jia *et al*., 2019), there are few datasets annotated based on these rules. Since irony has a humorous effect in the conversation processes, the paper considers irony detection as a sentiment detection task. Therefore, considering the multi-dimensional Valence-Arousal-Irony (VAI) intensity for irony sentences and context is more possible to identify the true meaning of ironic sentences and the emotional state of the social media user.

Based on Tang's (Tang & Chen, 2014) open data on irony sentences, the paper proposes

to extend sentence-level intensity of valence, arousal and irony, and context-level intensity of valence and arousal. This annotation method provides a way to judge the difference in context and semantics in irony sentences. By quantifying emotional indicators, the pattern of sentiment while using ironic sentence can be more easily understood. This augmented CDVAI dataset is the first dataset to do sentiment annotations for irony context.

Furthermore, this paper proposes deep learning models based on the pretrained Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al*., 2018) model to learn the dimensional VAI on the ironic sentences and dimensional VA on ironic contexts. The paper uses pre-trained BERT to extract hidden features of sentence or context, respectively. Then there are three methods to combine hidden features and predict VAI scores of sentence-level or VA scores of contexts-level: (1) using a linear layer to predict VAI and VA, respectively; (2) summing two hidden features from two encoders of sentence and context. (3) Concatenating two hidden features from two encoders of sentence and context. Furthermore, the paper constructs a token classification model to automatically predict the position of context. Then the predicted positions of context are used to replace the origin positions of context, and predict VAI scores of sentence-level or VA scores of contexts-level.

2. Related Works

Because of different research perspectives, the definition of irony is often adjusted. However, previous studies summarized a basic consensus in the process of exploring ironic sentences. “Irony is an expression in which the true meaning is the opposite of the literal meaning” (Li & Huang, 2020). Based on the above, the most common feature of irony is metaphor, which can make the literal meaning opposite to the true meaning of the sentence that the commenter wants to express. The form of ironic sentence can be expressed as using keywords of exaggeration with positive emotions to describe context with negative emotions. This emotional contrast makes the sentences have an ironic effect (Veale & Hao, 2010). Li *et al*. (2020) proposed an irony identification program (IIP) based on the grammatical structure of ironic sentences, which supports future studies to identify whether a sentence is ironic. The above research provides support for the definition of irony in the paper.

Irony sentences are not usually used in official documents. Thanks to the prevalence of social media, many ironic sentences have been posted online which has led researchers to collect and analyze ironic sentences on social media platforms (Lestari, 2019). Among the studies related to irony detection or sentiment detection. There are very few corpuses including VAI indicators. The possible reasons are that irony detection is not traditionally attributed to the domain of sentiment detection. However, irony has a humorous effect in conversation, which can result in specific emotional patterns for the writer and reader. Therefore, the paper considers irony detection as an emotion detection task. But most of the existing corpus are only included

valence and arousal (VA) or only include irony (I) indicator.

Recent studies have collected data on social media to build corpus. Preoţiuc-Pietro *et al*. (2016) used the Likert nine-point scale to annotate VA indicators for Facebook posts. They found that there is high correlation between VA. Bosco *et al*. (2013) annotate irony and emotional expression for Twitter tweets to establish the Senti-TUT corpus. Their corpus includes positive, negative emotions and irony, which considers the concept of valence. Ghosh *et al*. (2015) annotate figurative language such as irony, satire, and metaphor on a 11-point scale at SemEval-2015 Task 11. In addition, there are many constructions of VA or I corpus, but there are very few studies that comprehensively considers VAI.

The necessity of considering VAI indicators simultaneously is that there are correlations among the three indicators. Effects of irony on human emotions in conversation was found in the study of Pfeifer (Pfeifer & Lai, 2021). People who use irony are in a less negative and less excited state of mind. Existing VAI corpus was constructed by Xie *et al*. (2021). They found that stronger irony expressions may have lower valence (more negative) and higher arousal levels, respectively. However, since context is important information to construct ironic sentences which their study didn't consider. The biggest difference between Xie *et al*. (2022) and the paper is that the context is considered and annotated with VA score. To conclude, the above study proves that it is necessary to consider VAI together because of the correlations in these three indicators.

Irony Corpus built in Chinese such as Xiang *et al*. (2020) proposed Ciron dataset. Their dataset contains 8.7K Weibo posts. However, they annotated the intensity of ironic sentences in the corpus without considering context and other sentiment indicators. Existing corpus that include irony sentences and context is NTU Irony Corpus (Tang & Chen, 2014), but their corpus without other sentiment indicators. Lack of consideration of sentiment indicators is impossible to understand clearly on the emotional transitions and semantic changes in the sentences. Therefore, the corpus provided in this paper has a greater advantage in understanding the structure of ironic sentences and sentiment patterns.

In terms of the irony detection model, Rangwani *et al*. (2018) considered emojis on Twitter as a feature when annotating ironic sentences. They use Convolutional Neural Network (CNN) to pre-train the emoji and connect to a XGBoost model for classification. Naseem *et al*. (2020) proposed a T-Dice model based on the frame of the Transformer to detect valence and irony, then connected to Bi-directional Long Short-Term Memory (Bi-LSTM) to classify emotions. The accuracy of their model's prediction results exceeded the state-of-the-art methods of the time. Xiang *et al*. (2020) found that the performance of BERT is better than GRU in their experimental results on the Ciron dataset they built. Lu *et al*. (2020) improved the Bi-GRU model based on BERT in the Chinese sentiment analysis task to achieve the best results. To sum up, in recent years, no matter in sentiment or irony detection tasks. Models that can connect the

information of the entire sentence have achieved better results. Furthermore, models with attention mechanisms such as BERT or based on Transformers frame can make the model achieve better results. In summary, this paper will base on BERT to detect the VAI score of sentences and the VA score of the contexts.

3. CDVAI Dataset

The paper proposes to extend the NTU irony corpus to a Chinese dimensional valence-arousal-irony called CDVAI. The NTU irony corpus is the only Chinese corpus that includes ironic sentences and contexts. Therefore, the paper proposes to annotate the VAI intensity of the sentence-level and the VA intensity of the context-level, respectively. Li and Huang (2020) analyzed the sentence structure of Chinese irony based on the existing corpus. They summarized that context is an important information for detecting irony. Based on the sentence structure in the NTU irony corpus and their findings, the paper defines irony as "irony is an expression in which the true meaning is the opposite of the literal meaning." Context is the true meaning of the sentence (usually a negative description), while ironic keywords (usually positive descriptions) can make the literal meaning contrary to the context.

3.1 Dimensional VAI annotation

The paper annotated irony sentences with VAI intensity, and irony contexts with VA intensity. Every indicator was rated from 1 to 5 points. The detailed annotation judgement as follow:

- Valence: Lower valence scores indicate more negative emotions (1-2 points), whereas higher valence scores indicate more positive emotions (4-5 points), and 3 indicate neutral emotions, or inability to judge.
- Arousal: A score of 1 indicates the sentence is close to an objective description, or difficult to judge whether the sentence expresses excitement. A score of 2 indicates that the annotator can feel the low excitement expressed in the sentence, but there is no emotion word such as sad, annoyed, lost, happy, etc. in the sentence. A score of 3 and above indicate the annotator can feel the medium excitement expressed in the sentence, or with explicit emotional words or phrases to clearly describe the emotional state. A score of 4 indicates that the annotator can clearly feel strong excitement expressed in the sentence, such as madness, rage, etc. Furthermore, the sentence may contain violent words, such as aggressive language. A score of 5 indicates in addition to strong excitement, words with discrimination, hated, or words with obvious manic emotions. For example: “Great, the class report is going to be in the same group with that pathetic nerd!”.
- Irony: The annotator reads a sentence and judges whether the true meaning is the opposite of the literal meaning. Most of the sentences in NTU irony corpus use negative

descriptions as the context, and positive descriptions as the keywords to express irony. Irony intensity will be determined according to the gap between the positive intensity of irony keywords and the negative intensity of context. In this paper, the positive intensity of various ironic keywords appearing in the corpus is summarized as: wonderful > great > very good > good. A special case is "it's fine to get worse!", the true meaning in this case is the situation is already bad but the commenter doesn't want the situation to get worse, the ironic keywords "it's fine to" makes the literal meaning opposite to the true meaning. However, this case means the situation is already bad so the gap between positive intensity of irony keywords and the negative intensity of context is small. The larger the gap between the positive intensity of the ironic keyword and the negative intensity of the context, the higher the score of irony, and vice versa. A score of 1 indicates that the gap is very small, or the context is close to an objective description, which leads to hard judgement. For example: "Good, it's raining.". A score of 2 indicates that there is a small gap between ironic keywords and context. A score of 3 indicates that there is a moderate gap between the ironic keywords and the context. A score of 4 indicates that there is a big gap between the ironic keywords and the context. A score of 5 indicates that there is a great gap between ironic keywords and context. The sentence may contain discriminatory or morally unacceptable metaphors, such as sexual innuendo.

3.2 Annotated result analysis

There are 1004 sentences in NTU Irony Corpus, and 843 sentences with an ironic context. Each sentence was annotated by three annotators. The annotators consist of postgraduate students and an undergraduate student, all of them are native Chinese speakers and ages between 20 and 25. Due to the subjective judgement bias of different annotators, the paper uses the average of 3 annotators as the gold standard. The paper using scores to annotate VAI is more reasonable. Human perception of emotional intensity is closer to continuous scores than classification. The meaning of the annotating criterion in the paper is to concretize the definition of intensity of VAI and set the standard score line. Continued from above, the traditional method which is used to evaluate the agreement between annotators such as Cohen's kappa doesn't conform to the hypothesis of the paper. So, the paper uses mean absolute error (MAE) to evaluate the annotation consistency. At the sentence level, the MAE of the three annotators ranged from 0.05 to 0.31 in valence, 0.25 to 0.41 in arousal, 0.22 to 0.56 in irony. At the context level, the MAE of the three annotators ranged from 0.07 to 0.4 in valence, 0.15 to 0.65 in arousal. From the above, the MAE difference between of the three annotators is very small, which proves that the annotating is effective.

- **For example:**

Score of a sentence: valence: 1, arousal: 5, irony: 4

Score of a context: valence: 1, arousal: 5

Sentence: "*很好 (applause)工廠的廠務小姐已經來上班好多好多年了,跟我說她不會用 outlook 發會議通知!!ㄍㄋㄋ勒!!妳的薪水也給我我就幫你發通知!!*" ("*Very good (applause) The factory manager of the factory has been coming to work for many years. She told me that she doesn't know how to use Outlook to send meeting notices!! mother fucker!! Give me your salary and I will send the notices for you!!*")

Context: "*工廠的廠務小姐已經來上班好多好多年了,跟我說她不會用 outlook 發會議通知!!*" ("*The factory manager of the factory has been coming to work for many years. She told me that she doesn't know how to use Outlook to send meeting notices!!*")

Judgement: First, in terms of judging the score of valences, there are extremely negative emotions in this sentence such as "mother fucker!! Give me your salary and I will send a notice for you!!". Clearly, the emotions expressed by the swear words and complaints in the sentence are highly negative. Thus, valence is given a score of 1. In terms of judging the score of arousals, we can notice the abuse language and feel the emotion of manic. Thus, arousal is given a score of 5 points. In terms of judging the score of irony, the irony keyword "very good" is a weak positive description. However, according to the description of the sentence, the incident described in the context caused serious discomfort and negative emotions to the commenter. As we can see, there is a big gap between positive irony keyword and negative describe of context. Besides that, the sentence also contains sarcasm spans, such as "Give me your salary and I will send a notice for you.", so it is given a high score of 4 points in irony.

3.3 Statistics of Annotated Result

Table 1 shows the annotated result of CDVAI dataset in different levels and sentiment. Since the dataset is mainly ironic sentences, which results in valence scores that are all low (negative emotion) at sentence-level. While few valence scores of contexts are neutral at context-level. The sentences corresponding to these kinds of contexts often show low scores in valence and irony. There are many sentences containing emotions, which can be observed in the arousal scores centered on points 2, 3 and 4. The score of arousals at context-level is distributed to a lower score than sentence-level. The reason is that irony keywords usually have exaggerated expressions, resulting in a higher arousal. The distribution of the score is like arousal. Gap between positive irony keyword and negative context are usually small, which can be observed in the irony scores centered on points 1, 2 and 3.

Table 1. Score frequency of all sentiments.

Level	Sentiment	0	1	2	3	4	5
Sentence	Valence	0	380	624	0	0	0
	Arousal	0	60	406	369	150	46
	Irony	0	181	428	310	75	20
Context	Valence	0	302	516	25	0	0
	Arousal	56	279	264	161	76	26

4. Model Performance Evaluation

To validate the annotation consistency and the validity of the proposed CDVAI dataset in the paper, the paper constructs deep learning models to predict the VAI score of sentence-level and VA scores of context-level. Table 2 shows the general statistics of CDVAI dataset. The paper uses stratified sampling to split the dataset into training, validation, and the testing set. The ratio of training set and testing set is 7:3, and validation set is split from training set which ratio is 9:1.

Table 2. Statistics of the proposed CDVAI dataset.

dataset	Sentence-level	Context-level
Training set	632	531
Validation set	71	59
Testing set	301	253

4.1 Prediction Model

This paper uses pre-trained BERT models as an encoder to extract hidden state of sentences and contexts, then connected to a linear layer to perform score prediction. There are three methods to obtain final hidden features such as (1) M1: After input sentence and context into the encoder, the hidden features of the sentence are used to predict sentence VAI score through a linear layer. The hidden features of context are used to predict context VA score. (2) M2: The position of the context in the sentence has been located. After input sentence and context into the encoder, the hidden features at the context position are summed, then predict sentence VAI and context VA scores. (3) M3: After input sentence and context into the encoder, concatenate two hidden features of sentence and context then predict sentence VAI and context VA scores. Above processes are the first part of the experiment in this paper. The second part of the experiment, the paper attempted to create a model to predict context span automatically. The paper uses the pre-trained BERT models as encoders, and then the output of encoder with linear layer to predict the span of context in the sentence. Finally, the predicting context will replace the origin context in the first part of the experiment, with the predicted context of the proposed model, then repeat

the process of the first experiment.

The paper compares BERT models pre-trained on a Chinese corpus to find the best results. The pre-trained models are as follow:

- PM1: *hfl/chinese-macbert-base* uses Wikipedia simplified and traditional Chinese as the corpus to train the model. (Cui *et al*., 2020)
- PM2: *shibing624/macbert4csc-base-chinese* using the SIGHAN typo correction corpus to train the model. (Cui *et al.*, 2020)
- PM3: *uer/chinese_roberta_L-4_H-256* uses UER toolkit and CLUECorpus2020 to train the model. (Turc *et al*., 2019)
- PM4: *IDEA-CCNL/Erlangshen-Ubert-110M-Chinese* uses datasets from a variety of tasks for open-source UBERT. (Wang *et al*., 2022)
- PM5: *IDEA-CCNL/Erlangshen-Ubert-330M-Chinese* uses datasets from a variety of tasks for open-source UBERT.
- PM6: *IDEA-CCNL/Erlangshen-UniMC-RoBERTa-110M-Chinese* uses 13 supervised datasets to train the model. (Yang *et al.*, 2022)

4.2 Experimental Settings

The proposed CDVAI dataset includes the annotation of irony context to allow the model to understand contextual emotional changes during the training process. The paper uses a variety of modified pre-trained BERT models as the experimental encoder. The parameters are shown in Table 3. Each pre-trained model uses the same parameters, except the learning rate. Since context contains less information than sentences, a smaller learning rate should be tried. The context span prediction model in the second part of the experiment were tried smaller learning rate due to the difficulty to learn the span of context in the sentence.

Table 3. Parameter settings of BERT models.

Parameter	Value
Optimizer	Adam
Learning rate - sentence-level	4e-4, 4e-5, 4e-6
Learning rate - context-level	4e-5, 4e-6, 4e-7
Learning rate – span prediction	43e-6, 45e-6
Number of epochs	50

4.3 First Part of Experimental Results

The prediction performance of dimensional VAI score on sentence-level is shown in Table 4. First, the performance of valence prediction is quite good. All MAEs are about 0.4, no matter what approach in this paper. However, the paper can still discover that M1 got the greatest performance, which indicates more complex hidden features don't get better result in valence. The reason is detecting the score of valences is relatively easy in the task, so more complex hidden features cause worse results. The performance of arousal prediction is a bit worse than valence, which indicates arousal is relatively difficult to learn. All MAEs are about 0.6, however M1 does not have the greatest approach on all models. M2 and M3 make the performance progress at PM3. PM4, MP5 and PM6 improve performance while using M2 or M3. Finally, the performance of irony prediction is a bit better than arousal. All MAEs are about 0.5 to 0.6, which means our annotated method to judge irony is effective. M2 and M3 are more helpful to improve the performance of PM2, PM4, PM5 and PM6, which indicate these models can deal with complex hidden features better. Overall, the result of sentence-level VAI is quite well, but M2 and M3 doesn't show significantly helpful while predict VAI scores.

Table 4. Prediction performance of dimensional VAI score on sentence-level.

Model	Valence			Arousal			Irony		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
PM1	0.346	0.421	0.390	0.521	0.649	0.639	0.521	0.577	0.603
PM2	0.380	0.421	0.390	0.619	0.649	0.639	0.601	0.577	0.603
PM3	0.371	0.410	0.371	0.643	0.603	0.596	0.538	0.570	0.566
PM4	0.380	0.412	0.390	0.619	0.614	0.639	0.601	0.572	0.603
PM5	0.376	0.381	0.420	0.615	0.616	0.610	0.575	0.591	0.559
PM6	0.380	0.412	0.390	0.619	0.614	0.639	0.601	0.577	0.603

The prediction performance of dimensional VA score on context-level is shown in Table 5. The context-level valence also performs quite well. Overall, MAE is around 0.4. However, M2 and M3 improve the performance significantly. Among them, M3 provides an even better effect. This shows our approaches are more effective on context-level. The reason may be the complex relation of sentence and context, which shows that the true sentiment pattern of ironic sentences requires a judgment of the context first then combined with the whole sentence to understand. This result can also be seen in arousal. However, M2 showed more effective help in predicting arousal scores. The inference is the information of context itself is more important than the whole sentence while predicting arousal scores, and this effect significantly.

Table 5. Prediction performance of dimensional VA score on context-level.

Model	Valence			Arousal		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**
PM1	0.431	0.408	0.389	0.819	0.649	0.787
PM2	0.413	0.408	0.389	0.796	0.649	0.787
PM3	0.431	0.468	0.427	0.798	0.603	0.829
PM4	0.413	0.415	0.389	0.796	0.614	0.787
PM5	0.426	0.463	0.428	0.815	0.616	0.834
PM6	0.413	0.415	0.389	0.796	0.614	0.787

Analysis of the above shows that M2 and M3 can improve the performance on context-level significantly. However, they don't seem quite helpful on sentence-level. In summary, depending on the choice of pre-trained model, context information can improve performance while predicting VAI score in sentence. Results on context-level shows that understanding the true sentiment pattern of ironic sentences requires to combine sentence and context information.

4.4 Second Part of Experimental Results

Due to the lack of context annotation in previous study. The paper proposes a model to predict the irony context span automatically. The paper proposes to fine-tuning PM1 to PM6 to compare prediction performances. But the performances of the model are hard to accept. So, the paper adds a new pre-trained model to solve this problem, which is PM7: *IDEA-CCNL/Erlangshen-DeBERTa-v2-97M-Chinese* (He *et al*., 2020) to improve the model. The results show in Table 6.

Table 6. Prediction performance of context span predict in ironic sentences.

Model	Indicators		
	Precision	**Recall**	**F1**
PM1	0.373	0.302	0.333
PM2	0.349	0.274	0.307
PM3	0.329	0.218	0.262
PM4	0.373	0.309	0.331
PM5	0.436	0.352	0.390
PM6	0.373	0.298	0.338
PM7	0.438	0.377	0.405

The paper uses PM7 to predict context span, then replace origin context span with the predicted context span and execute the same process of above experiment to evaluate the

availability. The span predict model results on sentence-level are shown in Table 7. Compared with the first part of the experiment, the MAE of valence on sentence-level on M2 is making progress. The reason may be that although the predicted context spans are not correct, they contain more emotion information words accidently. The MAE of arousal on sentence-level becomes larger on M2, however the MAE reduces on M3. The reason may be the noise of the context improves the performance. The same situation occurs with irony. This discovery is quite surprising that the information of whole context may not be the important one to improve the prediction performance but the critical part or words in the context.

Table 7. Prediction performance of dimensional VAI score on sentence-level in part 2 experiment.

Model	Valence		Arousal		Irony	
	M2	M3	M2	M3	M2	M3
PM1	0.337	0.403	0.631	0.585	0.575	0.638
PM2	0.337	0.403	0.631	0.585	0.575	0.638
PM3	0.384	0.376	0.618	0.648	0.606	0.613
PM4	0.392	0.403	0.677	0.585	0.594	0.638
PM5	0.383	0.364	0.607	0.609	0.561	0.574
PM6	0.392	0.403	0.677	0.585	0.594	0.638

Finally, the span predict model results on context-level are shown in Table 8. Since the context span of the model predictions cannot be fully correct. Therefore, the main purpose of this part of the experiment is to examine the effect of VA score prediction with a biased context span. Compared with the first part of the experiment, the MAE of valence on context-level decreases a little. The MAE of arousal decreases quite a lot. This proves that the correction of context span matters.

Table 8. Prediction performance of dimensional VAI score on context-level in second part of experiment.

Model	Valence		Arousal	
	M2	M3	M2	M3
PM1	0.419	0.447	0.819	0.824
PM2	0.419	0.447	0.819	0.824
PM3	0.471	0.436	0.867	0.810
PM4	0.435	0.447	0.844	0.824
PM5	0.442	0.451	0.801	0.844
PM6	0.435	0.447	0.844	0.824

4.5 Error Analysis

Based on the performance of the model, the PM3 model has well performance in experiments. The paper presents an incorrect prediction case, as follows:

Sentence: "*很好...連喇叭都壞了 X-(" ("Very good.... even the speakers are broken X-(")*

Context: "*連喇叭都壞了" ("even the speakers are broken*")

Judgement: The prediction results are shown in Table 9. The reason why the model judges the valence to be 1.71 on sentence-level, may be that it judges "連", "壞了" ("*even, broken*") as negative words. However, the post only indicated that the speakers are broken, which is usually not perceived as highly negative. The lack of common sense may have led to the failure to detect its valence correctly. In terms of irony, the prediction score is relatively large. It is speculated that because the judgment of valence is relatively negative and the term "很好" ("*very good*") is positive, there is a large emotional gap. The model therefore yields a higher irony score. However, the sentence has no other span that emphasizes irony, so the annotated score is lower.

Table 9. Prediction results of the example

	Sentence-level			Context-level	
	V	**A**	**I**	**V**	**A**
Annotated	2	3	1	2	3
Predicted	1.71	3.45	1.94	1.63	1.97

5. Conclusion

This paper established the CDVAI dataset which extended from NTU irony corpus. The CDVAI dataset contains multi-dimensional sentiment annotation and irony context sentiment annotation, which is helpful for developing Chinese irony detection methods that allow the model to learn sentimental patterns in ironic sentence and context. The experimental results showed that the annotation of CDVAI dataset provides a learning direction for the BERT based model to understand the irony structure and sentiment contrast between sentence-level and context-level. Using M3 can improve performance significantly. The paper has summarized our experiment results below. First, M2 and M3 don't show significantly helpful while predicting VAI scores. However, in the second part of the experiment that the information of the whole context may not be important to improve the prediction performance but the critical part or words in the context. Second, M2 and M3 show significant improvement in predicting score of context-level, which proves the sentiment pattern of ironic context needs to combine sentence information. Finally, the sentiment in ironic contexts is harder to learn for the model, which needs correct spans of context to improve the performance.

The weakness of the CDVAI dataset is that the corpus is relatively small and excludes the whole ironic grammatical structure. Nevertheless, the paper is suitable to use as guide data to obtain more samples or as a template for annotation guidelines. Furthermore, the proposed CDVAI dataset could be combined with other ironic corpora to extend the training sample size. Furthermore, the model can be improved in the future.

Acknowledgments

This research was partially supported by the Ministry of Science and Technology (MOST), Taiwan (Grant numbers: MOST 110-2221-E-031-004, and MOST 111-2221-E-031-004MY3).

References

Bosco, C., Patti, V., & Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, *28*(2), 55-63.

Colston, H. L. (2019). Irony as indirectness cross-linguistically: On the scope of generic mechanisms. In *Indirect Reports and Pragmatics in the World Languages* (pp. 109-131). Springer.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. arXiv preprint arXiv:2004.13922

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. Version 2.

Dimovska, J., Angelovska, M., Gjorgjevikj, D., & Madjarov, G. (2018). Sarcasm and irony detection in English tweets. In *Proceedings of the International Conference on Telecommunications*, 120-131. https://doi.org/I:10.1007/978-3-030-00825-3_11

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., & Reyes, A. (2015). Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 470-478.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

Jia, X., Deng, Z., Min, F., & Liu, D. (2019). Three-way decisions based feature fusion for Chinese irony detection. *International Journal of Approximate Reasoning*, 113, 324-335. https://doi.org/10.1016/j.ijar.2019.07.010

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2018). Automatic sarcasm detection: A survey. *ACM Computing Surveys*, *50*(5), 1-22. https://doi.org/10.1145/3124420

Kumar, H., & Harish, B. (2019). Automatic irony detection using feature fusion and ensemble classifier. International Journal of Interactive Multimedia & Artificial Intelligence, 5(7). 70-79. https://doi.org/10.9781/ijimai.2019.07.002

Lestari, W. (2019). Irony analysis of Memes on Instagram social media. *PIONEER: Journal of Language and Literature*, *10*(2), 114-123. https://doi.org/10.36841/pioneer.v10i2.192

Li, A.-R., & Huang, C.-R. (2020). A method of modern Chinese irony detection. *From Minimal Contrast to Meaning Construct* (pp. 273-288). Springer.

Long, Y., Xiang, R., Lu, Q., Huang, C.-R., & Li, M. (2019). Improving attention model based on cognition grounded data for sentiment analysis. *IEEE transactions on affective computing*, *12*(4), 900-912. https://doi.org/10.1109/TAFFC.2019.2903056

Lu, Q., Zhu, Z., Xu, F., Zhang, D., Wu, W., & Guo, Q. (2020). Bi-GRU sentiment classification for Chinese based on grammar rules and BERT. International Journal of Computational Intelligence Systems, 13(1), 538-548. https://doi.org/10.2991/ijcis.d.200423.001

Naseem, U., Razzak, I., Eklund, P., & Musial, K. (2020). Towards improved deep contextual embedding for the identification of irony and sarcasm. In *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*, 1-7. https://doi.org/10.1109/IJCNN48605.2020.9207237

Pfeifer, V. A., & Lai, V. T. (2021). The comprehension of irony in high and low emotional contexts. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *75*(2), 120-125. https://doi.org/10.1037/cep0000250

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., & Shulman, E. (2016). Modelling valence and arousal in facebook posts. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 9-15. https://doi.org/10.18653/v1/W16-0404

Rangwani, H., Kulshreshtha, D., & Singh, A. K. (2018). Nlprl-iitbhu at semeval-2018 task 3: Combining linguistic features and emoji pre-trained cnn for irony detection in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, 638-642. https://doi.org/10.18653/v1/S18-1104

Tang, Y.-j., & Chen, H.-H. (2014). Chinese irony corpus construction and ironic structure analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1269-1278.

Turc, I., Chang, M. W., Lee, K., & Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962.

Veale, T., & Hao, Y. (2010). Detecting ironic intent in creative comparisons. In *Proceedings of 19th European Conference on Artificial Intelligence*, 765-770. https://doi.org/10.3233/978-1-60750-606-5-765

Wang, J., Zhang, Y., Zhang, L., Yang, P., Gao, X., Wu, Z., Dong, X., He, J., Zhuo, J., Yang, Q., Huang, Y., Li, X., Wu, Y., Lu, J., Zhu, X., Chen, W., Han, T., Pan, K., Wang, R.,... & Zhang, J. (2022). Fengshenbang 1.0: Being the foundation of Chinese cognitive intelligence. arXiv preprint arXiv:2209.02970

Xiang, R., Gao, X., Long, Y., Li, A., Chersoni, E., Lu, Q., & Huang, C.-R. (2020). Ciron: a new benchmark dataset for Chinese irony detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 5714-5720.

Xie, H., Lin, W., Lin, S., Wang, J., & Yu, L.-C. (2021). A multi-dimensional relation model for dimensional sentiment analysis. *Information Sciences*, *579*, 832-844. https://doi.org/10.1016/j.ins.2021.08.052

Yang, P., Wang, J., Gan, R., Zhu, X., Zhang, L., Wu, Z., Gao, X., Zhang, J., & Sakai, T. (2022). Zero-shot learners for natural language understanding via a unified multiple choice perspective. arXiv preprint arXiv:2210.08590.