# A Dimensional Valence-Arousal-Irony Dataset for Chinese Sentence and Context

**Sheng-Wei Huang**
Soochow University
Dept. of Data Science
kiwihwung11@gmail.com

**Wei-Yi Chung**
Soochow University
Dept. of Data Science
hwork0511@gmail.com

**Yu-Hsuan Wu**
Soochow University
Dept. of Data Science
ikaroskasane@gmail.com

**Chen-Chia Yu**
The University of Edinburgh
School of PPLS
alisonyu119@gmail.com

**Jheng-Long Wu**
Soochow University
Dept. of Data Science
jlwu@gm.scu.edu.tw

## Abstract

Chinese multi-dimensional sentiment detection is a challenging task with a considerable impact on semantic understanding. Past irony datasets are utilized to annotate sentiment type of whole sentences of irony. It does not provide the corresponding intensity of valence and arousal on the sentences and context. However, an ironic statement is defined as a statement whose apparent meaning is the opposite of its actual meaning. This means that in order to understand the actual meaning of a sentence, contextual information is needed. Therefore, the dimensional sentiment intensities of ironic sentences and context are important issues in the natural language processing field. This paper creates the extended NTU irony corpus, which includes valence, arousal and irony intensities on sentence-level; and valence and arousal intensities on context-level, called Chinese Dimensional Valence-Arousal-Irony (CDVAI) dataset. Therefore, this paper analyzes the annotation difference between the human annotators and uses a deep learning model such as BERT to evaluate the prediction performances on CDVAI corpus.

Keywords: Irony annotation, Dimensional valance-arousal-irony, Sentiment analysis, Deep learning

## 1 Introduction

The popularity of social media has made the exchange of opinions more frequent, and users not only use narratives but also widely use irony, metaphors and other special expressions when commenting on online forums. In the past, the literature has compiled research on irony detection (Joshi et al., 2017). Although most of the literature lacks a clear and consistent definition of irony, one of the most common features of irony is the inversion of the literal meaning and the semantic turn of the context. 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 sarcastic expressions (Veale & Hao, 2010). For example "Great, it's raining, but I didn't bring an umbrella....", the context "it's raining, but I didn't bring an umbrella...." is a negative emotion, keyword "Great" is a positive emotion. This emotional contrast enables the expression of irony. In order to improve the performance on the irony recognition task, this study argues that context must be considered to match the characteristics of ironic sentences. As the grammatical structure of the above-mentioned irony suggests, irony emotion detection is quite difficult in natural language processing (NLP).

As a result, past research on irony detection is rare, and the work of emotion analysis turned to the study of characteristics of irony language (Colston, 2019). With the development of machine learning, some studies have gradually begun to use its methods to predict the degree of irony (Chia et al., 2021). However, due to the limitation of annotation, most of them predict only the degree of irony of the whole sentence (Dimovska et al., 2018) instead of considering the expression differences of the context as mentioned above.

In order to improve machine learning of identifying the intensity of irony, scholars have proposed to annotate these structural features or use feature selection to screen important irony spans when studying irony in the English language. (Kumar & Harish, 2019). However, grammatical structures can differ in different languages, and the improvement of irony detection performance cannot be handled in the same way. Long et al. (2019) proposed the usage of capitalized words as a hint of irony in English. Such a method is not suitable for learning features of irony in Chinese, as the notion of capitalization does not exist in the Chinese language. In conclusion, while these rules have been thoroughly studied in English, their applicability to Chinese is an inappropriate approach. Although some scholars are studying Chinese irony rules (Jia et al., 2019), there are few datasets that are quantified and annotated based on these rules. In conclusion, considering the multi-dimensional Valence-Arousal-Irony (VAI) Irony Sentences and Context Dataset, it is possible to identify the true meaning of ironic sentences and the emotional state of the sender, which also contributes to the field of Chinese NLP.

The construction of the existing VAI corpus was carried out with the efforts of Xie et al. (2021). Its contribution is to show that the VAI indicator has the characteristics of mutual influence. The biggest difference between the corpus established in this study and Xie et al. is that the context is considered and annotated with VA.

Based on Tang's (Tang & Chen, 2014) open data on irony sentences, this study proposes to add sentence-level valence, arousal and irony annotations, and context-level valance annotations to the ironic sentences provided by the dataset. This annotation method provides a way to judge the difference in context and semantics in the subsequent analysis of irony sentences. By quantifying emotional indicators, the degree of irony is more clearly understood. This augmented CDVAI dataset is the first dataset to do sentiment annotations for irony context.

Furthermore, this paper proposes a deep learning model based on the work of Devlin et al. (2018), Bidirectional Encoder Representations from Transformers (BERT) to learn the dimensional VAI on the ironic sentences and dimensional VA on ironic contexts. The experimental models include (1) using a linear layer with pre-trained BERT to predict dimensional VAI on sentence and context; (2) sum hidden features of corresponding the context from pre-trained BERT (3) concatenate two hidden features of BERT from sentence and context, respectively.

## 2 Related works

Metaphor is a feature of irony, which can be expressed as the use of exaggerated keywords with positive emotions to describe things with negative emotions, which also makes the apparent meaning of the sentence opposite to the actual meaning that the speaker expects or wants to convey. It is also frequently used in satirical sentences. Since irony is not commonly used in official documents, most researchers turn to social media platforms for data collection and analysis (Lestari, 2019). Due to different research perspectives, the definition of irony is often adjusted. However, researchers have reached a basic consensus in the process of exploration, that is, the basic feature of irony. "Irony is an expression in which the true meaning is the opposite of the literal meaning" (Li & Huang, 2020) Li et al. (2020) proposed an Irony Identification Program (IIP) to identify whether a sentence is ironic during the annotation process. Using IIP, they studied the semantic relations in the grammatical structure of irony according to the context. The above research provides support for the definition of irony in this study.

English has few corpora for VAI, and most of them are only for VA (valence and arousal). In a recent study, Preoţiuc-Pietro et al. (2016) established and annotated the VA tags for Facebook posts. They used the Likert nine-point scale to annotate and found that the two indicators of VA had a high correlation. In addition, in the construction of the irony corpus, Bosco et al. (2013) proposed the corpus Senti-TUT to mark the irony and emotional expression of tweets on Twitter. Their work includes positive and negative emotions, emotional expression, and irony. The corpus considers the concept of valence instead of focusing only on irony. Gosh et al. (2015) Annotate figurative language such as irony, satire, and metaphor on a 11-point scale at SemEval-2015 Task 11 and recruit annotators using a crowdsourcing platform. In addition, there are still many foreign languages for the construction of VA or I corpus, but literature that

comprehensively considers all three aspects (which is VAI corpus) is very rare.

While few studies consider and label all three indicators simultaneously (VAI), there is a correlation among the three indicators, and the following studies demonstrate the need to do so. The effect of irony on human emotions was found in the study of Pfeifer (Pfeifer & Lai, 2021). Regardless of contextual emotion, people who use irony are considered to be in a less negative and less excited state of mind. The study by Xie et al. (Xie et al., 2021). found that stronger irony expressions may have lower valence (more negative) and higher arousal levels, respectively.

Research on Chinese Irony Corpus Xiang et al. (2020) constructed a corpus for irony. The Ciron dataset they built contains 8.7K Weibo posts. The study only annotated the degree of irony of sentences in the corpus without considering the context. NTU Irony Corpus (Tang & Chen, 2014) has released that it only provides the ironic sentences and context, but without other sentiment scores such as sentence-level VAI. Lack of carry out more subtle emotional labeling for the internal grammatical structure, which 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.

In the follow-up application in the field of irony, Rangwani et al. (2018) considered emojis, which are often used on Twitter, as a factor when annotating ironic sentences to improve the results. CNN (Convolutional Neural Network) is implemented to pre-train the emoji, and XGBoost model is applied for classification. Naseem et al. (2020) proposed a T-Dice model based on the transformer model to judge post valence, irony and irony classification. It was then connected to Bi-LSTM (Bi-directional Long Short-Term Memory) to classify emotions, and its accuracy surpassed the most advanced methods at that time. Xiang et al Xiang et al. found that the effect of BERT is better than that of GRU in the experimental results of 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 analysis or in irony recognition tasks, LSTM and other models that can connect the information of the entire sentence have achieved better results, and based on models with attention mechanisms such as BERT or transformers can make the overall model achieves the best results. In summary, this paper will use BERT as the basis to identify the VAI of the sentence and the VA of the context.

## 3 CDVAI dataset

This paper annotates and extends the NTU irony corpus to a dimensional valence-arousal-irony, called CDVAI. NTU irony corpus provides ironic sentences and their ironic context. The annotation tasks are the VAI intensity of the sentence and the VA intensity of the context, respectively. Li and Huang (Li & Huang, 2020) analyzed the sentence structure of Chinese irony based on the existing corpus, and proposed that context is an important information for judging irony. Based on its findings and the sentence structure within the NTU irony corpus, this study 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) are required to make the literal meaning contrary to the context.

### 3.1 Dimensional VAI annotation

The paper proposes that the VAI score were rated from 1 to 5. The detailed annotation processes 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 is neutral, representing no emotion or inability to judge .
- **Arousal:** A lower arousal score indicates a lower degree of intensity. A score of 0 is an objective description of absolutely no emotion. For example: "I am a student." A score of 1 is close to an objective description, or it is more difficult to judge whether there is emotion. A score of 2 means that the annotator can feel the emotion from the sentence, but there is no emotion word. A score of 3 and above will be given to posts with explicit emotional words or phrases. Emotional words such as sad,

annoyed, lost, happy, etc. can clearly describe the emotional state. A score of 4 means that the annotator can clearly feel strong emotions in the sentence, such as madness, rage, excitement, etc. In addition to exaggerated rhetoric, posts may contain violent words, such as aggressive language. A score of 5 indicates extreme choice of words, words with discrimination, hatred, or words with obvious manic emotions. For example: "Great, the class report is going to be with that pathetic nerd!"

- **Irony:** The judging criterion of irony is as follows. 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 that constitute irony sentences. This study believes that the judgment of irony intensity can be determined according to the difference between the positive degree of irony keywords and the negative degree of context. In this paper, the positive degree of various ironic keywords appearing in the corpus is summarized as: wonderful > great > very good > good. The larger the gap between the positive degree of the ironic keyword and the negative degree of the context, the higher the degree of irony, and vice versa. A score of 1 indicates that the irony keyword of the sentence has a small gap with the context, or the context is close to an objective description. Example: Good, it's raining. A score of 2 indicates that there is a moderate gap between ironic keywords and context. A score of 3 means that there is a clear gap between the ironic keywords and the context. A score of 4 means 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

After screening the NTU Irony Corpus, the paper left a total of 1004 sentences, of which 843 sentences with irony context needed to be annotated. Each sample 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 intrinsic bias of subjectivity of different annotators, taking the average of 3 annotators as the gold standard.

This study believes that it is most reasonable to label the three indicators of VAI with a score system, because human cognition of emotional intensity is closer to continuous scores than classification. From the unavoidable bias among annotators, we know that even if the scoring criteria are well-defined, there are differences in judging the same sentence. The meaning of the labeling criterion in this study is to set the standard score line and to concretize the vague definition of intensity. Therefore, the traditional classification consensus algorithm such as kappa value does not meet the assumptions of this study, so the mean absolute error (MAE) metric is used to evaluate the annotation quality of each annotator. The MAE among the three annotators ranged from 0.05 to 0.31 in sentence-level valance, 0.25 to 0.41 in arousal, 0.22 to 0.56 in irony. The valance at the context level is between 0.07 and 0.4. Arousal is between 0.15 and 0.65. It can be seen from the above that the difference between the labeling scores of the three annotators is very small, which proves that the labeling 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**: In terms of judging the valence, this post contains extremely negative emotions, such as "*mother fucker!! Give me your salary and I will sent a notice for you!!*". It is clear that the emotions manifested by the swear words and complaints are highly negative. As for the Arousal label, the post contains a clear sense of manic and abuse language. Thus, the Arousal label is given a score of 5 points. The post carries the irony keyword "*very good*", which is a minimal positive description. However, according to the description of the sender, the incident was judged to be a serious one because it caused serious discomfort and negative emotions. In addition to the irony keyword, this sentence contains other irony 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. Since the sentiment of the context is like the whole, the dimensional VA score is the same as the sentence.

### 3.3 Statistics of annotated result

The previous chapters have described the importance of the structure and context of ironic sentences. Next, this study extends the sentiment labeling based on Tang's (Tang & Chen, 2014) open data on irony sentences, which is the only dataset with labelled irony contexts. There are a total of 1004 sentences, of which 843 sentences have context. Table 1 shows the annotated CDVAI dataset in different levels and sentiment. Since the data is mainly ironic, the valence values are all negative. Considering that there are many emotions in the sentence, the arousal focuses on points 2, 3 and 4. The arousal of context is distributed to a lower score after the irony keyword is removed, because irony keywords usually have exaggerated expressions, resulting in a higher arousal. The irony distribution is more even, but because higher scores indicate more serious factual differences, there are still relatively few high scores.

| 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 |

Table 1: Score frequency of all sentiments.

## 4 Model performance evaluation

In order to verify the labeling consistency of this study and the feasibility of irony detection, this study uses a deep learning model to predict indicators. Table 2 shows the general statistics of CDVAI dataset. As a benchmark data for deep learning model, the dataset is split using Stratified sampling to get the training, validation, and the testing set. The ratio of train set and test set is 7:3, and validation set is split from train set and the ratio is 9:1.

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

Table 2: Statistics of the proposed CDVAI dataset.

### 4.1 Prediction models

The dataset already has manually annotated contexts. Then this paper uses the pretrained BERT model as an encoder and performs VAI score prediction. The sentence will enter the encoder to obtain hidden features and to predict the VAI score through the linear layer. The VA score prediction on context-level has three approaches. (1) M1: After entering the encoder, the VA score is predicted through a linear layer. (2) M2: The position of the context in the sentence can be located, so after entering the encoder, the hidden features of the location of the context is used to enter the linear layer to predict VA score. (3) M3: Enter the sentence and context into the encoder, respectively, and concatenate two hidden features from two hidden features and enter the linear layer to predict the VA score.

This study compares four BERT models pre-trained on a Chinese corpus and two on a multilingual corpus to find the best results, such as PM1: *ckiplab/albert-base-chinese* is trained using the Zhwiki corpus; PM2: *hfl/chinese-macbert-base* uses Wikipedia simplified and

traditional Chinese as the corpus to train the model; PM3: *shibing624/macbert4csc-base-chinese* is trained using the SIGHAN typo correction corpus; and PM4: *uer/chinese_roberta_L-4_H-256* is pre-trained from UER-py. Using multilingual corpus to train models, such as PM5: *bert-base-multilingual-uncased* is trained using the Wikipedia corpus, and PM6: *flax-community/alberti-bert-base-multilingual-case* is trained using the PULPO literary poetry corpus.

## 4.2 Experimental settings

Since the annotation of the CDVAI dataset uses the irony context in it to enhance the model's understanding of contextual emotional changes, it is necessary to use the context span feature for additional context VA prediction. Thus, the paper uses BERT as the experimental mode. The parameters are shown in Table 3. In addition to predicting the VAI of the sentence, the average feature of the context span is also used to predict the VA of the context. Each pre-trained model uses the same parameters, except the learning rate value, context has a smaller starting value than the sentence, which is due to the information of the context is less than the sentence, so a smaller learning rate should be tried.

| Parameter | Value |
|---|---|
| Optimizer | Adam |
| Learning rate - sentence-level | 4e-4, 4e-5, 4e-6 |
| Learning rate - context-level | 4e-5, 4e-6, 4e-7 |
| Number of epochs | 50 |

Table 3: Parameter settings of BERT models.

## 4.3 Experimental results

The prediction performance of dimensional VAI score on sentence-level is shown in Table 4. The MAEs of all models are less than 0.4. The model has the best MAE indicator. MAE is all greater than 0.5, one possible reason for the relatively poor performance is that the determination of excitement is not limited to the intensity of the word, but is annotated according to the overall situation of the described event. Therefore, it is difficult for the model to learn the degree of arousal. The irony result is similar to arousal, the MAEs of all models are greater than 0.5. The possible reason why it is difficult for the model to learn irony is that the degree of irony is not limited

to a single ironic word, there may be multiple ironic information in a sentence.

| Model | Valence | Arousal | Irony |
|---|---|---|---|
| PM1 | **0.319** | 0.532 | 0.548 |
| PM2 | 0.347 | 0.522 | 0.522 |
| PM3 | 0.339 | 0.526 | 0.532 |
| PM4 | 0.334 | 0.532 | 0.533 |
| PM5 | 0.353 | 0.536 | 0.520 |
| PM6 | 0.332 | 0.529 | 0.526 |

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

The prediction performance of dimensional VA score on context-level is shown in Table 5. Valence of context-level also performs quite well. Regardless of the method, the MAE is around 0.45. However, it is worth noting that the second method is not better than the first method to extract the hidden features of the corresponding position of the context and sum up the prediction. However, the prediction effect of M3 after connecting the hidden features of the sentence and context is better in some model scores. However, in most results, the effect of M1 is still better. It is speculated that the reason is that M2 and M3 will increase the amount of noise when predicting the value. The overall effect of Arousal of context is worse than that of valence, and the MAE of any method is about 0.8. However, it is worth noting that M3 achieves better results in most models in arousal prediction. And the difference is very small with M1. It can be speculated that M3 considers the information of the entire sentence and context to help predict the degree of arousal. Compared with M1, M2 has little difference in the results of most models. It can be seen that M2 is less helpful for predicting the degree of arousal.

| Model | Valence | | | Arousal | | |
|---|---|---|---|---|---|---|
| | M1 | M2 | M3 | M1 | M2 | M3 |
| PM1 | 0.438 | 0.452 | 0.456 | 0.813 | 0.835 | 0.787 |
| PM2 | 0.431 | 0.451 | 0.422 | 0.819 | 0.817 | 0.793 |
| PM3 | 0.401 | 0.451 | 0.422 | 0.743 | 0.817 | 0.793 |
| PM4 | 0.417 | 0.463 | 0.438 | 0.767 | 0.857 | 0.788 |
| PM5 | 0.425 | 0.438 | 0.388 | 0.829 | 0.824 | 0.799 |
| PM6 | 0.429 | 0.433 | 0.432 | 0.847 | 0.84 | 0.806 |

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

Analysis of the above cases shows that irony detection is a challenging task. The main problem encountered by the BERT model is difficult to infer the speaker's intention. The lack of common sense carried by the model also affects its judgement. There is also no way to judge the importance and severity of the incidents described in the posts, which makes it difficult for the model to identify the level of irony accurately.

### 4.4 Error analysis

Based on the performance of BERT model, we present a few incorrect prediction cases below.

- **Example:**

  **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 6. The reason why the model judges the valence to be "1.21" 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.

|  | 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 |

Table 6: Predict results of example 1.

## 5 Conclusion

The paper introduced the CDVAI dataset which extended from NTU irony corpus. It contains multi-index sentiment annotation and irony context sentiment annotation, which is helpful for developing Chinese irony computational methods of detection that allow the model to learn sentimental changes and differences in context. The experimental results showed that the annotation of CDVAI dataset provides a learning direction for the BERT model as the basic model, so that the model can understand the irony structure. And the third method proposed in this paper, connecting the sentence with the context and then predicting, can effectively improve the effect of the model predicting arousal of context.

CDVAI dataset does not have enough examples for machine learning experiments, and there is no way to cover a considerable number of fields and complete ironic sentence patterns. Nevertheless, the paper is suitable to use as guide data to obtain more samples or as a template for labeling guidelines. Furthermore, the proposed CDVAI dataset could be combined with other ironic corpora to extend the training sample size so that machine learning algorithms would be improved in the future.

### References

Castor, A. and Pollux, L. E. (1992). The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

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

Grandchercheur, L.B. (1983). Vers une modélisation cognitive de l'être et du néant. In S.G Paris, G.M. Olson, & H.W. Stevenson (Eds.), *Fondement des Sciences Cognitives*. Hillsdale, NJ: Lawrence Erlbaum Associates (pp. 6—38).

Strötgen, J. and Gertz, M. (2012). Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)* (pp.3746–3753).

Superman, S., Batman, B., Catwoman, C., and Spiderman, S. (2000). Superheroes experiences

with books. *The Phantom Editors Associates, Gotham City, 20th edition*.

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.

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

Dimovska, J., Angelovska, M., Gjorgjevikj, D., & Madjarov, G. (2018). Sarcasm and irony detection in english tweets. *International Conference on Telecommunications* (pp. 120-131). Springer, Cham.

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.

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR), 50*(5), 1-22.

Kumar, H., & Harish, B. (2019). Automatic irony detection using feature fusion and ensemble classifier. *International Journal of Interactive Multimedia & Artificial Intelligence, 5*(7).

Lestari, W. (2019). Irony analysis of Memes on Instagram social media. *PIONEER: Journal of Language and Literature, 10*(2), 114-123.

Li, A.-R., & Huang, C.-R. (2020). A method of modern Chinese irony detection. *In 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*.

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*.

Naseem, U., Razzak, I., Eklund, P., & Musial, K. (2020). Towards improved deep contextual embedding for the identification of irony and sarcasm. *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.

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*.

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

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. *Proceedings of the 12th international workshop on semantic evaluation* (pp. 638-642).

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

Veale, T., & Hao, Y. (2010). Detecting ironic intent in creative comparisons. In *ECAI 2010* (pp. 765-770). IOS Press.

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. *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 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.