Online Opinion Conflict Interaction Recognition Based on Dependent Multi-Task Deep Learning

Xiaodong Feng[1], **Zishuai Shang**[2], **Rui-Jie Zhu**[3]
[1]Sun Yat-sen University, China
[2]University of Electronic Science and Technology of China, China
[3]UC Santa Cruz, USA
fengxd5@mail.sysu.edu.cn, shangzs@alu.uestc.edu.cn, ridger@ucsc.edu

Abstract

Expression of conflicting opinions often appears in the interaction process among different users in various online communities, and effectively recognizing opinion conflict interactions is of great significance. Compared to traditional online opinion mining tasks, such as stance/sentiment classification, the opinion conflict interaction recognition shows unique and challenging characteristics that it is determined based on the interaction between the differences in emotion expression and the consistency of thematic content given two opinions. Thus, the paper tries to propose a model that can effectively model its unique feature and recognize different interaction categories, that is, Opinion Conflict Interaction Recognition based on **C**ausal **I**nference-enhanced dependent **M**ulti-Task **D**eep **L**earning, noted as **CIMDL**. The model introduced causal inference-enhanced multi-task learning and Co-Attention interaction mechanism in addition to the pre-training language model-based text embedding and deep neural network-based feature extraction. We construct two benchmark datasets for the newly proposed task, and conduct extensive experiments to demonstrate the advantages of the proposed approach over different state-of-art baselines.

Code & Datasets — https://github.com/zss019/CIMDL
Online Appendix — https://github.com/zss019/CIMDL/blob/main/appendix.pdf

Introduction

Online knowledge communities (such as knowledge Q&A communities and science weblogs) on basis of Web2.0 and mobile Internet technologies have become important channels for knowledge acquisition, sharing, and exchange. In the free and open online platform, due to the differences in personal knowledge background and personal traits, users would express inconsistent opinions with each other, forming opinion conflict interactions. For example, in the Q&A knowledge community, conflict opinions often appear in different answers and comments on a certain question; different bloggers often express different opinions on the same knowledge viewpoint. There exists a critical impact of opinion conflict interactions on the operation of online communities as well as social governance, since it would greatly influence users' online participation and the spreading of public opinion (Yang et al. 2022). Thu, effectively identifying opinion conflict interactions among users is of great significance, which would provide useful practice implications.

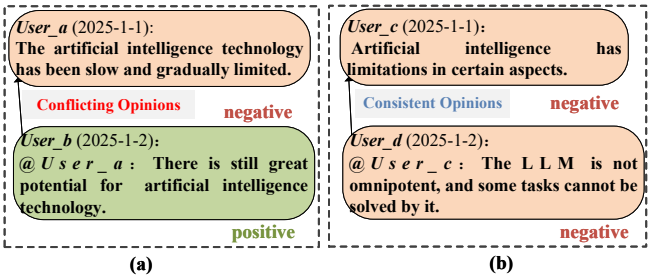

Figure 1: Examples show the uniqueness of opinion conflict interaction compared to sentiment or stance classification of a single text. (a) the positive sentiment (or support stance) towards something conflicts with the negative emotional views about the similar thing. (b) the negative sentiment (opposite stance) towards something supports the negative emotional views about the similar thing.

However, it is challenging to recognize opinion conflict interactions as it is quite different from the traditional opinion mining tasks. First, we note that there exists extensive approaches for online controversy detection (Benslimane et al. 2023; Wang et al. 2021). However, these methods primarily focus on macro-level analysis. They try to measure the overall conflict or controversy degree within a set of event-related posts or comments. Consequently, they cannot be directly applied to recognize the conflict interaction between a pair of two opinion texts from a micro perspective. Second, text semantic matching tasks (Wang et al. 2022; Yu et al. 2024b) aim to measure the matching degree between two texts. However, these methods primarily assess content or topic consistency in search and Q&A scenarios. They fail to capture semantic opinion differences toward identical or distinct entities. Third, and more importantly, opinion conflict interaction recognition differs from the single text-based opinion mining task, such as sentiment or stance classification (Zhang et al. 2024). As shown in Figure 1, the positive sentiment (or support stance) towards something can conflict with the negative emotional views about the similar thing (a), or the negative sentiment (opposite stance)

towards something would support the negative emotional views about the similar thing (b).

Overall, compared with the focus on semantic relevance between two texts in text matching tasks and the focus on emotional expression in stance detection tasks, the opinion conflict interaction presents a unique feature of combining emotional expression differences and thematic content relevance. It is not completely equivalent to emotional stance detection of the online opinion or the association of thematic content between opinions, but needs to be determined based on the combination of the differences in emotion expression and the consistency of thematic content in the given two opinions, as in Figure 1. Therefore, we can not directly employ the traditional stance detection or text matching methods, highlighting the research gap that needs to be filled. The key challenge is how to model the interaction between emotional expression and thematic content of the two opinions given, motivating the development of an effective method.

Thus, this study proposes a model to capture the unique characteristics of opinion conflict interactions among users when classifying it. We name it as Opinion Conflict Interaction Recognition based on Causal Inference-enhanced dependent Multi-Task Deep Learning, noted as **CIMDL**. The overall structure is based on end-to-end deep neural networks that have been widely used for text mining tasks recently. Specifically, the input text embedding of two opinion texts comes from the pre-trained language model. To model the dependency of emotional expression and thematic content of the two opinions on the final classification, we feed the embedding into Recurrent Neural Network (RNN) and Self-Attention to encode the sentiment feature and thematic content of the opinions. Using the multi-task learning framework, the model would fit the sentiment categories of the two input opinions in addition to the target label of opinion conflict interaction. Then, a Co-Attention interaction module is to model the interaction of the representation containing sentiment expression and thematic content between the input two opinion texts. More significantly, motivated by the logically dependent multi-task learning with causal inference (Chen et al. 2020), we introduced a causal transfer module to capture the dependency of emotional expression and semantic matching on the final conflict interaction label as discussed above (in Figure 1). Finally, it was fused with the sentence semantic representation to perform the final classification. We construct two benchmark datasets from two popular online communities in China, *Zhihu.com*[1] and *ScienceNnet.cn*[2], for the new proposed task, and conduct extensive experiments to demonstrate the advantages of the proposed approach over different baselines.

This work establishes a novel theoretical framework for opinion conflict interaction recognition by formally defining the task, introducing a causal inference-enhanced multi-task deep learning to model sentiment-content interplay, and validating its effectiveness through newly constructed benchmarks, thereby advancing research beyond traditional stance detection and semantic matching paradigms. In summary, our contributions are as follows.

- We defined a new opinion conflict interaction recognition task, which tries to recognize the conflict interaction between a pair of two user's opinions. It is quite different from existing text mining tasks, such as stance detection, semantic matching or overall controversy detection.
- We proposed an opinion conflict interaction recognition model based on causal inference-enhanced multi-task deep learning (**CIMDL**) to model the interaction between the differences in emotion expression and the consistency of thematic content given two opinions.
- We build two benchmark datasets for the newly defined task, and extensive experiments reveal the superior performance of the proposed method overall baselines.

Related Works

Controversy Detection of Online Content

Previous studies have focused on the controversy detection of online content on different platforms, such as knowledge collaborative editing platforms, news websites, and social media. According to the different types of data used, the methods can be divided into content-based methods, structure-based methods, and hybrid methods.

Content-based methods utilize the content features of the online information to identify the controversy within it, including dictionary-based methods (Pennacchiotti and Popescu 2010) and machine learning or deep learning-based methods (Das, Lavoie, and Magdon-Ismail 2016; Beelen, Kanoulas, and Van De Velde 2017; Koncar, Walk, and Helic 2021) to extract the features of the content. The structure-based approach mainly utilizes the interaction network between the online information and published users, and identifies possible controversy based on different structural features of the network, such as graph partitioning modularity (Guerra et al. 2013), graph random walk (Garimella et al. 2018) and graph motif (Coletto et al. 2017). Recent studies usually construct hybrid methods that consider both the online content and the interaction network between users or information, using traditional social network analysis methodologies (Emamgholizadeh et al. 2020), machine learning techniques (Popescu and Pennacchiotti 2010) and deep learning methods with graph neural networks (Benslimane et al. 2023; Wang et al. 2021).

However, these studies mostly focus on macro-level content controversy in online information, such as a single web page, a collection of tweets for an event/topic or comments on the same tweet. As a result, they cannot be applied directly to recognize the conflicting interaction in pairwise user opinions from the micro-perspective of user interaction.

Text Matching

Text matching mainly focuses on traditional information retrieval and Q&A context. Although our task is different from the text matching task, they share the same problem form of assigning a label given a pair of two input texts. Therefore, methods in text matching, especially feature representation and feature learning based on deep neural networks, can provide insights for the method design

[1]https://www.zhihu.com/

[2]https://sciencenet.cn/

for the new task. The early work on text matching applied traditional vector space model and text similarity upon word frequency distribution, such as TF-IDF and BM25. The emergence of text embedding, especially deep learning (Ward 2014; Lin and Lin 2023) and the pre-trained language model (such as BERT (Devlin et al. 2019) and Sentence-BERT (Reimers and Gurevych 2019)), has improved text representation and similarity computation. In addition, interaction-based deep learning models do not directly compute the similarity of text representations, but model the interaction between two text representations to enhance the performance of text matching, such as the CNN model to extract features from the interaction matrix (Pang et al. 2016), the Co-Attention (Yu et al. 2021) or dual attention (Wang et al. 2022) mechanism to extract the interaction feature, and data augmentation based on Gaussian noise and noise mask signal (Wang et al. 2024).

Problem Definition

Let $\{T_x, T_y\}$ be a pair of interactive opinion texts between two users, such as the comment and its reply in online communities. Specifically, $T_x = \{w_x^1, w_x^2, \ldots, w_x^n\}$ and $T_y = \{w_y^1, w_y^2, \ldots, w_y^m\}$. w_x^i and w_y^i denote the individual word in sequences T_x and T_y, respectively. n and m represent the number of words in T_x and T_y. The objective of the Opinion Conflict Interaction Recognition task is to train a classifier that leverages the textual content of these two opinion texts. The classifier automatically categorizes the interaction type of $\{T_x, T_y\}$, specifically recognizing whether it is a conflicting interaction or other types. The classifier outputs a label $y_{(x,y)} \in \{c_1, c_2, \ldots, c_k\}$, where c_i denotes predefined opinion interaction categories such as '*Support*', '*Conflict*', or '*Neutral*'. Here, k is the number of categories.

Methodology

Overall Architecture

The overall architecture of CIMDL is shown in Figure 2. It end-to-end recognizes opinion conflict interactions and extracts multi-dimensional features via hierarchical, specialized encoders. By integrating causal inference and multi-task learning, it leverages fine-grained sentiment insights to achieve accurate opinion conflict recognition.

In CIMDL, the Text Representation & Encoding module is first used to obtain word embeddings of input text pairs. The text pairs are also fed into the Sentence Semantic Matching (SSM) module to obtain holistic-level representations. In addition, word embedding sequences are fed into the Sentiment & Content Feature Extraction module to derive sentiment-specific and content-specific features. These features are forwarded to the Multi-task Learning with Causal Inference module, returning sentiment category and generating the causal fusion representations. The latter are then fused with the semantic matching representations from the SSM module. Finally, the fused representations are fed into the Final Classification module to output the opinion interaction category.

Text Representation & Encoding

The input texts T_x and T_y are respectively encoded into sequences of word embedding vectors by a fine-tuned RoBERTa model (Cui et al. 2021). The hidden representation from the last transformer layer of RoBERTa serves as the word-level representations, denoted as $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$ and $Y = [y_1, y_2, \ldots, y_n] \in \mathbb{R}^{d \times n}$. Here, d represents the dimensionality of word vectors (768 for RoBERTa), and n is the sequence length after augmented alignment.

Sentiment & Content Feature Extraction

Based on the word-level representations, this module extracts semantic sentiment features and thematic content features using specialized deep neural networks.

Sentiment Feature Encoder Conflicting online opinions usually convey specific sentiments. To capture these nuanced sentiments, a bidirectional GRU (Bi-GRU) network is used to encode the text representation sequence X and Y.

$$\begin{aligned} X' &= \overrightarrow{\mathrm{GRU}}(X) \oplus \overleftarrow{\mathrm{GRU}}(X), \\ Y' &= \overrightarrow{\mathrm{GRU}}(Y) \oplus \overleftarrow{\mathrm{GRU}}(Y). \end{aligned} \tag{1}$$

Here, $\overrightarrow{\mathrm{GRU}}$ and $\overleftarrow{\mathrm{GRU}}$ denote the forward and backward GRU hidden states, respectively. The symbol '$\oplus$' denotes the concatenation operation along the feature dimension. To emphasize salient sentiment information, we apply an attention-based weighted aggregation to the outputs of the Bi-GRU. The process is formalized as follows:

$$\begin{aligned} X_w &= \mathrm{softmax}(WX'), \quad X_{\text{sen}} = X' X_w^{\mathsf{T}}, \\ Y_w &= \mathrm{softmax}(WY'), \quad Y_{\text{sen}} = Y' Y_w^{\mathsf{T}}. \end{aligned} \tag{2}$$

$X_w \in \mathbb{R}^{n \times n}$ and $Y_w \in \mathbb{R}^{n \times n}$ are attention weight matrices derived from X' and Y'. The learnable $W \in \mathbb{R}^{n \times d_w}$ maps the features encoded by Bi-GRU to the attention space, where d_w is the feature dimension outputted by Bi-GRU.

Thematic Content Feature Encoder We use the Self-Attention mechanism (Vaswani et al. 2017) to capture word relationships and emphasize key content. The text representations X and Y are first mapped into their respective Q, K, and V matrices through independent linear layers. Subsequently, the attention weights are calculated as follows:

$$\mathrm{Weight} = \mathrm{softmax}\left(\frac{Q \cdot K^{\mathsf{T}}}{\sqrt{d_K}}\right). \tag{3}$$

Here, d_K denotes the dimension of the matrix K. The Weight_x and Weight_y, as variants of Weight for X and Y respectively, are then used to compute context-aware content representations X_{content} and Y_{content}:

$$\begin{aligned} X_{\text{content}} &= \mathrm{LayerNorm}(\mathrm{Weight}_x V_x), \\ Y_{\text{content}} &= \mathrm{LayerNorm}(\mathrm{Weight}_y V_y), \end{aligned} \tag{4}$$

where $\mathrm{LayerNorm}(\cdot)$ denotes layer normalization for output stabilization; V_x and V_y denote the V values after the linear layer on X and Y respectively.

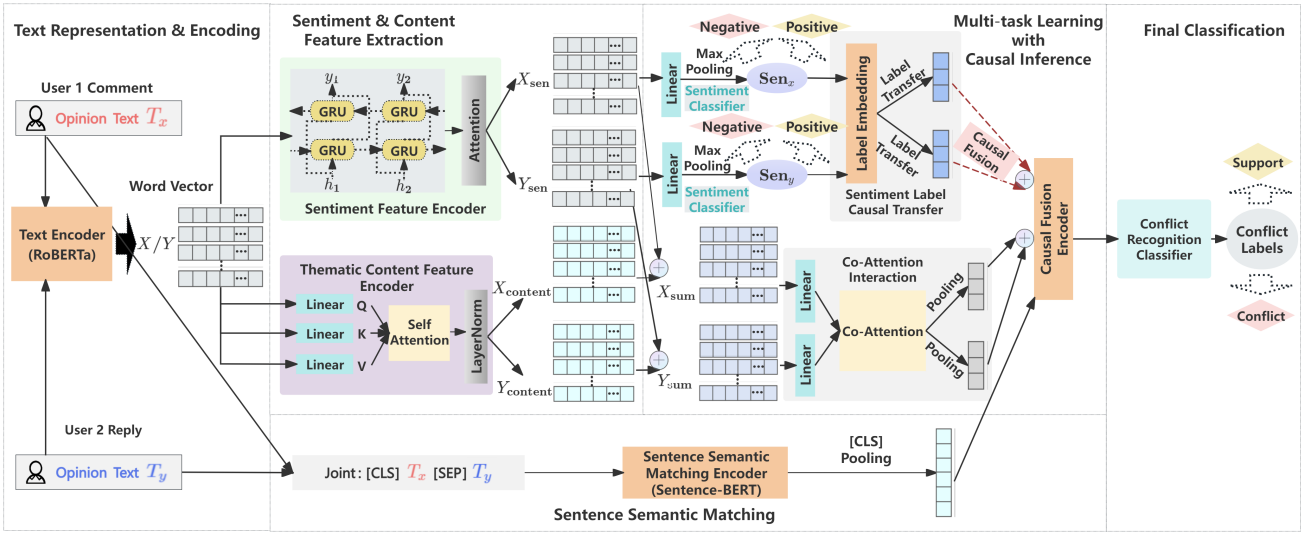

Figure 2: The overall structure of CIMDL model.

Multi-task Learning with Causal Inference

To leverage the close relationship between sentiment expression and opinion conflict interactions, while fully utilizing the sentiment and content features extracted in the earlier stage, we propose a framework that integrates causal inference within a dependent multi-task learning paradigm. The framework models the dependency between sentiment classification of each text and the final opinion conflict interaction classification, thereby providing a more nuanced understanding of opinion conflict interactions.

Multi-task learning leverages inter-task dependencies to enhance overall model performance. Specifically, we treat opinion conflict interaction recognition as the main task and sentiment classification as an auxiliary task. Additionally, we incorporate Co-Attention interaction and causal transfer components to enhance the expressivity of the model.

Sentiment Classification as an Auxiliary Task We classify the sentiment feature representations extracted from the Sentiment Feature Encoder into sentiment categories, and use the annotated sentiment labels as the supervisory signal for training the classifier (the details of the annotation are shown in the Benchmark Dataset Construction section). By leveraging the inherent relationship between sentiment classification output and the final opinion interaction label, this task contributes to improving the performance of the main task during joint model training.

Specifically, the sentiment feature vectors X_{sen} and Y_{sen} are mapped to a lower-dimensional space using a linear layer, followed by Max-Pooling to refine them into X'_{sen} and Y'_{sen}. They are then fed into a shared Multilayer Perceptron (MLP) classifier, which independently returns the sentiment category (0 or 1) for each text:

$$\begin{aligned} X'_{\text{sen}} &= \text{MaxPooling}\big(\text{Linear}(X_{\text{sen}})\big), \\ Y'_{\text{sen}} &= \text{MaxPooling}\big(\text{Linear}(Y_{\text{sen}})\big), \end{aligned} \tag{5}$$

$$\text{Sen}_x, \text{Sen}_y = \text{MLP}(X'_{\text{sen}}, Y'_{\text{sen}}). \tag{6}$$

Co-Attention Interaction To deeply mine the interactions between text pair $\{T_x, T_y\}$, we integrate the sentiment and thematic content features, and employ the Co-Attention mechanism (Tay, Luu, and Hui 2018) to capture fine-grained key interaction features at the word-sequence level. Sentiment representation matrices X_{sen} and Y_{sen} are concatenated with their corresponding content representation X_{content} and Y_{content}, along the feature dimension. The concatenated representations are then processed through a linear layer followed by GELU activation, returning X_{sum} and Y_{sum}.

$$\begin{aligned} X_{\text{sum}} &= \text{GELU}\left(\text{Linear}(X_{\text{content}} \oplus X_{\text{sen}})\right), \\ Y_{\text{sum}} &= \text{GELU}\left(\text{Linear}(Y_{\text{content}} \oplus Y_{\text{sen}})\right). \end{aligned} \tag{7}$$

Then, X_{sum} and Y_{sum} serve as Query and Value sequences alternately in Co-Attention mechanism (Lu et al. 2016; Ren et al. 2024). X_{sum} first served as Query, while Y_{sum} as both Key and Value, to compute the Co-Attention weight matrix, as:

$$\text{CoWeight}_x = \text{softmax}(Y_{\text{sum}}^{\mathsf{T}} X_{\text{sum}}). \tag{8}$$

CoWeight_x represents the relevance of the specific position of T_x on different positions of T_y. Finally, the Value sequence Y_{sum} is weighted by CoWeight_x to produce the Co-Attention output representation $\text{CoAttOut}_x \in \mathbb{R}^{d' \times n}$:

$$\text{CoAttOut}_x = Y_{\text{sum}} \text{CoWeight}_x, \tag{9}$$

where d' denotes the dimension size of X_{sum} and Y_{sum}. After that, we swap the roles of X_{sum} and Y_{sum}, to obtain the attention output representation $\text{CoAttOut}_y \in \mathbb{R}^{d' \times n}$.

Causal Transfer To further integrate sentiment representations into opinion conflict interaction recognition, we incorporate a causal transfer mechanism inspired by (Chen

et al. 2020) into multi-task learning framework to capture the dependency between sentiment labels and the opinion interaction label, i.e., the causal effects of sentiment labels on opinion interaction label. Grounded in causal inference, this mechanism models causal inter-dependencies between tasks using intermediate variables, to enhance information transfer between tasks and overall performance of the model. The mechanism consists of two key components.

• **Label Embedding** Sentiment labels Sen_x and Sen_y are embedded into dense vectors, as:

$$E_x^{\mathrm{sen}} = W_{\mathrm{sen}}\mathrm{Sen}_x^*,\ E_y^{\mathrm{sen}} = W_{\mathrm{sen}}\mathrm{Sen}_y^*. \tag{10}$$

The sentiment labels ($\mathrm{Label}_{\mathrm{sen}}$ by (14)) are converted into one-hot vectors (Sen_x^* and Sen_y^*) and then mapped to a continuous embedding space using a learnable weight matrix $W_{\mathrm{sen}} \in \mathbb{R}^{d_{\mathrm{sen}} \times C_{\mathrm{sen}}}$. Here, d_{sen} represents the embedding dimension and C_{sen} is the number of sentiment categories.

• **Label Transfer** To achieve causal fusion, the sentiment label embeddings $\mathrm{E}_x^{\mathrm{sen}}$ and $\mathrm{E}_y^{\mathrm{sen}}$ are respectively concatenated with the output of the Co-Attention interaction $\mathrm{CoAttOut}_x$ and $\mathrm{CoAttOut}_y$ after pooling (an analysis of pooling strategies is shown in the experiment section as Figure 4) to get the causal fusion representations F_x and F_y.

$$\begin{aligned} F_x &= \mathrm{Pool}(\mathrm{CoAttOut_x}) \oplus \mathrm{E_x^{sen}}, \\ F_y &= \mathrm{Pool}(\mathrm{CoAttOut_y}) \oplus \mathrm{E_y^{sen}}. \end{aligned} \tag{11}$$

Sentence Semantic Matching

Word-level vector sequences form the basis for extracting sentiment and content features, as well as the downstream Co-Attention interaction computation. However, sentence-level information is also vital for recognizing opinion conflict interactions. Prior research has shown that incorporating sentence-level representations enhances the model's ability to capture global semantics in text matching tasks (Yu et al. 2024a). Similarly, the global semantic context adds an additional layer of granularity, which is valuable for the model to comprehend the opinion-related content.

Thus, we introduced a Sentence Semantic Matching (SSM) module, which captures semantic relationships between text pairs from the sentence-level perspective, complementing the word-level interaction representation. Sentence-BERT (Reimers and Gurevych 2019), pre-trained on Chinese online community data [3], is applied as the encoder for semantic matching representation. Sentence-BERT extends BERT architecture by incorporating Siamese and triplet network for parameter updates, enabling it to capture entire semantic representations effectively.

Specifically, text sequences T_x and T_y are concatenated with a [SEP] token into T_{xy}, then encoded by Sentence-BERT. The embedding of the [CLS] token at the beginning of the sequence, denoted as XY_{cls}, is extracted and passed through a linear transformation layer to produce the sentence-level semantic matching representation SSM_{xy}.

[3]https://huggingface.co/DMetaSoul/sbert-chinese-qmc-domain-v1

Final Classification and Model Training F_x, F_y and SSM_{xy} are concatenated as the final representation vector O_{causal}, activated by GELU.

$$O_{\mathrm{causal}} = \mathrm{GELU}\left(F_x \oplus F_y \oplus \mathrm{SSM}_{xy}\right). \tag{12}$$

O_{causal} is passed through a Multi-Layer Perceptron (MLP), serving as the causal fusion encoder, to produce the final opinion interaction representation $O_{\mathrm{conflict}} \in \mathbb{R}^k$:

$$O_{\mathrm{conflict}} = \mathrm{MLP}(O_{\mathrm{causal}}). \tag{13}$$

Here, k represents the number of classes. O_{conflict} is fed into a softmax classifier to generate a probability distribution over predefined interaction categories. The category with the maximum probability, denoted as $\mathrm{Label}_{\mathrm{out}}$, represents the output label.

As shown in Figure 2, the model outputs three labels for each pair of opinion interaction texts, i.e., the opinion interaction label and the sentiment labels for the two texts. The training process optimizes the model based on these labels. While the sentiment label embeddings in (6) are derived from the ground-truth sentiment labels during training, it is the predicted labels during testing. To address this train-test discrepancy (Chen et al. 2020), we employ the Gumbel-softmax method (Lu et al. 2017) to infer the sentiment labels in unseen testing data to obtain label embedding in (10). This method approximates unknown labels by leveraging Gumbel random sampling based on the predicted probability distribution of categories, using a re-parameter trick to approximate the multinomial sampling by:

$$\mathrm{Label}_{\mathrm{sen}} = \mathrm{sofmax}((\pi_i + g)/\tau). \tag{14}$$

Here, π_i represents the prior probability of the sentiment label, i.e., the probability vector obtained from (6), g is a Gumbel-distributed random variable, and τ ($\tau > 0$) is the temperature parameter. As τ increases, $\mathrm{Label}_{\mathrm{sen}}$ transitions from a one-hot to a smoother distribution, facilitating gradient-based optimization during training while maintaining an effective label approximation in inference. During training, the sampled label ($\mathrm{Label}_{\mathrm{sen}}$) will replace the ground-truth sentiment label for label embedding in (10), and it also plays the same role during testing.

The model's loss function comprises the main loss $\mathrm{Loss}_{\mathrm{main}}$ on opinion conflict interaction recognition (main task) and the auxiliary loss $\mathrm{Loss}_{\mathrm{auxiliary}}$ on sentiment classification (auxiliary task), jointly optimizing these two tasks within the multi-task learning framework. Both losses employ cross-entropy, and the total loss is a weighted sum:

$$\mathrm{Loss}_{\mathrm{total}} = \mathrm{Loss}_{\mathrm{main}} + w_l \times \mathrm{Loss}_{\mathrm{auxiliary}}. \tag{15}$$

Where w_l is the fixed weight, and we analyze how it affects model performance in the experiment section. During training, the AdamW optimizer is used.

Experimental Results

Experimental Setup

Benchmark Dataset Construction To assess the proposed method, we constructed two datasets from two popular Chinese online knowledge communities, *Zhihu.com* and

ScienceNet.cn. *Zhihu* is the largest Q&A platform in China, and *ScienceNet* is the most influential Chinese online scientific community. We crawled users' opinion interaction records and manually annotated them. Zhihu dataset consists of comments and corresponding replies in Q&A thread on selected controversial topics. These interactions were annotated as "Support" (agreement between two opinions) and "Conflict" (disagreement or conflict between two opinions). ScienceNet dataset includes comments and replies on the web-blogs about scientific or academic topics. In addition to "Support" (S) and "Conflict" (C) , we have also annotated "Neutral" (N) for some smooth interaction since the communication in ScienceNet is smoother than in Zhihu and there exist many general discussion without obvious emotional support or conflict.

Each sample consists of a pair of sentences with three annotated labels. The opinion interaction labels were annotated by three trained students majoring in Information Systems, with disagreements resolved through majority voting, where the Cohen's Kappa values for Zhihu and ScienceNet are 0.75 and 0.67. The sentiment labels of two opinions were generated using one of the state-of-the-art Chinese sentiment analysis models from Baidu AI API [4], returning 0 (negative) or 1 (positive) for each opinion text. Zhihu dataset was split into training, validation, and test sets with 7:2:1, while 8:1:1 for ScienceNet dataset.

Table 1 summarizes dataset statistics. The average length of comments and replies in Zhihu dataset is shorter than in ScienceNet. This is mainly due to the informal and daily conversational nature of Zhihu, compared to the academic discussions on ScienceNet. Notably, despite its shorter average, the Zhihu dataset includes a higher proportion of longer samples (exceeding 300 characters) and a more skewed distribution of the text length. These distinctions provide varied evaluation scenarios for the proposed method and result in different hyperparameter settings in the experiments.

Evaluation Metrics We used several metrics for the evaluation. First, Accuracy (Acc) was to evaluate the overall classification performance. Furthermore, to address class imbalance issue, Macro-average Precision (P), Recall (R), and F1 are employed to equally weight all classes. We also report the F1 for each category to provide a detailed analysis of the model's performance across different categories.

Baselines We selected three sets of baselines for comparison. The first two sets measure opinion conflict interaction based on content matching between texts, where a fully connected layer is employed for classification after getting the pre-trained embeddings of the input texts from the language models. Prior work shows that pre-trained language models outperform traditional methods in text matching (Zhang et al. 2021; Song et al. 2022). Thus, our focus is on comparing against various pre-trained models.

- RoBERTa-based Models.
  - RoBERTa (Cui et al. 2021), the backbone for our CIMDL.
  - RoBERTa-Bi-LSTM, The RoBERTa embeddings were encoded with Bi-LSTM and the outputs were concatenated..
  - RoBERTa-Bi-LSTM-ATT, extended with a Co-Attention mechanism as the interaction module after RoBERTa-Bi-LSTM.
- More Pre-trained Language Models (PLMs).
  - Sentence-BERT (Reimers and Gurevych 2019), SimCSE (Gao, Yao, and Chen 2021). Following Reimers (Reimers and Gurevych 2019), with encoded and average-pooled interaction contents, $(u, v, |u - v|)$ is used for matching representation.
  - BERT (Devlin et al. 2019), MacBERT (Cui et al. 2020) and WoBERT (Su 2020), are used to encode the concatenated two texts with [SEP], and the embedding of [CLS] token is used for matching.
- Large Language Models (LLMs). Zero-shot learning with prompts is used to infer opinion interactions, where prompt details are shown in Online Appendix A.
  - Llama 3.3 (Grattafiori et al. 2024), as State-of-the-art open-source LLM.
  - Qwen-Max (Yang et al. 2025), DeepSeek-V3 (DeepSeek-AI et al. 2024), as SOTA Chinese LLMs.

Training Setting To accommodate the differing features of the two datasets, we tailored hyperparameters accordingly. For the Zhihu dataset, we configured a max text length of 100, 100 training epochs, and a 128-unit hidden layer. For the ScienceNet dataset, we set the max text length to 64, 60 training epochs, and a 64-unit hidden layer. Early stopping was employed to mitigate over-fitting. Training initiated with a 5e-4 learning rate, modulated by a cosine annealing scheduler that starts from a small value and linearly increases to the initial rate. The batch size is set to 128 for both. All models are built and run in a software environment comprising PyTorch 2.1.0 and CUDA 12.1, on a hardware setup featuring a 24GB RTX 3090 GPU.

Overall Performance Comparison

The overall comparison results are summarized in Table 2, which shows that the proposed CIMDL achieves the best performance for both datasets. In the first set of baselines, on the Zhihu dataset, the performance of the two extended models based on RoBERTa backbone was lower than original RoBERTa. However, our CIMDL enhances the RoBERTa encoder by integrating the hierarchical feature extraction modules and a multi-task learning framework with causal inference. This effectively improves its ability on the opinion conflict interaction recognition task, outperforming RoBERTa by 2.37% in Recall and 2.00% in F1.

Compared to the second group of baselines, which relies solely on text matching, CIMDL integrates both text matching as SSM module and sentiment analysis as the causal inference-enhanced multi-task learning. Additionally, CIMDL employs Co-Attention to extract the interaction between two texts, outperforming the overall matching

[4]https://ai.baidu.com/tech/nlp_apply/sentiment_classify

Dataset Name	Dataset size	Average length of commented text	Average length of reply text	% of Support (S)	% of Neutral (N)	% of Conflict (C)
Zhihu	9897	20.60	11.79	24.80	/	75.18
ScienceNet	18000	48.86	29.87	39.35	51.84	8.90

Table 1: Statistical information of the datasets.

Model	Zhihu						ScienceNet						
	Acc	P	R	F1			Acc	P	R	F1			
				S	C	avg				C	N	S	avg
RoBERTa	86.46	81.31	78.95	68.67	91.36	80.02	73.90	63.13	57.44	19.20	77.61	77.46	58.09
RoBERTa-LSTM	84.38	78.21	75.58	63.41	90.07	76.74	75.15	66.53	58.08	20.58	79.44	77.47	59.16
RoBERTa-LSTM-ATT	85.81	80.50	77.52	66.67	90.98	78.83	74.60	65.42	57.78	20.33	78.53	77.57	58.81
BERT	85.55	79.88	77.75	66.67	90.77	78.72	73.80	62.37	55.67	12.39	78.04	76.61	55.68
MacBERT	84.90	78.97	76.52	64.85	90.38	77.61	74.60	65.03	56.52	14.91	78.93	77.06	56.97
WoBERT	87.24	82.26	80.66	71.01	91.82	81.41	74.25	65.37	57.15	17.95	78.03	77.33	57.77
Sentence-BERT	85.16	80.56	74.28	62.50	90.75	76.62	73.00	61.52	57.16	20.30	76.97	76.44	57.90
SimCSE	85.94	80.83	77.40	66.67	91.09	78.88	74.85	65.63	58.29	21.05	78.44	78.32	59.27
Qwen-Max	73.44	60.92	59.61	37.04	83.17	60.10	55.60	54.07	**60.48**	40.20	44.23	68.98	51.14
DeepSeek-V3	73.70	63.26	63.77	44.20	82.79	63.50	56.40	52.96	60.27	36.93	49.22	70.00	52.05
llama3.3	69.79	60.06	61.64	41.41	79.65	60.53	53.40	53.16	56.49	**40.41**	38.55	66.35	48.44
CIMDL	**87.63**	**82.81**	**81.32**	**71.98**	**92.06**	**82.02**	**76.20**	**70.65**	59.50	24.27	**79.97**	**78.78**	**61.00**

Table 2: Overall Performance Comparison. The values in bold represent the best in each column.

method based on BERT, WoBERT, and MacBERT. Furthermore, the SSM module captures semantic information at a broader level, providing more comprehensive insights than Sentence-BERT and SimCSE.

CIMDL significantly outperforms large language models (LLMs) on both datasets. This shows that while LLMs exhibit advanced zero-shot learning and reasoning ability, their language understanding capabilities are still insufficient for fine-grained opinion conflict interaction identification without task-specific training. Compared to LLMs' training and fine-tuning, our model offers higher training efficiency and lower computational overhead. This highlights the effectiveness and practicality of the CIMDL method in the era of LLMs.

Extensive Experiments

Difference from Sentiment Classification Task The multi-task learning method employed in our CIMDL incorporates sentiment labels. To demonstrate the differences (shown in Figure 1) between the proposed opinion conflict interaction recognition task and typical sentiment classification/sentiment matching task, we conduct an experiment of directly inferring the opinion conflict label with annotated sentiment labels. In the first two settings, we respectively map the 'Negative' ('Positive') sentiment of T_x or T_y to the 'Conflict' ('Support') label of the opinion interaction. Also, in another sentiment matching setting, the inconsistent (consistent) sentiment labels of T_x and T_y indicated 'Conflict' ('Support') label. These sentiment-induced labels in the test set were compared with the ground-truth opinion interaction labels using the above metrics. The experiments were performed on the Zhihu dataset, as in Table 3.

Setting	Zhihu Acc	P	R	F1
Sen_x	72.04	61.98	63.13	62.43
Sen_y	72.26	57.78	56.13	56.51
Sentiment Matching	32.66	42.54	41.29	32.40
CIMDL	**87.63**	**82.81**	**81.32**	**71.98**

Table 3: Comparison to sentiment classification task on Zhihu dataset.

The results show that using sentiment label or its matching as predicting labels for opinion conflict interaction recognition falls short of task requirements. This highlights the distinct nature of sentiment classification/matching versus opinion conflict interaction at the task level, where Figure 1 shows the examples on it.

Ablation Study To evaluate the effects of different components in our model, we conducted a ablation analysis. As shown in Table 4, these variants sequentially removing the Sentence Semantic Matching (SSM), Causal Transfer (CT), and Co-Attention Interaction from CIMDL. In the last variants, all three components were removed. Results indicate that CIMDL significantly outperforms all variants, confirming the collective contribution of each component to improving opinion conflict interaction recognition.

The SSM module provides holistic semantic matching information for opinion conflict interaction recognition. Its re-

Model	Zhihu						ScienceNet						
	Acc	P	R	F1			Acc	P	R	F1			
				S	C	avg				C	N	S	avg
-w/o SSM	85.55	79.56	78.96	67.83	90.68	79.25	74.90	68.20	58.16	21.28	78.65	77.64	59.19
-w/o CT	84.11	77.41	77.63	65.34	89.70	77.52	74.85	65.24	58.25	21.43	79.12	77.39	59.31
-w/o ATT	85.03	79.74	75.20	63.49	90.58	77.04	75.70	**71.93**	58.44	21.24	79.58	78.00	59.61
-w/o above	84.90	78.72	77.53	65.88	90.30	78.09	75.15	68.06	58.88	**24.70**	79.12	77.58	60.47
CIMDL	**87.63**	**82.81**	**81.32**	**71.98**	**92.06**	**82.02**	**76.20**	70.65	**59.50**	24.27	**79.97**	**78.78**	**61.00**

Table 4: Experimental results of ablation study.

moval led to F1-score drops of 2.77% and 1.81% on the two datasets, underscoring the complementary role of the holistic-level representations to the word-level representations utilized elsewhere. The CT component integrates sentiment category representations learned from the sentiment classification task into downstream classification, via label embedding and label transfer. The Precision and F1-score on the two datasets drop by 5.40%/4.50% and 5.41%/1.69% after removing . This demonstrates the effectiveness of the causal inference-based approach, and highlights the value of sentiment information for opinion interaction recognition. The Co-Attention Interaction component captures fine-grained dependencies between interacting contents. Experimental results show that key features extracted by Co-Attention interaction significantly enhance the performance.

Effect of the Feature Extraction Method In CIMDL, we use a hierarchical and specialized feature extraction strategy, employing Bi-GRU with attention mechanisms for sentiment features, Self-Attention for thematic content features, and a pre-trained Sentence-BERT with frozen embeddings as the feature extractor for the SSM module. To validate the effectiveness of these feature extraction methods, we conducted two groups of comparative experiments. For sentiment and thematic content feature extraction, we tested three variant configurations: Bi-GRU with attention mechanisms for both features (referred to as BiGRU), Self-Attention for both features (Self-ATT), the exchange strategy (Exchange), i.e., Self-Attention for sentiment feature, Bi-GRU with attention mechanisms for thematic content. For the SSM module's feature extraction, we evaluated two additional setups, i.e., replacing Sentence-BERT with RoBERTa fine-tuned on social media dataset (RoBERTa-FT) [5] and fine-tuning Sentence-BERT on our training data (SBERT-FT).

As shown in Table 5, CIMDL outperforms all variants on the two groups of tests, demonstrating the superiority of the feature extraction design in the methodology of CIMDL.

Effect of Hyper-parameters

The Weight of Auxiliary Loss In CIMDL, we introduce a weight w_l to balance the contribution of the auxiliary task (sentiment classification) to the total loss. To assess the impact of w_l, we conducted a set of experiments on two datasets, varying w_l from 0.2 to 0.8 with interval of 0.1 while keeping other hyper-parameters fixed. The results shown in Figure 3, demonstrated that both the Accuracy and Precision improved as w_l increased from 0.2 to 0.5 on the ScienceNet dataset, but then declined when w_l increases to 0.8. A similar trend is observed in the Zhihu dataset. Thus, optimal performance is achieved with a balanced auxiliary loss weight ($w_l \approx 0.5$), effectively leveraging sentiment classification to enhance opinion conflict interaction recognition.

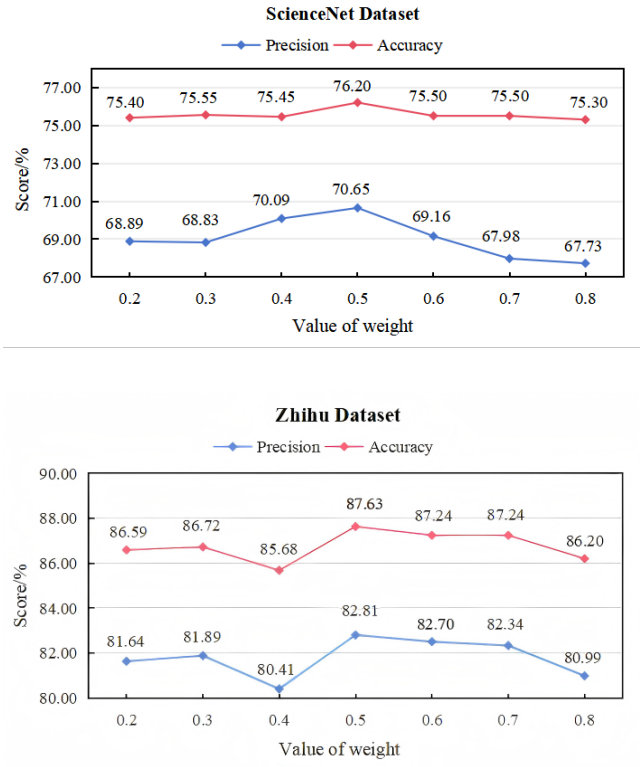

Figure 3: The effect of w_l on two datasets.

Pooling Strategy Selection In the final stage of the Causal Transfer, we applied pooling operations to the Co-Attention

[5] https://huggingface.co/Jiabo/Roberta_Chinese_sentiment

Model	Zhihu						ScienceNet						
	Acc	P	R	F1			Acc	P	R	F1			
				S	C	avg				C	N	S	avg
BiGRU	85.94	81.56	75.99	65.16	91.19	78.18	75.25	63.25	55.34	24.04	79.15	78.66	53.95
Self-ATT	86.85	82.56	78.19	68.34	91.70	80.02	75.40	67.66	58.34	20.92	79.41	78.01	59.45
Exchange	86.98	82.03	79.89	70.06	91.68	80.87	75.40	69.21	58.45	22.03	79.41	77.71	59.72
RoBERTa-FT	86.85	82.24	78.80	68.92	91.66	80.29	75.85	68.28	57.95	16.89	79.60	**79.01**	58.50
SBERT-FT	86.59	81.92	78.22	68.11	91.51	79.81	75.40	69.21	58.45	22.03	79.41	77.71	59.72
CIMDL	**87.63**	**82.81**	**81.32**	**71.98**	**92.06**	**82.02**	**76.20**	**70.65**	**59.50**	**24.27**	**79.97**	78.78	**61.00**

Table 5: The comparative experimental results with different feature extraction methods.The experimental results of the Exchange and Fine-tuned SBERT, on the ScienceNet dataset are exactly the same.

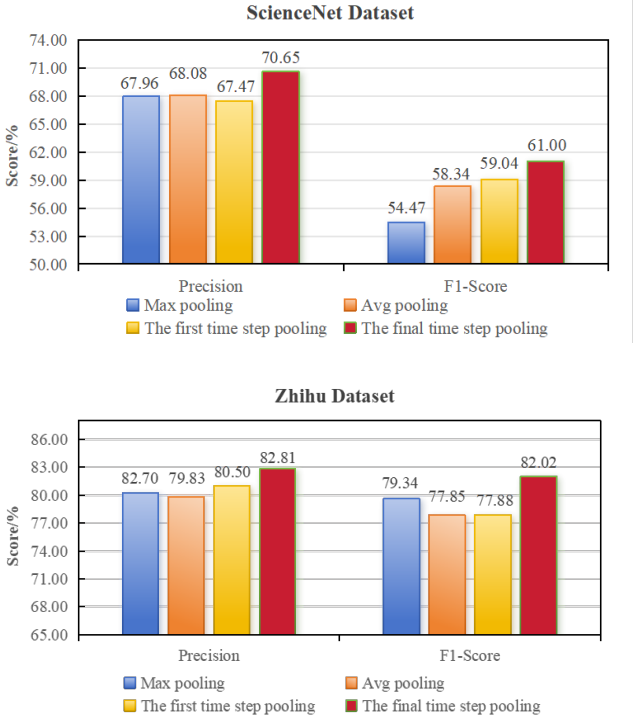

Figure 4: Performance of different pooling strategies on the two datasets.

output vectors $\mathrm{CoAttOut}_x$ and $\mathrm{CoAttOut}_y$, concatenating them with sentiment label embeddings. Specifically, we adopted a last-step pooling strategy, which extracts representation of the final time step from the sequence. To assess the effect of different pooling strategies on model performance, we compared four pooling strategies, i.e., Max pooling, Average pooling, First-step pooling (extract the [CLS] token), and the Last-step pooling. Figure 4 compared the Precision and F1-score for each strategy. Across both datasets, the Last-step pooling consistently achieved the highest Precision and F1-score. However, Max pooling and Average pooling were found to discard some fine-grained interaction features captured by the Co-Attention mechanism. Furthermore, the [CLS] token pooling introduced redundancy when combined with the SSM_{xy}, as both focus on holistic-level representations. Compared to these alternatives, the last-step representation effectively retained fine-grained features from Co-Attention while complementing the holistic semantic information from SSM_{xy}, achieving a balanced and non-redundant integration of hierarchical representations.

Case Analysis

To demonstrate the efficacy of CIMDL in recognizing opinion conflict interactions, we also present several representative cases from the ScienceNet dataset, shown in Online Appendix B. The selected typical cases showed that by integrating sentiment analysis and content understanding, CIMDL effectively discerns various categories of opinion interactions and outperforms the models that rely solely on text matching.

Social Implication

At the platform level, our proposed task and method enable the recognition of comment pairs that exhibit opinion conflict interactions. This capability allows online platforms to detect emerging interpersonal conflicts in a timely manner. By facilitating early intervention, platforms can potentially reduce the escalation of disputes and maintain a harmonious community environment. Furthermore, the automatic recognition of such interactions can help prioritize content for human moderators, thereby reducing manual review efforts and improving the efficiency of moderation workflows.

From a broader societal perspective, the generalizability of our method—without reliance on topic- or event-specific training data—enables cross-topic analysis of conflict interaction patterns and suggests a potential application in identifying emerging social controversies through large-scale discourse analysis. Such insights could support public institutions and policymakers in detecting contentious public issues at an early stage and in informing responsive governance strategies. Our work also provides insight for the growing body of research that leverages machine learning and AI-based methods to understand social dynamics in digital spaces.

Conclusion

This study theoretically advances opinion analysis by formally defining opinion conflict interaction as a novel

task and establishing a causal multi-task learning framework that captures the interplay between emotional divergence and thematic consistency, moving beyond traditional stance/sentiment classification and text matching paradigms. We propose a novel dependent multi-task deep learning model for this unique task. It extracts multi-level representations and leverages microscopic sentiment insight to assist in recognizing opinion interaction categories, within a causal inference-enhanced multi-task learning framework. The proposed model demonstrates superior performance over state-of-the-art baselines (including PLMs and LLMs) on newly constructed benchmarks.

This research enables more precise detection of opinion conflicts in online communities through its emotion-content interaction analysis, offering community managers and public opinion managers a scalable tool to proactively identify escalating conflicts, mitigate toxic interactions, and monitor conflict trends. Given the practical significance, this work has limitations on the incorporation of more semantic elements in the opinion and potential application based on it, and future work will incorporate additional micro-level information, beyond sentiment, and extend the application to real-world scenarios such as public opinion analysis and online user behavior modeling.

Acknowledgements

The work is supported by The National Social Science Fund of China under grant no.25BTQ025.

References

Beelen, K.; Kanoulas, E.; and Van De Velde, B. 2017. Detecting controversies in online news media. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1069–1072.

Benslimane, S.; Azé, J.; Bringay, S.; Servajean, M.; and Mollevi, C. 2023. A text and GNN based controversy detection method on social media. *World Wide Web*, 26: 799–825.

Chen, W.; Tian, J.; Xiao, L.; He, H.; and Jin, Y. 2020. Exploring logically dependent multi-task learning with causal inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2213–2225.

Coletto, M.; Garimella, K.; Gionis, A.; and Lucchese, C. 2017. A motif-based approach for identifying controversy. In *Proceedings of the international AAAI conference on web and social media*, volume 11, 496–499.

Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; and Hu, G. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 657–668. Association for Computational Linguistics.

Cui, Y.; Che, W.; Liu, T.; Qin, B.; and Yang, Z. 2021. Pre-Training With Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3504–3514.

Das, S.; Lavoie, A.; and Magdon-Ismail, M. 2016. Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion. *ACM Transactions on the Web (TWEB)*, 10: 1–25.

DeepSeek-AI; Liu, A.; Feng, B.; Xue, B.; and et al. 2024. DeepSeek-V3 Technical Report. ArXiv:2412.19437 [cs].

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Emamgholizadeh, H.; Nourizade, M.; Tajbakhsh, M. S.; Hashminezhad, M.; and Esfahani, F. N. 2020. A framework for quantifying controversy of social network debates using attributed networks: biased random walk (BRW). *Social Network Analysis and Mining*, 10: 90.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910. Association for Computational Linguistics.

Garimella, K.; Morales, G. D. F.; Gionis, A.; and Mathioudakis, M. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1: 1–27.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; and et al. 2024. The Llama 3 Herd of Models. ArXiv:2407.21783 [cs].

Guerra, P.; Meira Jr, W.; Cardie, C.; and Kleinberg, R. 2013. A measure of polarization on social media networks based on community boundaries. In *Proceedings of the international AAAI conference on web and social media*, volume 7, 215–224.

Koncar, P.; Walk, S.; and Helic, D. 2021. Analysis and prediction of multilingual controversy on reddit. In *Proceedings of the 13th ACM Web Science Conference 2021*, 215–224.

Lin, S.-C.; and Lin, J. 2023. A Dense Representation Framework for Lexical and Semantic Matching. *ACM Transactions on Information Systems*, 41.

Lu, J.; Kannan, A.; Yang, J.; Parikh, D.; and Batra, D. 2017. Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 289–297.

Pang, L.; Lan, Y.; Guo, J.; Xu, J.; Wan, S.; and Cheng, X. 2016. Text matching as image recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Pennacchiotti, M.; and Popescu, A.-M. 2010. Detecting controversies in Twitter: a first study. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics in a world of social media*, 31–32.

Popescu, A.-M.; and Pennacchiotti, M. 2010. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, 1873–1876.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Ren, G.; Diao, L.; Guo, F.; and Hong, T. 2024. A co-attention based multi-modal fusion network for review helpfulness prediction. *Information Processing & Management*, 61: 103573.

Song, J.; Liang, D.; Li, R.; Li, Y.; Wang, S.; Peng, M.; Wu, W.; and Yu, Y. 2022. Improving Semantic Matching through Dependency-Enhanced Pre-trained Model with Adaptive Fusion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 45–57. Association for Computational Linguistics.

Su, J. 2020. WoBERT: Word-based Chinese BERT model - ZhuiyiAI. Technical report.

Tay, Y.; Luu, A. T.; and Hui, S. C. 2018. Hermitian Co-Attention Networks for Text Matching in Asymmetrical Domains. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 4425–4431. International Joint Conferences on Artificial Intelligence Organization.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, H.; Song, X.; Zhou, B.; Wang, Y.; Gao, L.; and Jia, Y. 2021. MSSF-GCN: Multi-scale Structural and Semantic Information Fusion Graph Convolutional Network for Controversy Detection. In *Web Information Systems Engineering–WISE 2021: 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, Proceedings, Part I 22*, 394–402. Springer.

Wang, S.; Liang, D.; Song, J.; Li, Y.; and Wu, W. 2022. DABERT: Dual Attention Enhanced BERT for Semantic Matching. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1645–1654.

Wang, Y.; Zhang, B.; Liu, W.; Cai, J.; and Zhang, H. 2024. STMAP: A novel semantic text matching model augmented with embedding perturbations. *Information Processing & Management*, 61: 103576.

Ward, R. 2014. Semantic modelling with long-short-term memory for information retrieval. *arXiv preprint arXiv:1412.6629*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; and et al. 2025. Qwen2.5 Technical Report. ArXiv:2412.15115 [cs].

Yang, S.; Brossard, D.; Scheufele, D. A.; and Xenos, M. A. 2022. The science of YouTube: What factors influence user engagement with online science videos? *PLOS ONE*, 17(5): 1–19.

Yu, C.; Xue, H.; Jiang, Y.; An, L.; and Li, G. 2021. A simple and efficient text matching model based on deep interaction. *Information Processing & Management*, 58: 102738.

Yu, H.; Pan, W.; Fan, X.; and Li, H. 2024a. Multi-Granularity Fusion Text Semantic Matching Based on WoBERT. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 11766–11775. ELRA and ICCL.

Yu, L.; Liu, B.; Lin, Q.; Zhao, X.; and Che, C. 2024b. Semantic Similarity Matching for Patent Documents Using Ensemble BERT-related Model and Novel Text Processing Method. *arXiv preprint arXiv:2401.06782*.

Zhang, C.; Zhou, Z.; Peng, X.; and Xu, K. 2024. Doubleh: Twitter user stance detection via bipartite graph neural networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1766–1778.

Zhang, K.; Wu, L.; Lv, G.; Wang, M.; Chen, E.; and Ruan, S. 2021. Making the Relation Matters: Relation of Relation Learning Network for Sentence Semantic Matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35: 14411–14419.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see the second half of the abstract and the end of the Introduction section.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, the Methodology Section shows the main claims in detail.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes, see the second paragraph in the Conclusion section.
   (f) Did you discuss any potential negative societal impacts of your work? NA
   (g) Did you discuss any potential misuse of your work? NA
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we have provided an anonymous GitHub code repository to comply with anonymity in the footnote on Page 1.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see Experimental Setup in the Experiments Result section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, error bars were not reported because the experiments were run with a fixed random seed to ensure reproducibility, and multiple runs with different seeds were not conducted.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see Training Setting in the Experiments Result section.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, we have proven the effectiveness of the proposed method through comparative experiments with baseline, ablation learning, feature extraction method analysis and hyper-parameters analysis, see the Experiments Result section.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see Case Analysis in the Online Appendix B section.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, all public code and models used are cited in the paper.
   (b) Did you mention the license of the assets? No, as we ae using publicly available codes.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, we have provided an anonymous GitHub code repository including our new dataset and model in Page 1.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, we crawl the data following the robot protocol.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA, in the released data, we dele let all the personal information.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? We share the URL to the dataset an code repository, sticking to the FAIR guidelines.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, it includes a detailed datasheet file.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA