CodEOE: A Benchmark for Jointly Extracting Cross-Document Events and Opinions from Social Media

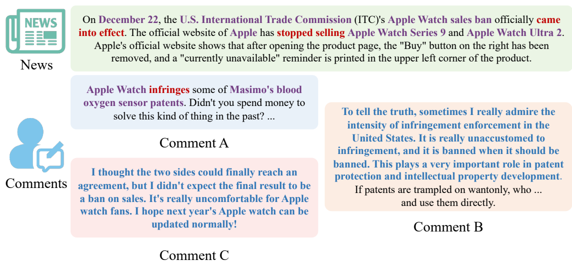
Anonymous ACL submission

Abstract

Event extraction (EE) and opinion or sentiment analysis have been extensively studied within recent decades, but their joint research remains an under-explored area. To bridge the gap in event-level opinion and sentiment analysis, we introduce the Cross-Document Event-Opinion Extraction (CodEOE) task, which aims to capture complex event-opinion interactions and jointly extract event triggers, event arguments, opinions and sentiment polarities towards events from multiple documents. The CodEOE task requires a model extracting trigger-argument pairs and trigger-opinion-sentiment triplets by understanding cross-document contexts. We manually construct a high-quality bilingual CodEOE dataset in both Chinese and English with 6,000+ trigger-argument pairs and 4,000+ trigger-opinion-sentiment triplets. We develop an end-to-end model based on the grid-tagging method to benchmark the task, which can effectively perform cross-document context understanding and achieve pair and triplet prediction. The results of our model surpass those of two strong baselines and are comparable to large language models. We hope that this new benchmark will advance research on event-level opinion and sentiment analysis. Our data and code are available here for peer review.

1 Introduction

Extracting structured information from unstructured text is a fundamental challenge in natural language processing (NLP). Event extraction (EE) (Doddington et al., 2004), a pivotal task in this domain, aims to transform unstructured text into trigger-argument structures. Simultaneously, the need for machines to understand human opinions and sentiments has driven extensive research in sentiment analysis (McDonald et al., 2007; Cambria et al., 2017). Aspect-based sentiment analysis (ABSA), a prominent subfield, focuses on detecting



(a)

Event Trigger: came into effect			
Arguments	December 22	U.S. International Trade Commission	Apple Watch sales ban
Type	Date	Organization	Other
Event Trigger: stopped selling			
Arguments	Apple	Apple Watch Series 9	Apple Watch Ultra 2
Type	Organization	Object	Object
Event Trigger: infringes			
Arguments	Apple Watch	Masimo's blood oxygen sensor patents	
Type	Object	Other	

Event Trigger: came into effect	
Opinion	I thought the two sides could finally reach an agreement... It's really uncomfortable for Apple watch fans. I hope next year's Apple watch can be updated normally!
Sentiment	Negative
Opinion	To tell the truth, sometimes I really admire the intensity of infringement enforcement in the United States. It ... patent protection and intellectual property development.
Sentiment	Positive

(b)

Figure 1: An example of the cross-document event-opinion extraction (CodEOE) task. (a) illustrates a news document along with multiple user comments. Event triggers and arguments as well as related opinions are highlighted in red, purple and blue. (b) presents the extracted event triggers, their corresponding arguments and types, as well as event-specific opinion expressions and sentiment polarities across all documents.

fine-grained sentiment orientations toward specific targets.

Existing EE research primarily focuses on sentence-level and document-level tasks. For sentence-level event extraction (SEE), the ACE2005 dataset (Doddington et al., 2004) has been widely used, establishing EE as a mainstream task and inspiring numerous studies (Chen et al., 2015; Liu et al., 2018; Wadden et al., 2019; Wang et al., 2022). To overcome sentence boundary limitations, document-level event extraction (DEE) has further advanced this field. Datasets like RAMS (Ebner et al., 2020) and WIKIEVENTS (Li et al., 2021) catalyze a wave of research, including span-based (Liu et al., 2017; Zhang et al., 2020; Liu et al., 2023) and generation-based

(Li et al., 2021; Wei et al., 2021) approaches. Gao et al. (2024) proposes the Cross-Document Event Extraction (CDEE) task, extending EE to the cross-document level. However, these studies predominantly emphasize isolated event information, overlooking the equally valuable exploration of opinions and sentiments expressed about events, particularly in social media contexts.

Similarly, ABSA originates with aspect-based sentiment analysis (Tang et al., 2016; Fan et al., 2018; Li et al., 2019). Opinion terms and category elements are later incorporated, leading to triplet and quadruple extraction of ABSA (Chen et al., 2022; Li et al., 2024; Cai et al., 2021; Zhang et al., 2021; Fei et al., 2022; Li et al., 2023). However, these works mainly focus on sentiment analysis of specific aspects, limiting their application to domains like restaurants, laptops, and mobile phones. In contrast, events convey richer information than noun-based aspects and are prevalent on social media. Compared to ABSA, event-level opinion and sentiment analysis have broader applications.

In this paper, we address the gap in event-level opinion and sentiment analysis by proposing the task of **Cross-Document Event-Opinion Extraction (CodEOE)**. The objective of CodEOE task is to detect event trigger-argument pairs and event trigger-opinion-sentiment triplets from a given news article and multiple associated comment documents. As illustrated in Figure 1, multiple social media users provide comments on a hot news. The task aims to extract eight trigger-argument pairs, such as (‘*came into effect*’, ‘*December 22*’) and (‘*stopped selling*’, ‘*Apple*’), and two trigger-opinion-sentiment triplets, such as (‘*came into effect*’, ‘*I thought ... updated normally!*’, ‘*Negative*’).

This design is motivated by two considerations: **1)** Event arguments associated with a trigger are typically numerous and unordered. Directly constructing trigger-argument-opinion-sentiment quadruples would exacerbate data sparsity. **2)** The relations between arguments and opinions for an event are relatively loose, making their direct coupling in a quadruple less suitable. Our event-centered design ensures that the extracted opinions and sentiments are grounded in specific events.

To support this task, we manually annotate a novel CodEOE dataset. We collect approximately 30,000 trending news items from Chinese social media, each item with a news article and several associated comment documents. Two experienced annotators follow an iteratively refined guideline to label key elements, including event triggers, arguments, opinions, and sentiment polarities, ensuring systematic and consistent annotations. After rigorous data selection, the dataset comprises 865 news articles and 6,236 associated comments. To enhance multilingual benchmarks for event-opinion analysis, we translate the Chinese corpus into English and re-annotate it. The dataset statistics show that each news article involves approximately seven comment documents, three events, and five opinions on different events.

To benchmark the CodEOE task, we propose an end-to-end framework, which integrates an interactive attention module and cross-document relative distance encoding. We employ a grid-filling method (Wu et al., 2020) for pair and triplet decoding and adopt a multi-task learning strategy with weighted loss functions. Experimental results demonstrate that our model achieves comparable performance to two large language models (LLMs), providing a strong baseline for future research.

To sum up, this work contributes in threefold:

- We extend opinion and sentiment analysis to the event level by introducing the cross-document event-opinion extraction (CodEOE) task, which includes trigger-argument pair and trigger-opinion-sentiment triplet extraction.
- We construct a high-quality bilingual CodEOE dataset in both Chinese and English, addressing the research gap in event-level opinion and sentiment analysis.
- We propose an end-to-end framework to benchmark this task. Our method achieves comparable performance to large language models on the CodEOE task, effectively handling multi-document contexts and recognizing intricate event-opinion relations.

2 Related Work

2.1 Event Extraction

Event extraction can be categorized into sentence-level, document-level, and cross-document level. For sentence-level event extraction (SEE), Automatic Content Extraction (ACE2005) (Doddington et al., 2004) has facilitated numerous breakthrough studies (Lu et al., 2021; Liu et al., 2018; Xu et al., 2023; Wadden et al., 2019; Wang et al., 2022). Later, Deng et al. (2022) proposed the Title2Event dataset, applying open event extraction (OpenEE) to news headlines for the first time.

The latest attention has been placed on document-level event extraction (DEE). Ebner et al. (2020) introduced the Roles Across Multiple Sentences (RAMS) dataset. Li et al. (2021) proposed a new document-level event extraction benchmark dataset, WIKIEVENTS. The mainstream methods for DEE typically include span-based methods (Liu et al., 2017; Zhang et al., 2020; Yang et al., 2023; Liu et al., 2023) and generation-based methods (Li et al., 2021; Du et al., 2021; Wei et al., 2021). Recently, prompt-based (Ma et al., 2022; Nguyen et al., 2023; Liu et al., 2024; He et al., 2023; Zeng et al., 2022) and QA-based methods (Liu et al., 2020; Li et al., 2020; Du and Cardie, 2020; Lu et al., 2023; Hong and Liu, 2024) have also been employed to guide models in event extraction. Moreover, Gao et al. (2024) introduced the Cross-Document Event Extraction (CDEE) task.

2.2 Opinion Mining and Sentiment Analysis

Opinion mining and sentiment analysis (SA) are pivotal research topics in the NLP community, particularly the ABSA task. The original ABSA task aimed at classifying the sentiment polarity of given aspects (Tang et al., 2016; Fan et al., 2018; Li et al., 2019). Subsequently, researchers proposed various composite ABSA-related tasks, such as aspect-opinion pair extraction (Zhao et al., 2020; Wu et al., 2021), aspect sentiment triplet extraction (Peng et al., 2020; Chen et al., 2021, 2022; Li et al., 2024), and structured opinion mining (Shi et al., 2022; Wu et al., 2022). To further refine ABSA tasks, aspect-category-opinion-sentiment quadruple extraction (Cai et al., 2021; Zhang et al., 2021; Fei et al., 2022) and comparative opinion quintuple extraction (Liu et al., 2021) have also garnered considerable attention. Recently, Li et al. (2023) introduced the dialogue-level aspect-based sentiment quadruple extraction task. Furthermore, some works focus on event-based sentiment analysis without opinion terms (Zhou et al., 2013; Jagdale et al., 2016; Petrescu et al., 2019; Zhang et al., 2022).

3 Data Construction

To further analyze the relations between events and opinions, we construct a new dataset sourced from Weibo hot searches. This dataset is designed to jointly analyze the triggers and arguments of events, as well as the opinion clauses in news articles or comments.

3.1 Data Collection and Preprocessing

To facilitate event-oriented opinion analysis, we construct a new dataset to promote the task of joint extraction of events and opinions. The original data is collected from Weibo[1], China's largest social media platform. Considering the timeliness, importance, and social impact of news events, we select posts and comments related to major news events from Weibo's trending topics, totaling about 30,000 hot search data entries. Each entry includes a news article and several related comments, ranging from December 2023 to July 2024. Initially, we exclude news that does not contain real-world event information, such as discussions on event topics, government reports, and personal statements. Comments containing commercial advertisements, spam content, personal attacks, or other discourse unrelated to the core event theme are filtered out to ensure the relevance of textual analysis. Subsequently, we normalize the expressions in the news and comments, identifying abusive or inappropriate remarks through manual inspection. We limit the maximum number of comments per news article to 20 to achieve better controllable modeling. After rigorous data cleaning, we obtain a final dataset comprising 865 news articles and their 6,236 related comments.

3.2 Annotation Framework

We summarize some crucial parts of the annotation standards, mainly divided into event annotation and opinion annotation. The details about the annotation standard are shown in Appendix §A.1.

The annotation process is carried out by two experienced graduate students, who are familiarized with the specific requirements and complexities of the event extraction task through specialized training. The annotation work follows a set of detailed guidelines[2] that has been iteratively optimized, clearly defining key elements such as event triggers, event arguments, opinions, and sentiment polarities to ensure systematic and consistent annotations.

The annotators strictly adhere to these guidelines during the annotation process, precisely identifying and categorizing event and opinion information in the text. Additionally, to ensure the quality of the annotations, we implement strict quality control measures, including but not limited to double an-

[1]https://weibo.com/
[2]https://anonymous.4open.science/r/CodEOE-08BD

Table 1: Data statistics of CodEOE. 'Com.' refers to comment. 'Tri.', 'Arg.' and 'Opi.' refer to event trigger, event argument and opinion terms, respectively. 'Tri-Arg' refers to trigger-argument pairs. 'Tri-Opi-Senti' refers to trigger-opinion-sentiment triplets.

		Documents		Items			Pairs & Triplets	
		News	Com.	Tri.	Arg.	Opi.	Tri-Arg	Tri-Opi-Senti
ZH	train	690	5,011	2,011	4,721	3,654	5,167	3,654
	valid	88	612	241	565	444	623	444
	test	87	613	248	635	441	705	441
	total	865	6,236	2,500	5,921	4,539	6,495	4,539
EN	train	681	4,870	1,990	4,672	3,562	5,139	3,562
	valid	87	594	234	563	429	625	429
	test	87	608	253	633	448	704	448
	total	855	6,072	2,477	5,868	4,439	6,468	4,439

notations and random checks, as well as regular annotation review meetings. The annotation process is divided into span-level and relation-level steps, which is shown in Appendix §A.2.

To ensure annotation quality at both span and relational levels, we adopt a two-stage evaluation. For span consistency, the Cohen's Kappa score reaches **0.95** through exact span boundary alignment. We also calculate the Cohen's Kappa score across all pairs and triplets, which is **0.83**, indicating a high level of consistency in our annotated corpus. For instances with inconsistent annotations, we determine the final annotation results through detailed consistency check meetings conducted by a third expert with extensive experience. Furthermore, We construct an English version of the dataset. The details are shown in Appendix §C.

3.3 Data Analysis

We randomly divide the corpus into train/valid/test sets by the number of news articles, in the ratio of 8:1:1. As shown in Table 1, the Chinese version of the dataset contains 865 news articles, 6,236 comments, 6,495 trigger-argument pairs, and 4,539 trigger-opinion-sentiment triplets. The English version of the dataset includes 855 news articles, 6,072 comments, 6,468 trigger-argument pairs, and 4,439 trigger-opinion-sentiment triplets.

To comprehensively assess the characteristics of our dataset, we conduct a detailed statistical analysis. As shown in Table 2, in the Chinese dataset, each cross-document instance (consisting of a news article and its related comments) contains an average of 2.89 event triggers, 6.84 event arguments, and 5.25 opinions. Correspondingly, the English dataset instances contain an average of 2.9 event triggers, 6.86 event arguments, and 5.19 opinions. These statistics highlight the multi-event and multi-opinion nature of our dataset, posing challenges for the development and evaluation of complex information extraction models. More data statistics about polarity and topic distribution are shown in Appendix §B.

Table 2: Statistics related to triggers, arguments, opinions and their lengths. All lengths refer to the numbers of words. 'Com.' represents comment. 'per ins.' represents each instance with one news and several comments.

	ZH	EN
	Train / Valid / Test	Train / Valid / Test
News min len.	17 / 33 / 18	13 / 26 / 16
News max len.	494 / 409 / 453	398 / 351 / 344
News avg len.	159.39 / 166.17 / 154.42	131.08 / 136.59 / 129.98
Com. max len.	506 / 446 / 444	371 / 323 / 377
Com. avg len.	51.92 / 54.12 / 52.55	43.53 / 46.63 / 42.14
Tri. avg len.	2.76 / 2.62 / 2.62	1.61 / 1.5 / 1.59
Tri. per ins.	2.91 / 2.74 / 2.85	2.92 / 2.69 / 2.91
Arg. avg len.	4.65 / 4.7 / 4.66	3.26 / 3.21 / 3.23
Arg. per ins.	6.84 / 6.42 / 7.3	6.86 / 6.47 / 7.28
Opi. avg len.	32.24 / 32.05 / 32.24	28.20 / 28.05 / 28.15
Opi. per ins.	5.29 / 5.03 / 5.07	5.27 / 4.92 / 5.15

4 Methodology

We present an end-to-end model to accomplish the CodEOE task based on the grid-tagging method. Figure 2 shows an overview of the overall architecture of our end-to-end CodEOE framework.

4.1 Problem Definition

Given a news text N and a set of comments $C = \{c_1, c_2, \ldots, c_k\}$, we define a document set $D = \{N, C\}$ as input, where the number of the document set in D is $k+1$. Let T denote the set of event triggers, A the set of event arguments, O the set of opinions, and P the set of sentiment polarities. An event trigger t_i ($t_i \in T$) or event argument a_i ($a_i \in A$) consists of one word or multiple consecutive words within a sentence, while an opinion o_i ($o_i \in O$) includes one or more consecutive clauses. The sentiment of an opinion is denoted by p_i ($p_i \in P$), where $P = \{POS, NEU, NEG\}$, with POS, NEU, and NEG representing the positive, neutral and negative sentiments expressed by opinion o_i towards the event trigger t_j, respectively. The goal of the CodEOE task is to extract trigger-argument pairs $TAP = \{(t_i, a_i)\}_{i=1}^{|TAP|}$ and trigger-opinion-sentiment triplets $TOST = \{(t_i, o_i, p_i)\}_{i=1}^{|TOST|}$, where $|TAP|$ and $|TOST|$ denote the number of trigger-argument pairs and trigger-opinion-sentiment triplets, respectively.

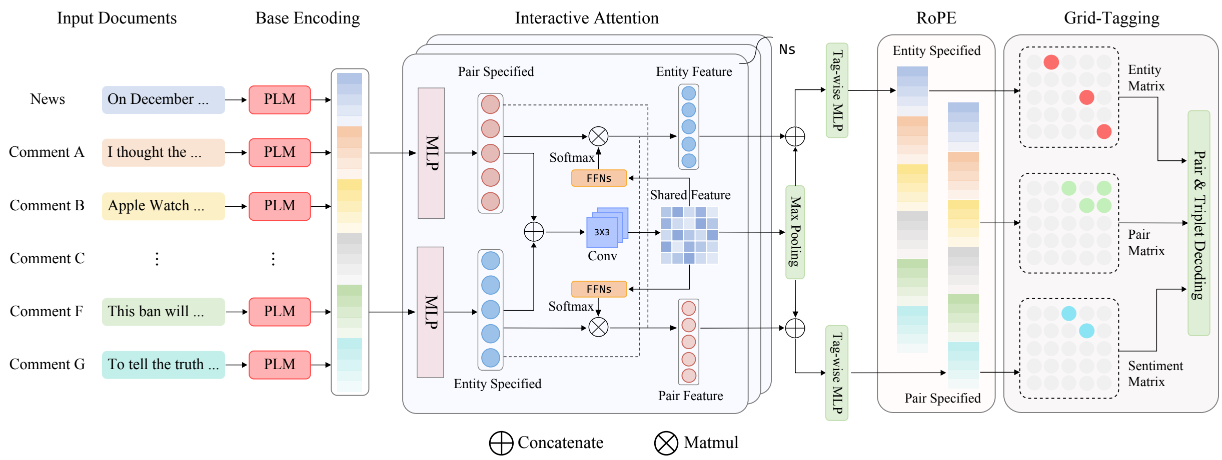


Figure 2: The overall framework of our CodEOE model. The base encoder first learns the base contextual representations of multiple input documents. The interactive attention module then captures task-specific features for span and pair. We further integrate rotary position embeddings (RoPE) to better understand cross-document relations. Finally, the system decodes all pairs and triplets based on grid-tagging labels.

4.2 Base Encoding

We utilize a pre-trained language model (PLM), such as BERT (Devlin et al., 2019), to encode the document set D. Since the length of D may significantly exceed the maximum length that BERT can handle, we encode each document d_i ($d_i \in D$) individually using separate PLMs. Specifically, we represent document d_1 ($d_1 = N$) as the news text, and d_j ($d_j \in C$) as the related comments. To prevent cross-document span extraction, we use [CLS] and [SEP] tokens to separate each document $d_i = \{w_1, w_2, \ldots, w_n\}$, n is the length of document d_i, and w_j represents the j-th token of d_i.

$$d_i' =< [\text{CLS}], w_1, w_2, ..., w_n, [\text{SEP}] >, \quad (1)$$

$$\boldsymbol{H}_i = \text{PLM}(d_i') = \boldsymbol{h}_{cls}, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_n, \boldsymbol{h}_{sep}, \quad (2)$$

where $\boldsymbol{h}_n$ is the contextual embedding of w_n.

4.3 Interactive Attention Module

The CodEOE task extracts TAP and $TOST$ through two steps: Entity Span Recognition (ESP) and Pair Recognition (PR), conceptualized as joint entity-relation extraction leveraging independent and shared feature spaces (Yan et al., 2021).

Firstly, we employ two independent MLPs to initialize the task-specific representations for ESP and PR. It is noteworthy that our Interactive Attention Module consists of L identical layers. The first layer uses initialized text embeddings, while the initial features of other layers come from the feature representations of the previous layer. We denote E, P, S as the entity feature, pair feature and shared feature, respectively. For simplicity, we define $F \in \{E, P\}$.

$$\boldsymbol{H}_{(l)}(F) = \begin{cases} \text{MLP}(\boldsymbol{H}), & l = 1 \\ \text{MLP}(\boldsymbol{\zeta}_{(l-1)}(F)), & else \end{cases} \quad (3)$$

where $\boldsymbol{H}_{(l)}(F) \in \mathbb{R}^{M \times d_{model}}$ represents the task-specific representation in the l-th layer. $\boldsymbol{H}$ is the representation of the multi-document sequence obtained by concatenating token representations of each document ($\boldsymbol{H}_i$ in Eq. (2)). $\boldsymbol{\zeta}_{(l-1)}(F)$ represents the updated task-specific representations from the previous layer, which are defined in Eq. (6). M is the token-level length of D. d_{model} is the dimension of the encoded hidden layer. l indicates the depth of the interactive attention module, and L represents the number of layers of the interactive attention module. $l \in \{1, 2, \ldots, L\}$.

We combine the task-specific representations of the two subtasks, and then obtain a shared feature map through a 3×3 convolutional layer.

$$\boldsymbol{H}_{(l)}(S) = \begin{cases} \text{Conv}([\boldsymbol{H}_{(l)}(E; P)]), & l = 1 \\ \text{Conv}([\boldsymbol{H}_{(l)}(E; P); \boldsymbol{H}_{(l-1)}(S))]), & else \end{cases} \quad (4)$$

where $\boldsymbol{H}_{(l)}(S) \in \mathbb{R}^{M \times M \times d_{share}}$ is the shared feature map representation between $\boldsymbol{H}_{(l)}(E)$ and $\boldsymbol{H}_{(l)}(P)$. d_{share} is the dimension of the shared representation.

The shared features $\boldsymbol{H}_{(l)}(S)$ are then used to calculate the interactive attention scores between tasks. We use feedforward neural networks (FFNs) and softmax activation function to calculate the interactive attention score.

$$\begin{aligned} \boldsymbol{\alpha}(E) &= \text{Softmax}(\text{FFNs}(\boldsymbol{H}_{(l)}(S))), \\ \boldsymbol{\alpha}(P) &= \text{Softmax}(\text{FFNs}(\boldsymbol{H}_{(l)}(S))^T), \end{aligned} \quad (5)$$

where $\boldsymbol{\alpha}(E) \in \mathbb{R}^{M\times M}$ and $\boldsymbol{\alpha}(P) \in \mathbb{R}^{M\times M}$ represent the Entity-to-Pair attention, and the Pair-to-Entity attention, respectively.

We then use matrix multiplication to integrate the interactive attention into the task-specific representations and incorporate the attention interaction information back into the initial representations of the two subtasks through residual connections:

$$\begin{aligned} \boldsymbol{\zeta}_{(l)}(E) &= \boldsymbol{H}_{(l)}(E) + \boldsymbol{\alpha}(E) \otimes \boldsymbol{H}_{(l)}(P), \\ \boldsymbol{\zeta}_{(l)}(P) &= \boldsymbol{H}_{(l)}(P) + \boldsymbol{\alpha}(P) \otimes \boldsymbol{H}_{(l)}(E), \end{aligned} \tag{6}$$

where $\otimes$ represents matrix multiplication.

After multiple layers of the interactive attention module, the task-specific representations $\boldsymbol{\zeta}_{(L)}(E)$ and $\boldsymbol{\zeta}_{(L)}(P)$ exchange useful information, promoting feature enhancement between the two subtasks.

4.4 Cross-Document Relative Distance

Limited by PLMs, we can only encode each document in the document set D individually, which may impair cross-document context understanding. To compensate for this, we integrate Rotary Position Embedding (RoPE) (Su et al., 2021) into the task-specific feature representations, which dynamically encodes the global relative distance across multiple documents.

Firstly, we perform Max-Pooling over the shared features across two tasks and concatenate the task-specific features with the shared features.

$$\begin{aligned} \boldsymbol{H}'(S) &= \text{Max-Pooling}(\boldsymbol{H}_{(L)}(S)), \\ \boldsymbol{H}'(F) &= [\boldsymbol{\zeta}_{(L)}(F); \boldsymbol{H}'(S)], \end{aligned} \tag{7}$$

We then employ a tag-wise MLP layer to obtain the final task-specific feature representations.

$$\boldsymbol{g}_i^t(F) = \text{MLP}^t(\boldsymbol{h}_i'(F)), \tag{8}$$

where $t \in \{tri, \dots, h2h, \dots, pos, \dots, \phi_{\text{ent}}\}$ represents our predefined special tags, shown in Appendix §D.1. ϕ_{ent} denotes the non-type label in the entity matrix. $\boldsymbol{h}_i'(F) \in \boldsymbol{H}'(F)$.

Finally, we apply rotary position embeddings to the task-specific feature representations.

$$\boldsymbol{u}_i^t(F) = \mathcal{R}(\theta, i)\boldsymbol{g}_i^t(F), \tag{9}$$

where $\mathcal{R}(\theta, i)$ is a positioning matrix parameterized by θ and the absolute index i of $\boldsymbol{g}_i^t$.

4.5 Pair Decoding

We calculate the unary score between any token pair based on the label t through each tag-wise representation $\boldsymbol{u}_i^t$:

$$s_{ij}^t = (\boldsymbol{u}_i^t)^T \boldsymbol{u}_j^t, \tag{10}$$

Where s_{ij}^t is the probability of the relation label type t between tokens w_i and w_j. We then apply a softmax layer over all elements in each matrix to determine the final relation label t. More details about grid-tagging scheme and pair decoding are shown in Appendix §D.1 and §D.2.

4.6 Multi-Task Learning

We employ the entity matrix, pair matrix, and sentiment matrix for task modeling, which are considered as three subtasks, thus the training objective of the model is to minimize the cross-entropy loss of each subtask:

$$\mathcal{L}_m = -\frac{1}{BN^2}\sum_{b=1}^{B}\sum_{i=1}^{N}\sum_{j=1}^{N} \boldsymbol{\omega}^m q_{ij}^m log(p_{ij}^m), \tag{11}$$

where $m \in \{ent, pair, senti\}$ represents a subtask, B is the total number of training data instances, N is the sum of the lengths of all document tokens in an instance's document set D. q_{ij}^m is the ground truth label, and p_{ij}^m is the predicted label. Due to the label imbalance, we adopt a tag-wise weighting vector $\boldsymbol{\omega}^m$ to alleviate this issue. The final loss is the weighted sum of the losses from the three subtasks.

$$\mathcal{L} = \alpha\mathcal{L}_{ent} + \beta\mathcal{L}_{pair} + \gamma\mathcal{L}_{senti}. \tag{12}$$

5 Experiments

5.1 Settings

We conduct experiments on our CodEOE dataset with the model proposed in Section 4. We focus on two main aspects of model performance: 1) span match, which concerns the boundaries of event triggers, event arguments, and opinion spans; and 2) pair & triplet extraction, which involves the detection of span pairs or triplets, including trigger-argument, trigger-opinion pairs, and trigger-opinion-sentiment triplets. We utilize both **Exact F1 (F1)** and **Partial F1 (PF1)** as our evaluation metrics. The details of our evaluation metrics are shown in Appendix §D.4.

For Chinese and English datasets, we utilize Chinese-Roberta-wwm-base (Cui et al., 2021) and

Table 3: Main Results on the CodEOE task. 'T/A/O' represent Event Trigger/Event Argument/Opinion, respectively.

		Span (F1)			Pair (F1)		Triplet (F1)	Span (PF1)			Pair (PF1)		Triplet (PF1)
		T	A	O	T-A	T-O	T-O-S	T	A	O	T-A	T-O	T-O-S
ZH	CRF-Extract-Classify	58.17	65.85	42.82	21.73	20.02	17.35	73.22	78.54	61.97	43.59	36.04	31.27
	InstructUIE	54.44	57.52	45.85	37.09	22.93	17.60	72.98	73.22	71.25	56.84	46.82	34.50
	Llama3-Chinese-8B	60.83	64.20	52.62	45.22	27.73	23.14	73.52	76.87	76.42	60.24	47.16	39.30
	Qwen2.5-7B-Instruct	59.24	63.99	**54.75**	42.99	30.21	23.15	72.28	75.50	76.70	60.99	49.51	**40.93**
	Ours	**67.47**	**70.05**	53.66	**50.82**	**31.76**	**25.81**	**74.25**	**80.22**	**76.99**	**61.47**	**51.84**	39.28
EN	CRF-Extract-Classify	60.36	64.14	45.98	22.91	18.42	14.74	68.24	71.21	58.66	32.19	30.05	26.44
	InstructUIE	56.98	59.07	46.65	42.54	23.62	17.76	65.66	65.69	72.80	47.82	39.08	29.89
	Llama3-8B-Instruct	58.30	60.42	**58.39**	40.00	30.41	23.79	68.09	71.20	69.83	52.57	41.99	33.01
	Qwen2.5-7B-Instruct	57.21	61.22	57.25	41.87	30.48	**24.01**	64.30	69.75	66.42	51.23	41.58	**33.91**
	Ours	**66.00**	**66.50**	53.38	**49.52**	**30.85**	23.78	**73.60**	**77.06**	**74.07**	**57.53**	**43.99**	32.62

Roberta-base (Liu et al., 2019) as our base encoders, respectively. The learning rate is set to 1e-5 for the Chinese dataset and 2e-5 for the English dataset. The testing results are obtained from models fine-tuned on the the validation set. All experiments are conducted using five different random seeds, and the reported scores represent the average of five runs. Other experimental settings are shown in Appendix §D.5.

5.2 Baselines

Since there is currently no model for joint event and opinion extraction, we consider re-implementing two strong baseline models for our CodEOE task, including CRF-Extract-Classify (Cai et al., 2021) and InstructUIE (Wang et al., 2023). Our experimental details about baselines are shown in D.3.

CRF-Extract-Classify takes the same PLMs as used in our model. InstructUIE uses mT5-base (Xue et al., 2021). Moreover, we conduct additional experiments with LLMs. We use Llama3-8B-Instruct (AI@Meta, 2024) and Llama3-Chinese-8B (Cui et al., 2023) to fine-tune on the English and Chinese datasets, respectively, using LLaMA-Factory (Zheng et al., 2024). We use Qwen2.5-7B-Instruct (Yang et al., 2024) to fine-tune on both datasets within the same framework. The input prompt template is shown in Appendix §E.

5.3 Main Results

Table 3 compares the performance of various models on the CodEOE task. Compared to baseline models, our proposed method achieves near-optimal results across all metrics, particularly excelling in pair and triplet extraction tasks.

Firstly, our method achieves nearly the highest exact F1 scores across the board for span categories. Specifically, in the Chinese dataset, it surpasses Llama3-Chinese-8B, the next best performer, by margins of 6.64, 5.85, and 1.04 points for triggers (T), arguments (A), and opinions (O), respectively. Notably, Qwen2.5-7B-Instruct shows competitive performance, particularly in opinion span detection with the highest F1 score of 54.75.

Table 4: Ablation Results (F1). 'w/o All': removing all three components (IA, RoPE and Wei.(ω^m)).

	ZH			EN		
	T-A	T-O	T-O-S	T-A	T-O	T-O-S
Ours	50.82	31.76	25.81	49.52	30.85	23.78
w/o IA	45.34	27.42	24.02	43.15	26.83	21.14
w/o RoPE	44.01	28.79	25.09	45.81	27.22	22.62
w/o Wei.(ω^m)	46.97	28.29	24.35	48.07	28.91	21.32
w/o IA & RoPE	39.69	26.77	22.47	42.15	27.07	22.01
w/o All	39.05	26.13	21.56	37.44	26.85	20.94

Secondly, our model shows its strength in more complex scenarios involving pair and triplet extraction. For example, in the Chinese dataset, our model leads Llama3-Chinese-8B by 5.60, 4.03, and 2.67 points in exact F1 scores for T-A, T-O, and T-O-S, respectively. Qwen2.5-7B-Instruct also demonstrates strong performance, particularly in the English dataset, achieving the highest F1 score for T-O-S.

Finally, while our model excels in PF1 scores, Qwen2.5-7B-Instruct demonstrates excellent performance in PF1 scores for T-O-S in both Chinese (40.93) and English (33.91) datasets. This result highlights the ability of LLMs to match ambiguous boundaries and recognize emotions, leading to superior performance in partial matching evaluations.

5.4 Ablation Study

We conduct ablation experiments on the CodEOE task. We progressively remove key components, including the Interactive Attention module (IA), Rotary Position Embeddings (RoPE), the weighting strategy (ω^m), and combinations of these components. Results in Table 4 provide valuable insights into the effectiveness of each component.

Removing the Interactive Attention module (w/o

Table 5: The impact of interactive attention module depths on the CodEOE dataset

Layers	ZH			EN		
	T-A	T-O	T-O-S	T-A	T-O	T-O-S
L=1	48.83	28.98	24.52	48.44	28.37	21.64
L=2	**51.08**	30.03	25.00	**49.52**	**30.85**	**23.78**
L=3	50.82	**31.76**	**25.81**	48.47	29.10	22.49
L=4	50.37	29.41	23.89	46.73	29.44	21.26

IA) causes substantial degradation across tasks (e.g., 4.34 F1 drop for T-O in Chinese), validating its critical role in modeling trigger-opinion dependencies. Eliminating Rotary Position Embeddings (RoPE) significantly impairs pair extraction (6.81 F1 decline for Chinese T-A), confirming its effectiveness in capturing cross-document positional relations. After removing the task-weighting strategy (w/o Wei.(ω^m)) reduces T-A performance by 3.85 F1 in Chinese dataset, indicating its necessity for balanced multi-task learning.

Component combinations reveal synergistic effects: concurrent removal of IA and RoPE causes catastrophic performance collapse (11.13 F1 drop for Chinese T-A), while removing all components yields the lowest scores. These results collectively establish the complementary nature of the inter-document interaction of IA, the positional awareness of RoPE, and adaptive task weighting in addressing challenges of the CodEOE task.

5.5 Further Analysis

Impact of the Interactive Attention Module Depths. As shown in Table 5, we conduct a further analysis on the number of layers in the Interactive Attention (IA) module. The results show that the model achieves the best performance with 3 layers on the Chinese dataset, while 2 layers yield the best results on the English dataset. This difference can be attributed to the distinct grammatical and syntactic characteristics of Chinese and English. Chinese sentences are more flexible in structure, with semantic relations often implicit in the context, requiring deeper interactive modeling to capture the complex relations between triggers and their associated elements. When the number of layers increases to 4, the model performance on both the Chinese and English datasets decreases, which can be attributed to noise accumulation or overfitting.

Analysis on the Number of Event Triggers. Table 2 highlights that instances containing multiple events are a key characteristic of the CodEOE dataset. To further investigate model performance under different numbers of events, we evaluate our model and two strong baselines (InstructUIE and Llama3-8B) on trigger-argument pair extraction and trigger-opinion pair extraction tasks, as shown in Figure 3. The results demonstrate that our model shows significant advantages in single-trigger scenarios, achieving much higher F1 scores than the other two models. However, as the number of triggers increases, the performance of all models declines, particularly in multi-trigger scenarios (3 triggers or more). This indicates that multi-trigger scenarios bring greater challenges for models to understand cross-document contexts. Additionally, in multi-trigger scenarios, Llama-3-8B slightly outperforms our model in trigger-opinion pair extraction, reflecting the strong understanding capability of large language models when handling long texts.

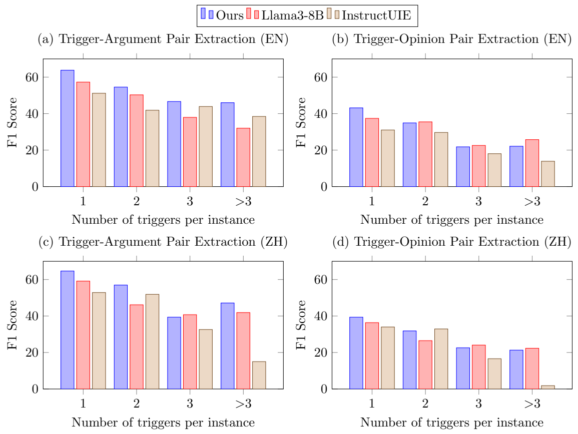


Figure 3: Results of pair extraction on instances containing different number of event triggers for English and Chinese Datasets.

We also conduct a case study and make a comparison with two strong baselines, which is shown in Appendix §D.6.

6 Conclusion

In this paper, we introduce a novel task of Cross-Document Event-Opinion Extraction (CodEOE), bridging the gap in the understanding of event-level opinions and sentiments. We manually construct a high-quality, bilingual dataset, providing a significant resource for research into cross-document semantic understanding. We benchmark the CodEOE task using an end-to-end model, which demonstrates robust capability in capturing cross-document contextual interactions. Experimental results reveal the challenges of the task, such as the diversity of opinion expressions and the complex relations between opinions and events.

Limitations

Our work has the following potential limitations. Firstly, our CodEOE dataset is collected from the social media platform, Weibo, which predominantly emphasize public events with immediate dissemination value. This could relatively limit coverage of events in specialized platforms such as News platforms and financial websites. Secondly, we only annotate the CodEOE dataset in two languages. We plan to extend multilingual support to enhance cultural and linguistic coverage.

Ethics Statement

This research utilizes data exclusively sourced from the publicly accessible platform, Weibo, ensuring no inclusion of personally identifiable information. We implement rigorous measures including diverse sampling strategies and manual verification processes to enhance data representativeness and reliability. The methodologies and dataset construction details are transparently documented to enable reproducibility, with the full dataset to be publicly released to support academic inquiry. We adhere to ethical standards in research and ensure compliance with institutional and national guidelines.

References

AI@Meta. 2024. Llama 3 model card.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. 2017. Affective computing and sentiment analysis. *A practical guide to sentiment analysis*, pages 1–10.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Yuqi Chen, Chen Keming, Xian Sun, and Zequn Zhang. 2022. A span-level bidirectional network for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4309, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhexue Chen, Hong Huang, Bang Liu, Xuanhua Shi, and Hai Jin. 2021. Semantic and syntactic enhanced aspect sentiment triplet extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1474–1483, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Haolin Deng, Yanan Zhang, Yangfan Zhang, Wangyang Ying, Changlong Yu, Jun Gao, Wei Wang, Xiaoling Bai, Nan Yang, Jin Ma, Xiang Chen, and Tianhua Zhou. 2022. Title2Event: Benchmarking open event extraction with a large-scale Chinese title dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6511–6524, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Xinya Du, Alexander Rush, and Claire Cardie. 2021. Template filling with generative transformers. In

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 909–914, Online. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442, Brussels, Belgium. Association for Computational Linguistics.

Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. 2022. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4121–4128. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Qiang Gao, Zixiang Meng, Bobo Li, Jun Zhou, Fei Li, Chong Teng, and Donghong Ji. 2024. Harvesting events from multiple sources: Towards a cross-document event extraction paradigm. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1913–1927, Bangkok, Thailand. Association for Computational Linguistics.

Yuxin He, Jingyue Hu, and Buzhou Tang. 2023. Revisiting event argument extraction: Can EAE models learn better when being aware of event co-occurrences? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12542–12556, Toronto, Canada. Association for Computational Linguistics.

Zijin Hong and Jian Liu. 2024. Towards better question generation in qa-based event extraction. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 9025–9038. Association for Computational Linguistics.

Rajkumar S. Jagdale, Vishal S. Shirsat, and Sachin N. Deshmukh. 2016. Sentiment analysis of events from twitter using open source tool.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. Diaasq: A benchmark of conversational aspect-based sentiment quadruple analysis. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13449–13467. Association for Computational Linguistics.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

You Li, Xupeng Zeng, Yixiao Zeng, and Yuming Lin. 2024. Enhanced packed marker with entity information for aspect sentiment triplet extraction. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 619–629, New York, NY, USA. Association for Computing Machinery.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1789–1798, Vancouver, Canada. Association for Computational Linguistics.

Wanlong Liu, Shaohuan Cheng, Dingyi Zeng, and Qu Hong. 2023. Enhancing document-level event argument extraction with contextual clues and role relevance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12908–12922, Toronto, Canada. Association for Computational Linguistics.

Wanlong Liu, Li Zhou, DingYi Zeng, Yichen Xiao, Shaohuan Cheng, Chen Zhang, Grandee Lee, Malu Zhang, and Wenyu Chen. 2024. Beyond single-event extraction: Towards efficient document-level multi-event argument extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9470–9487, Bangkok, Thailand. Association for Computational Linguistics.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Brussels, Belgium. Association for Computational Linguistics.

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

Ziheng Liu, Rui Xia, and Jianfei Yu. 2021. Comparative opinion quintuple extraction from product reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3955–3965, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Di Lu, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2023. Event extraction as question generation and answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1666–1688, Toronto, Canada. Association for Computational Linguistics.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6759–6774, Dublin, Ireland. Association for Computational Linguistics.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 432–439.

Chien Nguyen, Hieu Man, and Thien Nguyen. 2023. Contextualized soft prompts for extraction of event arguments. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4352–4361, Toronto, Canada. Association for Computational Linguistics.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8600–8607. AAAI Press.

Alexandru Petrescu, Ciprian-Octavian Truică, and Elena-Simona Apostol. 2019. Sentiment analysis of events in social media. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 143–149.

Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. 2022. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, Dublin, Ireland. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan. The COLING 2016 Organizing Committee.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. 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)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. DeepStruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823, Dublin, Ireland. Association for Computational Linguistics.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, and 1 others. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Kaiwen Wei, Xian Sun, Zequn Zhang, Jingyuan Zhang, Guo Zhi, and Li Jin. 2021. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4672–4682, Online. Association for Computational Linguistics.

Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. 2022. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11513–11521.

Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. 2021. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3957–3963. ijcai.org.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585, Online. Association for Computational Linguistics.

Zhiyang Xu, Jay Yoon Lee, and Lifu Huang. 2023. Learning from a friend: Improving event extraction via self-training with feedback from Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10421–10437, Toronto, Canada. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. A partition filter network for joint entity and relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 185–197, Punta Cana, Dominican Republic. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yuqing Yang, Qipeng Guo, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023. An AMR-based link prediction approach for document-level event argument extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12876–12889, Toronto, Canada. Association for Computational Linguistics.

Qi Zeng, Qiusi Zhan, and Heng Ji. 2022. EA^2E: Improving consistency with event awareness for document-level argument extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2649–2655, Seattle, United States. Association for Computational Linguistics.

Qi Zhang, Jie Zhou, Qin Chen, Qingchun Bai, and Liang He. 2022. Enhancing event-level sentiment analysis with structured arguments. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1944–1949, New York, NY, USA. Association for Computing Machinery.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhisong Zhang, Xiang Kong, Zhengzhong Liu, Xuezhe Ma, and Eduard Hovy. 2020. A two-step approach for implicit event argument detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7479–7485, Online. Association for Computational Linguistics.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. SpanMlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3239–3248, Online. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Xujuan Zhou, Xiaohui Tao, Jianming Yong, and Zhenyu Yang. 2013. Sentiment analysis on tweets for social events. In *Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 557–562.

A Annotation Details

A.1 Annotation Standard

Event Annotation: Given the diversity of hot news event types on social media, manual design of specific event schema is costly and time-consuming, and predefined event types often fail to capture the diversity of events originating from social media news. Similar to open event extraction (Deng et al., 2022), an event is defined as an action or a state of change which occurs in the real world. We avoid predefining event types or schemas, allowing models to flexibly adapt to diverse event types. We define seven types for event arguments: Location, Date, Organization, Person, Country, Object and Other. While event triggers remain type-agnostic to capture open-domain patterns, argument types serve solely as consistency anchors during boundary verification. The evaluation explicitly focuses on trigger-argument pair identification, excluding argument type labels from assessment metrics while maintaining rigorous evaluation of argument boundary accuracy and structural association.

Event annotation can be formalized as: $Event = \{Trigger, [Argument_1, Argu - ment_2, \ldots, Argument_n]\}$. The $Trigger$ constitutes the minimal text span structurally anchoring an event predicate, while an $Argument$ denotes any semantic role-bearing constituent fulfilling the predicate-argument structure linked to its corresponding $Trigger$.

Opinion Annotation: An opinion is an individual's emotional attitude or viewpoint towards an event. For opinion annotation, we observe that event-level opinions often could not be captured by simple words or phrases. Thus, we represent opinions at the clause level to better capture the complexity of expressions related to events. The sentiment of an opinion is categorized into $positive$, $negative$, and $neutral$. Opinion annotation can be formalized as: $Expression = \{Trigger, Opinion, Sentiment\}$, where $Opinion$ is the span expressing a viewpoint represented by one or several consecutive clauses. $Sentiment$ is the sentiment orientation of the $Opinion$ towards an event, which is represented by $Trigger$.

A.2 Annotation Process

Span-Level Annotation. The primary task for annotators is to identify and mark event triggers in the text, event-related arguments (such as involved persons, locations, times, etc.), and opinions and their sentiments related to the event. Firstly, annotators precisely locate the event triggers by marking their start and end positions in the text. Subsequently, all relevant arguments and their positions are identified and marked. Additionally, when expressing opinions related to events, annotators mark the clauses that express these opinions and categorize their sentiments into three types: $positive$, $negative$, or $neutral$.

Relation-Level Annotation. In the relation-level annotation, we treat the event trigger as the subject of the event, and annotators connect each event trigger with its associated event arguments. For each opinion, annotators link it to the event trigger it pertained to, and the sentiment polarity is assigned based on the expressed sentiment towards the event.

B Extended Data Statistics

B.1 Polarity Distribution

We analyze the distribution of sentiment polarities in the trigger-opinion-sentiment triplets within both the Chinese and English datasets. In the Chinese dataset, the proportions of positive, negative, and neutral sentiment of triplets are 27.3%, 46.7%, and 26.0%, respectively. Similarly, the English dataset shows a distribution of 27.1% positive, 46.9% negative, and 26.0% neutral sentiment of triplets. The distribution of sentiment polarities is relatively even, with no evident long-tail distribution. Negative sentiment constitutes the largest proportion. This may be related to the tendency of social media users to express negative emotions. Such a balanced distribution indicates that our data sampling is reasonable, which helps reduce biases when models process data across different sentiment categories.

B.2 Topic Distribution

Additionally, we segment our dataset into ten distinct topics, including Society, Sports, Disaster, Business, Politics, Technology, Finance, Entertainment, Military, and Else. As illustrated in figure 4, the Society topic comprises the highest proportion of data, reflecting the natural inclination of social media users to discuss societal events and underscoring the role of social media as a primary platform for public discourse. This topical distribution characteristic makes the dataset more aligned with real-world hot event scenarios, providing a practical context for research.

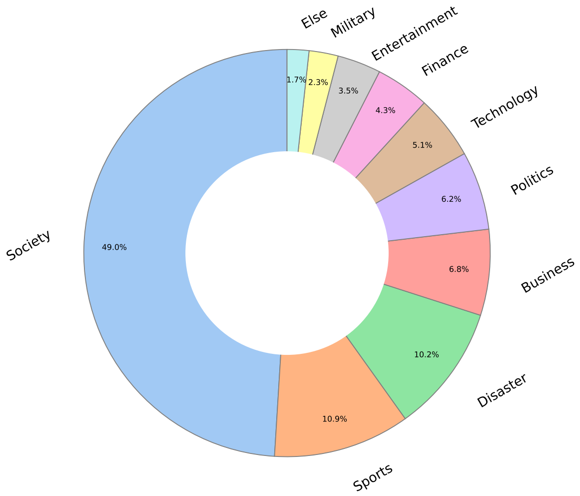


Figure 4: The distribution of topics in CodEOE.

Item	Text
Source	具体来看，易方达创业板ETF当日净申购4.79亿份，资金净流入7.68亿元，助推该ETF规模突破400亿元大关，达到405亿元。
Translated	Specifically, the E Fund ChiNext ET had a net subscription of 479 million shares that day, with a net inflow of 768 million yuan, boosting the scale of the ETF to exceed the 40 billion yuan mark, reaching 40.5 billion yuan.
Revision	Specifically, the E Fund ChiNext ET had a net subscription of 479 million shares that day, with a net inflow of 768 million yuan, boosting the scale of the ETF to exceed the 40 billion yuan threshold, reaching 40.5 billion yuan.
Source	华夏科创50ETF净申购6.53亿份，资金净流入5.06亿元
Translated	The net subscription of Huaxia Science and Technology Innovation 50ETF was 653 million shares, and the net inflow of funds was 506 million yuan.
Revision	The net subscription of ChinaAMC STAR 50 ETF was 653 million shares, and the net inflow of funds was 506 million yuan.

Figure 5: Two translation revision examples. The first one is a more appropriate expression. The second one addresses error correction for proper nouns.

C Specification on Data Construction

C.1 Parallel English Dataset Construction

To further the development of joint analysis of events and opinions, we also construct an English version of the dataset based on the Chinese corpus. This involved two steps: text translation and annotation projection.

Text Translation: We use Google Translate API[3] to convert the Chinese text into English. Despite the good performance of NMT (Neural Machine Translation), some errors still occur during the translation process. A significant reason for these errors is that our corpus, collected from social media, is filled with grammatically non-compliant sentences, which has brought challenges for the NMT system to produce correct and elegant translations. Thus, we meticulously revise the translations to eliminate errors and ensure readability. Figure 5 lists one of the errors and revision results.

Annotation Projection: After attempting to use the awesome-align automatic alignment tool (Dou and Neubig, 2021), we find its performance on aligning named entities unsatisfactory. Consequently, we resort to manually re-annotating the alignments, ultimately producing the annotated English corpus.

[3] https://cloud.google.com/translate

D Model and Experiment Specification

D.1 Grid-Tagging Scheme

The grid-tagging method (Wu et al., 2020; Li et al., 2023) has become increasingly popular in recent years for end-to-end information extraction models. we apply the grid-tagging method to our end-to-end extraction framework and redesign the labeling scheme to meet our needs.

We divide the labeling scheme into three blocks: entity span boundary detection, entity pair detection, and opinion sentiment detection.

Entity span boundary labels: We use tri, arg, and opi to denote the tagging relations between the head and tail of event triggers, event arguments, and opinion terms, respectively. For example, the arg between '*February*' and '*1*' denotes an event argument of '*February 1*' in Figure 6.

Entity pair labels: We use $h2h$ and $t2t$ labels, both of which align the head and tail tokens between a pair of entities in two types. For example, the head word of '*February*' (argument) and '*issued*' (trigger) is connected with $h2h$, while the tail word of '*1*' (argument) and '*issued*' (trigger) is connected with $t2t$, which is shown in Figure 6.

Opinion sentiment labels: We add a sentiment polarity label to the head and tail of the two entities in the trigger-opinion pair, indicating the sentiment expressed by the opinion towards a particular event. Sentiment polarity labels include pos, neg and neu. As shown in Figure 7, we assign a sentiment label between the heads and tails of triggers and opinions.

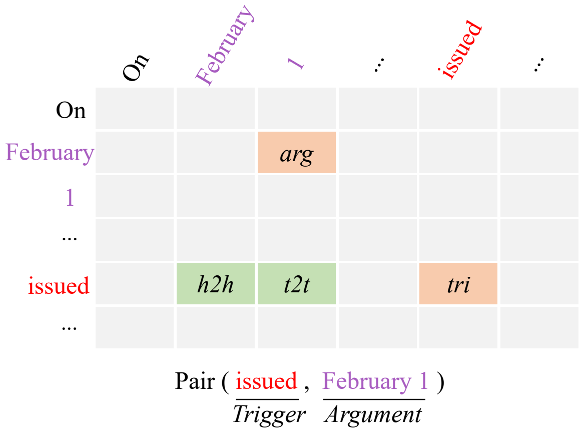


Figure 6: Tagging scheme for pair extraction

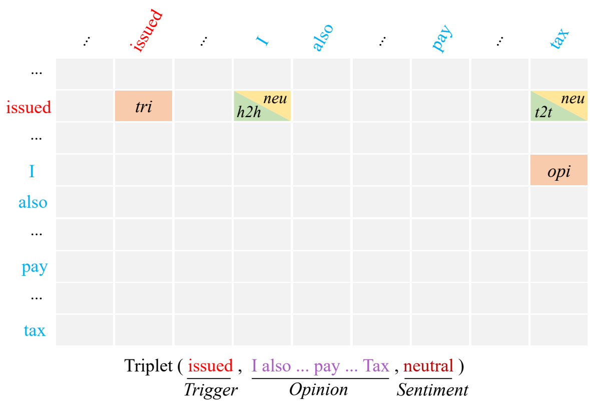


Figure 7: Tagging scheme for triplet extraction

D.2 Label Classification

After calculating s_{ij}^{t}, the probability of the relation label type t between tokens w_i and w_j in Eq. (10), we apply a softmax layer over all elements in each matrix to determine the final relation label t.

$$\begin{aligned} p_{ij}^{ent} &= \text{Softmax}([s_{ij}^{\phi ent}; s_{ij}^{tri}; s_{ij}^{arg}; s_{ij}^{opi}]), \\ p_{ij}^{pair} &= \text{Softmax}([s_{ij}^{h2h}; s_{ij}^{t2t}]), \\ p_{ij}^{senti} &= \text{Softmax}([s_{ij}^{pos}; s_{ij}^{neg}; s_{ij}^{neu}]), \end{aligned} \tag{13}$$

where p_{ij}^{ent}, p_{ij}^{pair} and $p_{ij}^{senti}$are the probabilities of each relation label between token w_i and token $w_j$in the entity matrix, pair matrix, and sentiment matrix, respectively. After obtaining all the labels in the grid, we decode the trigger-argument pairs and trigger-opinion-sentiment triplets according to the labeling scheme described in §D.1.

D.3 Baselines

Since there is currently no model for joint event and opinion extraction, we consider re-implementing two strong baseline models for our CodEOE task, including CRF-Extract-Classify (Cai et al., 2021) and InstructUIE (Wang et al., 2023).

- CRF-Extract-Classify is a two-stage pipeline model designed for the ABSA task. It first performs joint extraction of aspects and opinions, and then classifies the predicted category-sentiment based on the extracted aspect-opinion pairs in the second stage. To adapt to our CodEOE task, we modified the model. Specifically, we simplified the original aspect-category-opinion-sentiment quadruplet into a trigger-argument pair and a trigger-opinion-sentiment triplet. In the modified model, the trigger-argument and trigger-opinion are co-extracted in the first step, and then in the second step, the sentiment term is predicted based on the extracted trigger-opinion.

- InstructUIE is a unified information extraction framework that utilizes instruction tuning with large language models (LLMs). This approach enables the model to uniformly simulate various information extraction tasks and capture the interdependencies between tasks. Here we convert the pair and triplet extraction into relation extraction form and fine-tune the model using instructions for the relation extraction task.

D.4 Evaluation Metrics

We utilize both Exact F1 (F1) and Partial F1 (PF1) as our evaluation metrics.

Exact F1 evaluates the complete congruence between predictions and ground truth. For spans, a prediction is considered correct only if it precisely matches the start and end boundaries of an entity. For pairs, the prediction must accurately identify both two spans. For triplets, the prediction must not only match both spans but also correctly classify their sentiment polarity.

Partial F1 evaluates partial consistency between predictions and ground truth. Predictions are defined as tuple $p = \{p_1, p_2, \dots, p_n\}$, with n ($n \in \{1, 2, 3\}$) denotes span, pair, or triplet structures, respectively. For instance, a predicted trigger-argument-sentiment triplet may be represented as $p_{triplet} = \{p_{tri}, p_{opi}, p_{senti}\}$. For each prediction p and its best-matching ground truth g, the degree of match is quantified by calculating the length of the Longest Common Substring (LCS) between them. A prediction p is considered correct if the

LCS length for all p_i reaches at least a predetermined threshold τ (set to 0.5) of the corresponding g_i length. For triplets, in addition to span matching, the sentiment polarity p_{senti} of the prediction must also fully align with g_{senti}.

D.5 Extended Experimental Settings

We use AdamW algorithm for optimization. The hidden state dimension of Roberta is set to 768. The weight decay value is set to 0.01 and the warmup rate is set to 0.1. Within the Interactive Attention module, the dropout rate for the Multi-Layer Perceptron (MLP) and convolutional layers is set to 0.1. The hidden layer dimensions for the MLPs in Eq. (3) and Eq. (8) are set to 768 and 128, respectively. The tag-wise weight vector $\boldsymbol{\omega}^m$ is set to [1, 2, 2, 2]. α, β and γ in Eq. (12) are set to 1.5, 2.5 and 3.5, respectively. The batch size is set to 2 at multi-document level. The training epochs are set to 30 for both Chinese and English datasets. The train process adopts an early stopping strategy and the patience is set to 10. Experiments are run on one same Tesla A100 GPU.

D.6 Case Study

We conduct a case study and make a comparison with two strong baselines, InstructUIE and Llama3-8B-Instruct. As shown in Figure 8, our model consistently outperforms the baselines for trigger-argument and trigger-opinion pair extraction. For the trigger '*came into effect*', Llama3-8B-Instruct incorrectly merges two independent arguments, '*U.S. International Trade Commission*' and '*Apple Watch sales ban*', into a single long span. Similarly, for the opinion '*To tell the truth … property development.*', InstructUIE extracts an excessively long span that includes unnecessary contextual information. For sentiment classification of event-specific opinions, InstructUIE and Llama3-8B-Instruct exhibit varying degrees of misinterpretation. We attribute this to the complexity of the task, which requires models to not only identify the relations between triggers and opinions but also accurately understand the sentiment towards a specific trigger. This dual challenge of relation identification and sentiment analysis poses significant difficulties for current models.

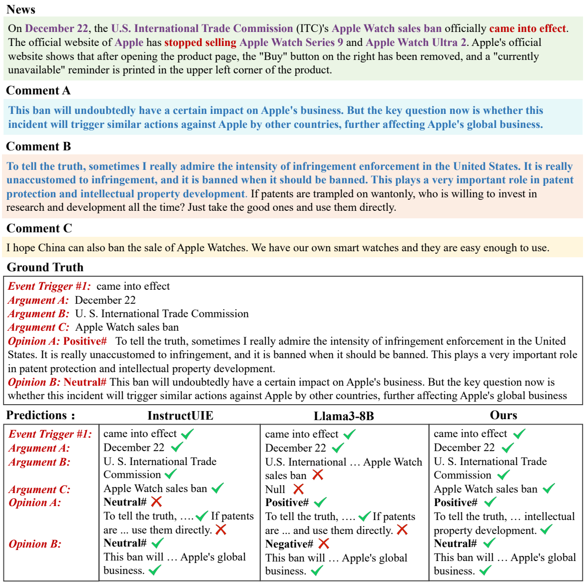


Figure 8: A test case from the CodEOE dataset focusing on the event trigger 'came into effect'.

E Input Prompt for LLMs

> Your task is to extract information from a news document and several comment texts. You will be provided with multiple documents. Your goal is to extract event and opinion information. Find the 'trigger word', which represents the main event or action; the 'argument', which represents the key entity or time related to the trigger word; and the 'opinion', which represents the view or description of the event associated with the trigger word. Understand whether there is a relationship between these pieces of information, and then organize the related information into 'trigger-argument pairs' and 'trigger-opinion-sentiment pairs'. Sentiment can be 'positive', 'negative', or 'neutral'. The output should be in the form of relationship pairs, with four types of relationships: trigger-argument, trigger-opinion-positive, trigger-opinion-negative, and trigger-opinion-neutral. The output format should be "relation1: word1, word2; relation2: word3, word4".
> Document input:
> document1: {...},
> document2: {...},
> ...