SUPPLEMENTARY MATERIAL
CODEOE: A BENCHMARK FOR JOINTLY EXTRACTING CROSS-DOCUMENT EVENTS AND OPINIONS FROM SOCIAL MEDIA

Zixuan Wang[†], *Yun Hu*[†], *Mengying Yuan*[†], *Kang He*[†], *Fei Li*[†*], *Huisheng Ma*[‡*], *Donghong Ji*[†], *Chong Teng*[†]

[†] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education,
School of Cyber Science and Engineering, Wuhan University, Wuhan, China,
{zixuanwang_nlp, lifei_csnlp}@whu.edu.cn
[‡] North China Institute of Computing Technology, Beijing, China

1. DATA CONSTRUCTION

1.1. 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.

1.2. Annotation Framework

1.2.1. Annotation Standard

We summarize some crucial parts of the annotation standards, mainly divided into event annotation and opinion annotation.

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 (1), 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: $Event = \{Trig-ger, [Argument_1, Argument_2, \ldots, Argument_n]\}$. 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 $Trigger$.

1.2.2. Annotation Process

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

*Corresponding author.
[1]https://weibo.com/

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 annotations and random checks, as well as regular annotation review meetings. The annotation process is divided into span-level and relation-level steps.

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.

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.

1.3. 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 1 lists one of the errors and revision results.

Annotation Projection: After attempting to use the awesome-align automatic alignment tool (2), we find its performance on aligning named entities unsatisfactory. Consequently, we resort to manually re-annotating the alignments, ultimately producing the annotated English corpus.

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.

Fig. 1. Two translation revision examples. The first one is a more appropriate expression. The second one addresses error correction for proper nouns.

2. EXTENDED DATA STATISTICS

2.1. Detailed statistical analysis

To comprehensively assess the characteristics of our dataset, we conduct a detailed statistical analysis. As shown in Table 1, 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.

2.2. 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

[2]https://anonymous.4open.science/r/CodEOE-08BD
[3]https://cloud.google.com/translate

Table 1. 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

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.

2.3. 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 2, 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.

3. MODEL AND EXPERIMENT SPECIFICATION

3.1. Grid-Tagging Scheme

The grid-tagging method (3; 4) 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

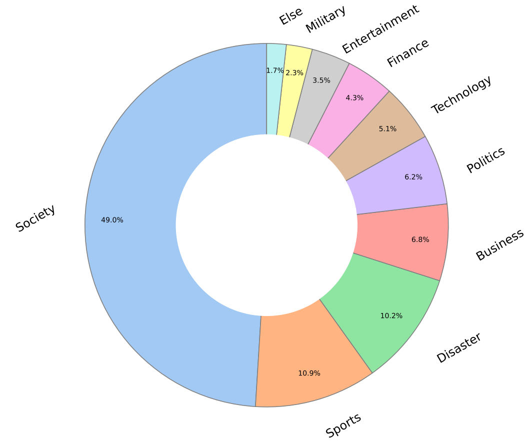

Fig. 2. The distribution of topics in CodEOE.

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 3.

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 3.

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 4, we assign a sentiment label between the heads and tails of triggers and opinions.

3.2. Label Classification

After calculating s_{ij}^{t}, the probability of the relation label type t between tokens w_i and w_j in Eq. (8), we apply a softmax layer over all elements in each matrix to determine the final relation label t.

$$\begin{aligned} p_{ij}^{ent} &= \text{Softmax}([s_{ij}^{\phi_{ent}}; s_{ij}^{tri}; s_{ij}^{arg}; s_{ij}^{opi}]), \\ p_{ij}^{pair} &= \text{Softmax}([s_{ij}^{h2h}; s_{ij}^{t2t}]), \\ p_{ij}^{senti} &= \text{Softmax}([s_{ij}^{pos}; s_{ij}^{neg}; s_{ij}^{neu}]), \end{aligned} \tag{1}$$

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-

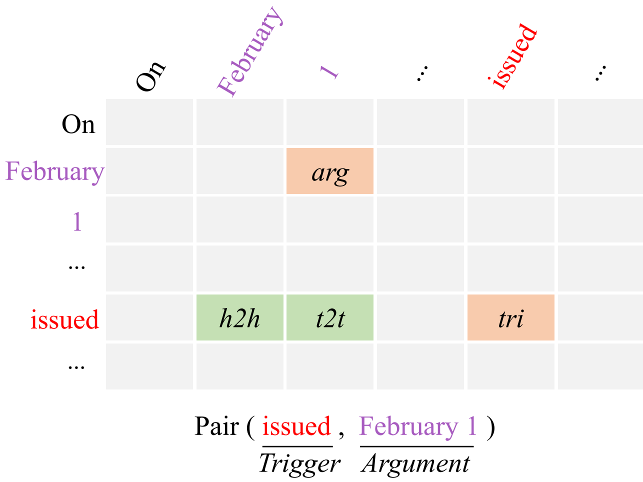

Fig. 3. Tagging scheme for pair extraction

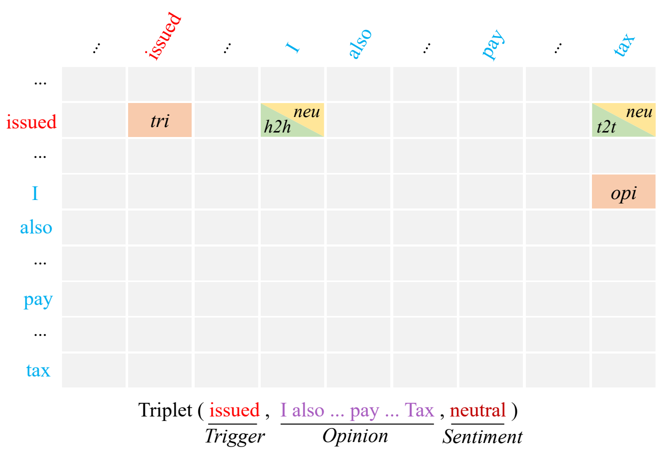

Fig. 4. Tagging scheme for triplet extraction

argument pairs and trigger-opinion-sentiment triplets according to the labeling scheme described in §3.1.

3.3. 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 (5) and InstructUIE (6).

- 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.

3.4. 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}.

3.5. 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. (7) are set to 768 and 128, respectively. The tag-wise weight vector ω^m is set to [1, 2, 2, 2]. α, β and γ in Eq. (10) 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.

3.6. Zero-shot and Few-shot experiments on LLMs

We perform Zero-Shot, 5-Shot, and 10-Shot evaluations (temperature=0.5) on GPT-4o-0806 (7), Claude-3.7-Sonnet (8) and DeepSeek-V3-0324 (9). The results are presented in the table below, where the results represent the mean scores from five independent runs.

As shown in Table 2, part of closed-source LLMs and open-source LLMs perform worse than fine-tuned models in zero-shot and few-shot settings. This suggests our dataset exhibits no data contamination with existing LLMs, and our task presents a challenge for current large models.

3.7. 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: {...},
> ...

3.8. Case Study

We conduct a case study and make a comparison with two strong baselines, InstructUIE and Llama3-8B-Instruct. As shown in Figure 5, 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.

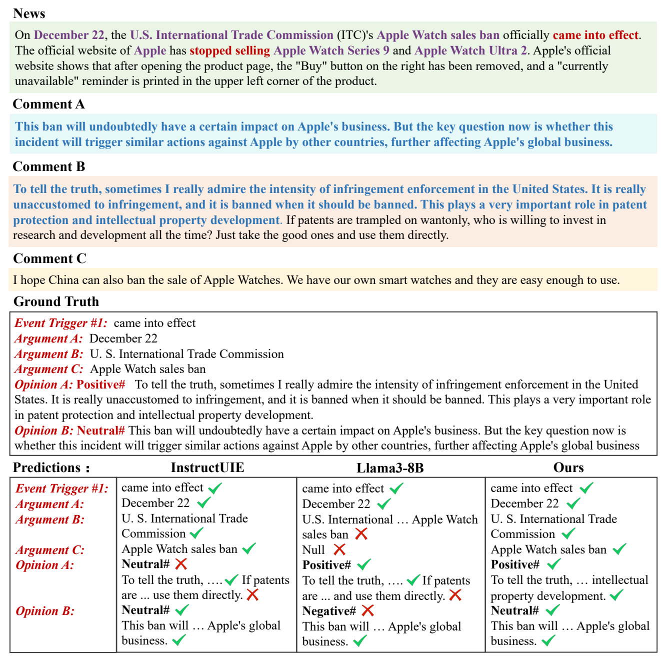

Fig. 5. A test case from the CodEOE dataset focusing on the event trigger 'came into effect'.

4. RELATED WORK

4.1. 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) (10) has facilitated numerous breakthrough studies(11; 12; 13; 14; 15). Later, Deng et al.(1) 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.(16) introduced the Roles Across Multiple Sentences (RAMS) dataset. Li et al.(17) proposed a new document-level event extraction benchmark dataset, WIKIEVENTS. The mainstream methods for DEE typically include span-based methods (18; 19; 20; 21) and generation-based methods (17; 22; 23). Recently, prompt-based (24; 25; 26; 27; 28) and QA-based methods (29; 30; 31; 32; 33) have also been employed to guide models in event extraction. Moreover, Gao et al.(34) introduced the Cross-Document Event Extraction (CDEE) task.

Table 2. Main Results on the CodEOE task. 'T/A/O' represent Event Trigger/Event Argument/Opinion, respectively.

			Prompt-based Methods						Supervised Methods					
		Setting	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	GPT-4o-0806	Zero-shot	27.42	27.90	19.64	15.74	4.62	3.02	40.04	47.53	39.29	23.61	11.01	6.93
		5-shot	45.88	49.54	51.43	26.25	19.21	11.70	63.80	64.02	76.60	36.29	34.34	23.84
		10-shot	48.25	48.35	55.07	28.80	22.74	14.94	68.45	64.71	78.71	36.61	35.32	23.86
	DeepSeek-V3-0324	Zero-shot	25.06	28.06	17.59	13.74	3.18	1.43	43.73	46.49	33.28	22.99	10.85	6.57
		5-shot	43.56	46.34	54.24	28.30	20.32	13.02	61.39	67.78	78.30	38.63	33.33	21.70
		10-shot	44.40	51.38	56.47	29.77	22.00	13.62	63.0	70.36	79.76	39.15	35.74	22.70
	Claude-3.7-Sonnet	Zero-shot	22.14	35.24	14.09	17.88	2.83	1.16	40.96	56.44	32.32	25.91	10.29	6.30
		5-shot	41.93	50.72	54.98	28.05	21.51	14.14	61.90	67.41	80.28	39.51	39.24	26.10
		10-shot	43.93	53.02	**56.52**	32.98	24.48	17.11	62.43	73.29	**81.39**	43.68	39.01	27.26
	CRF-Extract-Classify	Fine-tuned	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	Fine-tuned	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	Fine-tuned	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	Fine-tuned	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	Fine-tuned	**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	GPT-4o-0806	Zero-shot	22.81	25.38	8.54	13.93	1.86	1.24	37.86	45.79	24.84	19.38	4.80	3.25
		5-shot	41.82	39.09	56.26	25.77	24.71	15.54	56.62	56.45	64.52	33.94	32.91	21.71
		10-shot	45.95	44.91	55.82	29.06	24.45	20.02	59.16	62.91	64.69	38.05	35.66	26.70
	DeepSeek-V3-0324	Zero-shot	21.47	21.76	9.06	11.79	2.50	1.31	26.20	36.33	21.87	16.80	5.83	2.62
		5-shot	36.64	36.06	57.08	19.50	27.97	17.66	48.35	56.47	64.09	32.57	38.04	27.37
		10-shot	43.59	40.02	**58.60**	25.00	28.30	21.47	51.31	58.65	67.26	36.30	40.84	27.54
	Claude-3.7-Sonnet	Zero-shot	19.48	20.28	6.87	12.05	1.34	0.67	24.04	38.80	20.74	17.86	3.84	1.87
		5-shot	41.25	39.97	53.30	24.61	24.32	19.75	50.39	57.24	65.60	34.27	38.14	29.45
		10-shot	41.63	40.79	55.25	26.52	29.52	20.97	50.32	59.72	66.35	35.83	39.58	31.78
	CRF-Extract-Classify	Fine-tuned	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	Fine-tuned	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	Fine-tuned	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	Fine-tuned	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	Fine-tuned	**66.00**	**66.50**	53.38	**49.52**	**30.85**	23.78	**73.60**	**77.06**	**74.07**	**57.53**	**43.99**	32.62

4.2. 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 (35; 36; 37). Subsequently, researchers proposed various composite ABSA-related tasks, such as aspect-opinion pair extraction (38; 39), aspect sentiment triplet extraction (40; 41; 42; 43), and structured opinion mining (44; 45). To further refine ABSA tasks, aspect-category-opinion-sentiment quadruple extraction (5; 46; 47) and comparative opinion quintuple extraction (48) have also garnered considerable attention. Recently, Li et al.(4) introduced the dialogue-level aspect-based sentiment quadruple extraction task. Furthermore, some works focus on event-based sentiment analysis without opinion terms (49; 50; 51; 52).

5. 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

[1] 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, "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*, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 6511–6524, Association for Computational Linguistics.

[2] Zi-Yi Dou and Graham Neubig, "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*, Online, Apr. 2021, pp. 2112–2128, Association for Computational Linguistics.

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

[4] Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji, "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*. 2023, pp. 13449–13467, Association for Computational Linguistics.

[5] Hongjie Cai, Rui Xia, and Jianfei Yu, "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)*, Online, Aug. 2021, pp. 340–350, Association for Computational Linguistics.

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

[7] OpenAI, "Gpt-4o system card," 2024.

[8] Anthropic, "Claude 3.7 sonnet," 2025.

[9] DeepSeek-AI, "Deepseek-v3 technical report," 2024.

[10] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel, "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, May 2004, European Language Resources Association (ELRA).

[11] Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen, "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)*, Online, Aug. 2021, pp. 2795–2806, Association for Computational Linguistics.

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

[13] Zhiyang Xu, Jay Yoon Lee, and Lifu Huang, "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*, Toronto, Canada, July 2023, pp. 10421–10437, Association for Computational Linguistics.

[14] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi, "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)*, Hong Kong, China, Nov. 2019, pp. 5784–5789, Association for Computational Linguistics.

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

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

[17] Sha Li, Heng Ji, and Jiawei Han, "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*, Online, June 2021, pp. 894–908, Association for Computational Linguistics.

[18] Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao, "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)*, Vancouver, Canada, July 2017, pp. 1789–1798, Association for Computational Linguistics.

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

[20] Yuqing Yang, Qipeng Guo, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang, "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)*, Toronto, Canada, July 2023, pp. 12876–12889, Association for Computational Linguistics.

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

[22] Xinya Du, Alexander Rush, and Claire Cardie, "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*, Online, June 2021, pp. 909–914, Association for Computational Linguistics.

[23] Kaiwen Wei, Xian Sun, Zequn Zhang, Jingyuan Zhang, Guo Zhi, and Li Jin, "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)*, Online, Aug. 2021, pp. 4672–4682, Association for Computational Linguistics.

[24] Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao, "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)*, Dublin, Ireland, May 2022, pp. 6759–6774, Association for Computational Linguistics.

[25] Chien Nguyen, Hieu Man, and Thien Nguyen, "Contextualized soft prompts for extraction of event arguments," in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, July 2023, pp. 4352–4361, Association for Computational Linguistics.

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

[27] Yuxin He, Jingyue Hu, and Buzhou Tang, "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)*, Toronto, Canada, July 2023, pp. 12542–12556, Association for Computational Linguistics.

[28] Qi Zeng, Qiusi Zhan, and Heng Ji, "EA^{2}E: Improving consistency with event awareness for document-level argument extraction," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States, July 2022, pp. 2649–2655, Association for Computational Linguistics.

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

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

[31] Xinya Du and Claire Cardie, "Event extraction by answering (almost) natural questions," in *Proceedings of the 2020 Conference on Empirical Methods in Natural

Language Processing (EMNLP), Online, Nov. 2020, pp. 671–683, Association for Computational Linguistics.

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

[33] Zijin Hong and Jian Liu, "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*. 2024, pp. 9025–9038, Association for Computational Linguistics.

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

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

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

[37] Xin Li, Lidong Bing, Piji Li, and Wai Lam, "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*. 2019, AAAI'19/IAAI'19/EAAI'19, AAAI Press.

[38] He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue, "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*, Online, July 2020, pp. 3239–3248, Association for Computational Linguistics.

[39] Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li, "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*. 2021, pp. 3957–3963, ijcai.org.

[40] Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si, "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*. 2020, pp. 8600–8607, AAAI Press.

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

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

[43] You Li, Xupeng Zeng, Yixiao Zeng, and Yuming Lin, "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*, New York, NY, USA, 2024, SIGIR '24, p. 619–629, Association for Computing Machinery.

[44] Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji, "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)*, Dublin, Ireland, May 2022, pp. 4232–4241, Association for Computational Linguistics.

[45] Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji, "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*, 2022, vol. 36, pp. 11513–11521.

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

[47] Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji, "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*, Lud De Raedt, Ed. 7 2022, pp. 4121–4128, International Joint Conferences on Artificial Intelligence Organization, Main Track.

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

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

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

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

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