Appendix

1 Dataset

To better understand the experimental datasets, we provide detailed statistics of the corpus. Table 1(a) shows the statistics of the textual features, and Table 1(b) shows the categorical distribution of the datasets. Note that we exactly follow the experimental settings of the previous works [5, 1], and we also mainly focus on the results of the relative majority categories of CT-, CT+, and PS+.

Table 1: Statistics of the datasets.

(a) Statistics of textual features of the datasets.

Statistic	English	Chinese
Average document length	467.25	716.38
Average sentence length	14.73	29.00
Average sentence number in documents	17.49	15.56
Number of documents	1,727	4,649

(b) Statistics of categorical distribution of the datasets.

Statistic	English	Chinese
Number of CT- documents	279 (16.16%)	1,342 (28.87%)
Number of CT+ documents	1,150 (66.59%)	2,403 (51.69%)
Number of PS+ documents	274 (15.87%)	848 (18.24%)
Number of PS- documents	12 (0.69%)	36 (0.77%)
Number of Uu documents	12 (0.69%)	20 (0.43%)
Number of total documents	1,727	4,649

2 Hyperparameters

Our implementation is based on PyTorch[1], experimented on a machine with NVIDIA TESLA T4 GPUs. For the implementation, we adopt BERT-base model[2] as the textual encoder, which has 12-layers, 768-hiddens, and 12-heads. The initial learning rate is tuned in {2e-5, 3e-5, 5e-5} with linear decay. The batch size is tuned in $\{1, 2, 3, 4\}$ for both datasets. We use AdamW [3] to optimize the model parameters. The number of graph layers is tuned in $\{1, 2, ..., 6\}$. The harmonic factor β_1 and β_2 is tuned in $[0, 1]$ and [1e-4,1e-3], respectively. The scaling factor γ is tuned in [1e-4,1e-2]. For re-implementation, we report our optimal hyperparameters on English and Chinese datasets in Table 2. Note that the optimal hyperparameters are tuned on the validation set by grid search strategy according to the averaged Micro-F1 and Macro-F1 scores.

[1] https://pytorch.org/

[2] https://github.com/huggingface/transformers

Table 2: Optimal hyperparameters of our model.

Hyperparameter	English	Chinese
embedding dimension F	768	768
graph layer number L	2	2
scaling factor γ	1e-3	1e-3
learning rate	1e-5	2e-5
harmonic factor β_1	1	1
harmonic factor β_2	5e-4	5e-4
batch size	4	2
training epoch	15	15

3 Mutual Information Feature Selection

Mutual information (MI) [4] measurement is a canonical feature selection method applicable to supervised classification-style tasks. In this work, we adopt MI to mine textual keywords related to the sentence-level factuality categories. In general, MI mathematically measures how much information the presence/absence of a word t contributes to the sentence factuality category c, which can be formulated as follows:

$$\mathrm{I}(T;C) = \sum_{s_t \in \{1,0\}} \sum_{s_c \in \{1,0\}} P(s_t, s_c) \log \frac{P(s_t, s_c)}{P(s_t)P(s_c)} \tag{1}$$

where T is a random variable taking $s_t = 1$ when the sentence contains word t; and C is a random variable taking $s_c = 1$ when the sentence is in the category c. To implement the MI scores, we rewrite Eq. (1) using Maximum Likelihood Estimation (MLE):

$$\begin{aligned} &\mathrm{I}(T = t; C = c) \\ &= \frac{N_{11}}{N} \log \frac{N N_{11}}{N_{1.} N_{.1}} + \frac{N_{01}}{N} \log \frac{N N_{01}}{N_{0.} N_{.1}} \\ &+ \frac{N_{10}}{N} \log \frac{N N_{10}}{N_{1.} N_{.0}} + \frac{N_{00}}{N} \log \frac{N N_{00}}{N_{0.} N_{.0}}, \end{aligned} \tag{2}$$

where the N with subscripts denote the number of sentences having the value of s_t and s_c. For example, N_{10} is the number of sentences containing the word t (*i.e.*, $s_t = 1$) and not belonging to the category c (*i.e.*, $s_c = 0$). $N_{1.} = N_{10} + N_{11}$ is the number of all the sentences containing the word t (*i.e.*, $s_t = 1$). $N = N_{00} + N_{01} + N_{10} + N_{11}$ is the total number of sentences.

In the implementation, we first use BERT tokenizer to split words into subwords as preprocessing for the convenience of generating textual token features by BERT model. Then, we use MI feature selection method to derive (sub-)words associated with the sentence-level event factuality categories. Instead of removing stop words in traditional text classification researches [6], we find that some general stop words (such as adverbs or auxiliary words) reflect negative or speculative cues in our experiments, such as `not`, `may`, and

would, so we reserve these words and encourage the model to discriminate the importance of them by attention mechanism. Table 5 shows the selected keywords for the factuality categories.

To further investigate the details of the selected keywords in sentences, we conduct a feature selection case study. Table 6 shows examples of sentences with selected keywords. Note that the keywords are selected according to the overall MI scores without categories, since we do not obtain the categories in testing. As shown in the table, the selected keywords tend to reflect valuable sentence-level factuality cues for factuality prediction. With the selected keywords, we construct the document-level factuality hypergraph for better factuality understanding, as introduced in the main content of this paper.

4 Extensive Experiments

To better understand the property of our model, we conduct extensive experiments in this section.

Table 3: Experimental results (%) on model variants of the uncertain hypergraph networks.

Datasets	Methods	Micro-F1	Macro-F1	ΔAvg
English	Entire Model	**90.93**	**87.37**	-
	w/o auxiliary loss	89.94	85.75	↓ 1.31
	w/o attention	89.98	86.04	↓ 1.14
	w/o uncertainty	89.48	86.27	↓ 1.28
	w/o relation	89.42	84.83	↓ 2.03
	w/o hypergraph	86.05	79.44	↓ 6.41
	repl. HGCN	89.21	85.52	↓ 1.79
	repl. ULGN	88.52	84.02	↓ 2.88
	repl. GCN	88.22	84.21	↓ 2.94
	BERT Model	83.53	76.76	↓ 9.01
Chinese	Entire Model	**94.26**	**93.36**	-
	w/o auxiliary loss	93.80	93.21	↓ 0.31
	w/o attention	93.61	92.81	↓ 0.60
	w/o uncertainty	93.52	92.73	↓ 0.69
	w/o relation	93.46	92.76	↓ 0.70
	w/o hypergraph	87.43	86.34	↓ 6.93
	repl. HGCN	93.41	92.62	↓ 0.80
	repl. ULGN	93.39	92.57	↓ 0.83
	repl. GCN	92.75	92.31	↓ 1.28
	BERT Model	85.83	84.28	↓ 8.76

4.1 Analysis on Hypergraph Networks

To investigate the impact of each mechanism in the model, we run variants on the proposed hypergraph neural networks. Here, we provide the full experimental results on both the English and Chinese datasets. The results are shown in Table 3, which shows that:

(1) ***The "attention", "uncertainty", and "relation" mechanisms have a positive impact on the results.*** *"w/o attention"* removes the variance-based attention weights. This mechanism helps to detect valuable node features with higher confidence. *"w/o uncertainty"* removes the variances, keeping only the mean vectors as features. This mechanism measures the uncertainty of local features across sentences, helping to resolve the conflicting event factuality with the degree of feature uncertainty. *"w/o relation"* removes the relational learnable weights, considering all edges as the same relation. This mechanism helps to capture different semantic correlations in the graph, allowing feature aggregation in different ways. (2) ***The proposed hypergraph model can be more beneficial for the DocEFI task than previous typical graph models.*** *"repl. HGCN"* replaces the proposed model with vanilla HGCN. *"repl. ULGN"* and *"repl. GCN"* treat hyperedges as fully connected subgraphs, and on the graph adopt ULGN [1] and GCN respectively. It reflects the hypergraph is more inferential for DocEFI, and the uncertainty manner captures valuable information for event factuality. (3) ***The proposed hypergraph model generally performs well on both benchmarks in different languages.*** From the table, our model renders similar trends on both benchmarks, reflecting that our model is applicable in different language scenarios.

Table 4: Experimental results (%) on structure variants of the constructed document hypergraph.

Datasets	Methods	Micro-F1	Macro-F1	ΔAvg
English	Entire Graph (ours)	**90.93**	**87.37**	-
	w/o node: mention	87.95	83.67	↓ 3.34
	w/o node: sentence	90.23	86.46	↓ 0.81
	w/o node: keyword	89.80	85.61	↓ 1.45
	w/o edge: MOD	87.99	83.82	↓ 3.25
	w/o edge: SOD	89.88	85.91	↓ 1.26
	w/o edge: KOM	89.48	86.05	↓ 1.39
	w/o edge: KOS	89.73	86.33	↓ 1.12
	w/o edge: MS	90.39	86.65	↓ 0.63
	w/o edge: KK,MM,SS	90.19	86.47	↓ 0.82
	w/o hypergraph	86.05	79.44	↓ 6.41
Chinese	Entire Graph (ours)	**94.26**	**93.36**	-
	w/o node: mention	88.73	87.73	↓ 5.58
	w/o node: sentence	93.55	92.82	↓ 0.63
	w/o node: keyword	93.79	93.01	↓ 0.41
	w/o edge: MOD	88.34	87.38	↓ 5.95
	w/o edge: SOD	93.20	92.29	↓ 1.07
	w/o edge: KOM	93.69	93.19	↓ 0.37
	w/o edge: KOS	93.90	93.14	↓ 0.29
	w/o edge: MS	93.39	92.58	↓ 0.83
	w/o edge: KK,MM,SS	93.55	92.68	↓ 0.70
	w/o hypergraph	87.43	86.34	↓ 6.93

4.2 Analysis on Hypergraph Structures

To investigate the effect of each sub-structure in the hypergraph, we run variants on the hypergraph structure. Here we provide the full experimental results on both the English and Chinese datasets. The results are shown in Table 4, which shows that:

(1) ***All nodes retain useful information for the results.*** In particular, the mention nodes retain crucial event information, which is essential for document-level factuality identification. The keyword nodes are effective for the results, highlighting fine-grained factuality cues for the event in each sentence, such as "not", "no" for negative factuality, and "likely", "might" for speculative factuality. The sentence nodes preserve contextual information, which not only preserves factuality cues, but also preserves syntactic and semantic meanings in texts. (2) ***All edges provide useful semantic connections, especially the edges associated with the document node.*** For example, the *"MOD"* and *"SOD"* edges establish the connections between document-level and sentence-level nodes, which help to aggregate valuable information from local sentences to the global document. Furthermore, the *"KOM"* and *"KOS"* edges establish the connections between sentence-level and word-level nodes, which aggregate information from local keywords to the sentences involved. The *"MS"* edges further represent the connections between the mention

nodes and the sentence nodes, providing specific information about the corresponding relationship. The additional *"KK,MM,SS"* edges enclose the nodes with similar roles, also providing useful effects. (3) ***The document hypergraph explicitly strengthens the correlations among the related textual components.*** *"w/o hypergraph"* removes the hypergraph structures but remain uncertainty mechanism on node representations. The hypergraph serves as prior knowledge of the textual structures, benefiting to the document understanding, so we can observe the results decrease significantly when the document hypergraph is removed.

4.3 Impact of Graph Layer Number

To better understand the influence of graph layers, we perform experiments on the number of graph layers, as shown in Figure 1. From the figure we can see that:

(1) ***Our model achieves the best results when the number of graph layers is 2.*** We attribute the reason to the fact that applying more than one layer can help the model to access higher-order neighbors, thus benefiting higher-order messages passing along the hyperedges besides the direct neighbors. (2) ***The results decrease as the number of graph layers increases continuously.*** Such a phenomenon is usually observed in classical graph neural networks [2, 1]. We believe that the graph model may encounter the well-known oversmoothing problem [2] in the convolution process when the number of layers is too large. We leave this general problem for future studies.

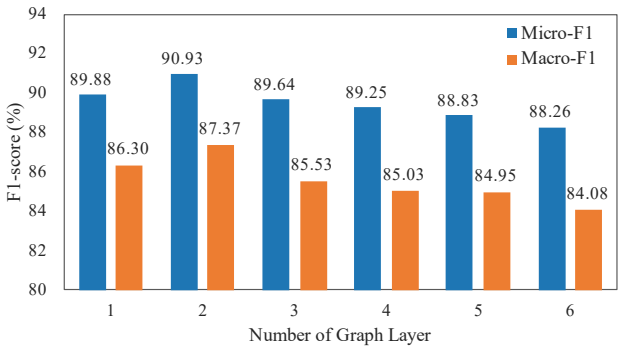

Figure 1: Impacts of number of graph layers on the English event factuality dataset.

References

[1] Pengfei Cao, Yubo Chen, Yuqing Yang, Kang Liu, and Jun Zhao, 'Uncertain local-to-global networks for document-level event factuality identification', in *Proceedings of EMNLP*, (2021).

[2] Thomas Kipf and Max Welling, 'Semi-supervised classification with graph convolutional networks', in *Proceedings of ICLR*, (2017).

[3] Ilya Loshchilov and Frank Hutter, 'Decoupled weight decay regularization', in *Proceedings of ICLR*, (2019).

[4] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, 'Introduction to information retrieval', (2005).

[5] Zhong Qian, Peifeng Li, Qiaoming Zhu, and Guodong Zhou, 'Document-level event factuality identification via adversarial neural network', in *Proceedings of NAACL*, (2019).

[6] C. Silva and B. Ribeiro, 'The importance of stop word removal on recall values in text categorization', *Proceedings of IJCNN*, (2003).

Table 5: Examples of the selected keywords corresponding to different factuality categories.

Dataset	Category	Top 10 words according to MI score
English	PS+	likely, could, expected, plans, be, would, to, may, can, that
	CT+	not, no, likely, n't, could, if, would, expected, or, but
	CT-	not, no, n't, any, do, failed, never, wo, currency, ban
	Overall	not, if, no, likely, n't, could, expected, would, or, do
Chinese	PS+	或, 可能, 拟, 是否, 预计, 希望, 涉嫌, 有望, 计划, 否认 (may, maybe, may, whether, expect, hope, involve, hopeful, plan, deny)
	CT+	否认, 或, 可能, 不, 传闻, 拒绝, 是否, 没有, 拟, 如果 (deny, may, maybe, no, hearsay, refuse, whether, no, may, if)
	CT-	否认, 不, 没有, 传闻, 不会, 未, 拒绝, 传言, 任何, 从未 (deny, no, no, hearsay, won't, not, refuse, hearsay, any, never)
	Overall	否认, 或, 可能, 如果, 不, 没有, 拟, 传闻, 是否, 不会 (deny, may, maybe, if, no, no, may, hearsay, whether, won't)

Table 6: Examples of the selected keywords in sentences of different factuality categories. Note that we select keywords in sentences according to the overall scores, since we cannot access to true event factuality categories of the document in testing phase.

Dataset	Examples		Mention	Category
English	Sentence	I am glad to see more and more countries involved in the forum and sharing their judicial reform experiences, which will no doubt **deepen** mutual understanding and accelerate regional prosperity	deepen	CT+
	Keywords	no, more, see, countries, forum, am, will, regional, which, mutual		
	Sentence	China will not be **misled** by comments of various kinds and will not give up fulfilling its due duties, foreign minister Wang Yi said on Sunday when asked about the Korean peninsula situation.	misled	CT-
	Keywords	not, china, asked, peninsula, foreign, korean, minister, sunday, comments, will		
	Sentence	Germany's social democratic party on Friday announced that it would start exploratory **talks** for another grand coalition government.	talks	PS+
	Keywords	would, government, party, union, conservatives, talks, with		
Chinese	Sentence	然而，约翰逊否认了以上全部指责，并指出："英国正在为加勒比地区前所未有的灾难投入前所未有的援助。" (However, Johnson denied all the above accusations and pointed out that "Britain is investing unprecedented **aid** in the unprecedented disaster in the Caribbean region.")	援助 (aid)	CT+
	Keywords	否认, 指出, 为, 并, 正在, 投入, 然而, 英国, 以上, 指责 (deny, point out, for, and, are, input, however, the UK, above, accuse)		
	Sentence	在回答记者提问时，他说自己从未考虑过辞职。 (In response to a reporter, he said he had never considered to **resign**.)	辞职 (resign)	CT-
	Keywords	考虑, 从未, 辞职, 过, 在, 他, 自己, 回答, 记者, 说 (consider, never, resign, ever, in, he, himself, answer, reporter, say)		
	Sentence	据报道，萨勒曼在4天的访问期间计划会见俄罗斯总统普京，讨论包括石油以及叙利亚在内的若干问题。 (It is reported that Salman plans to **meet** with Russian President Putin during his four-day visit to discuss several issues including oil and Syria.)	会见 (meet)	PS+
	Keywords	计划, 在, 普京, 包括, 据, 讨论, 报道, 叙利亚, 总统, 在内 (plan, in, Putin, include, accord, discussion, report, Syria, President, within)		