The First Span-Level Claim Detection Dataset for Traditional Chinese

Anonymous EMNLP submission

Abstract

The impact of misinformation is an increasingly pressing issue, and with the current method of manually fact-checking proving to be insufficient in keeping up with the sheer volume of misinformation being spread, it has become imperative to automate the fact-checking process. Our focus in this paper is on the first step of the fact-checking process, claim detection, and we aim to address this issue in traditional Chinese by introducing a claim detection dataset, which is the first dataset in traditional Chinese according to our knowledge. Unlike most existing claim detection datasets, we approach this as a span identification task, as we believe this to be a more accurate reflection of the way fact-checkers operate in real life. Using our dataset, we have fine-tuned BERT-based language models as a baseline to achieve automatic claim detection. In response to the rise of GPT, We also try to do achieve this task using GPT by zero-shot learning and compare it against fine-tuning.

1 Introduction

The rise of the internet and social media has simplified access to information but has also amplified misinformation. This situation necessitates fact-checking due to the harmful impact of false information. Numerous global fact-checking organizations like Taiwan FactCheck Center(TFC)[1], PolitiFact[2], and FullFact[3] have been globally established for fight against it.

The overwhelming amount of misinformation, however, challenges human fact-checkers, making automation critical. As proposed by (Barrón-Cedeño et al., 2020), automatic fact-checking involves a four-stage process: claim detection, verified claim retrieval, evidence retrieval, and claim verification, as shown in figure 1.

[1]https://tfc-taiwan.org.tw/
[2]https://www.politifact.com/
[3]https://fullfact.org/

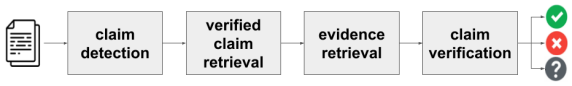

Figure 1: Fact-checking Pipeline

While automated fact-checking is a desirable objective, skepticism persists about machine-generated verdicts' trustworthiness. FullFact's investigation(Nakov et al., 2021) recommends partial automation to assist human fact-checkers. Our work primarily concentrates on automating the claim detection stage, which we believe can be integrated into their fact-checking processes. To accomplish this, we use supervised machine learning for training a claim detection model. The process necessitates a suitable dataset, which, for Traditional Chinese, didn't exist. Therefore, we created a dataset using rumors verified by TFC.

Differing from most of the previous works(Alam et al., 2021; Daxenberger et al., 2017; Gupta et al., 2021) that perceive claim detection as a classification task, we approach it as a span identification task, similar to (Sundriyal et al., 2022). Figure 2 illustrates one of our data instance. Our claim detection task is defined as "***Given an article, identify the spans that contain checkworthy claims***", and we refer checkworthy claim as "***A narrative whose veracity can be checked and whose dissemination could have negative consequences on society***".

根據文獻顯示，武漢病毒主要通過一種叫做Angiotensin-converting enzyme 2，也就是ACE-2的一種外肽酶，作為病毒的受體，來攻擊人體肺部組織！而東亞人體內的ACE-2的濃度，是印歐人的4倍到5倍！也就是說，一旦感染病毒，東亞人，特別是O系漢族的嚴重性，是印歐人的4倍！這是針對O系漢族的病毒！

Figure 2: An instance of our claim detection dataset. The highlighted narrative in red represents the claim span.

Claim detection, especially in non-English languages, is a pressing research topic due to the lack

of resources. We specifically focus on developing a Traditional Chinese claim detection dataset, currently absent in literature. Our main contributions are building and openly release the first Traditional Chinese claim detection dataset[4].

2 Related Work

In this section, we provide an overview of related work on fact checking and claim detection, with a focus on the datasets that have been proposed.

2.1 Fact-Checking

Fact-checking, a process typically undertaken by journalists or experts, verifies the accuracy of public discourse claims. This process involves stages such as claim detection, evidence retrieval, and fact verification (Barrón-Cedeño et al., 2020). The annual FEVER (Fact Extraction and VERification) workshop, initiated in 2018, focuses on automating fact-checking[5]. Two dataset versions were released for this purpose, one in 2018 (Thorne et al., 2018) and another in 2021 (Aly et al., 2021), each with shared tasks concentrating on fact verification in the fact-checking pipeline.

Moreover, some approaches aim to handle the entire fact-checking pipeline at once. ClaimBuster (Hassan et al., 2017), an automated system, performs end-to-end fact-checking, comprising submodules like Claim Monitor, Claim Spotter, Claim Matcher, and Claim Checker.

2.2 Claim Detection

Claim detection identifies claims within a text. Prior than fact-checking, It's also a subtask in argument mining, which focusing on extracting arguments from text. Levy et al. (Aharoni et al., 2014) pioneered claim detection, releasing a dataset for argument mining, which includes claims and supporting evidence from Wikipedia articles on controversial topics. They later developed a logistic regression classifier, focusing on features essential for claim detection (Levy et al., 2014).

ClaimBuster's Claim Spotter module (Hassan et al., 2017) trains an SVM to classify sentences as claims or non-claims. They build a claim detection dataset of 20,617 sentences from US presidential debates spanning 1960 to 2012. Other fact-checking works have adopted this system as a basline(Gangi Reddy et al., 2022; Li et al., 2022).

The annual CheckThat! Lab competition [6] since 2018 addresses three subtasks: claim detection, evidence retrieval, and fake news detection. The open-source dataset used for these tasks and its construction process offer valuable insights (Alam et al., 2021).

In contrast to treating claim detection as a binary classification task, (Sundriyal et al., 2022) interprets claim detection as a span identification task. They re-annotated tweets from the dataset proposed by (Gupta et al., 2021) to generate the exact claim span, which named CURT. Having dataset, they train a model to pinpoint the claim boundary in the tweet. This approach aligns most closely with our task objective.

3 Dataset

In this section, we provide a comprehensive description of our dataset construction process, from data sourcing to labeling.

3.1 Data Collection & Filtering

We sourced our data from the Taiwan FactCheck Center (TFC). TFC collects rumors disseminated online, primarily via two popular social media platforms in Taiwan, LINE and Facebook. After fact-checking these rumors, TFC issues a detailed report for each rumor, outlining their verification process. These reports, published on their website[7], also include the original rumor in the background section, which served as our raw data source. We collected this data using TFC's API, ranging from July 2018 to March 2022, yielding a total of 2914 rumors.

Next, we implemented a filtering process to refine our dataset. Initially, we excluded multimodal rumors—those containing non-text information—as our focus is strictly on text-based claim detection. Subsequently, we removed rumors containing URL links because the claims could either reference the context of the link or be embedded within the link, rendering the text insufficient for accurate claim detection. Lastly, rumors without any Chinese characters were eliminated. This filtration resulted in a collection of 1178 rumors for further annotation.

3.2 Annotation

We established an annotation process, as depicted in Figure 3. Annotators reviewed each rumor, scru-

[4] https://github.com/jason50706/CDDTC
[5] FEVER workshop website
[6] CLEF CheckThat! Lab website
[7] https://tfc-taiwan.org.tw/articles/8721

tinizing the sentences for check-worthy claims. If a check-worthy claim was identified, its precise boundary was highlighted.

To aid annotators, we developed two guiding questions. The first question—"Can the narrative be verified through supporting or refuting evidence?"—served to determine if the narrative qualifies as a claim. If the response was affirmative, we posed the second question—"Could the claim's validity impact society?"—to ascertain the claim's check-worthiness. The annotation process was facilitated by a platform developed by Academia Sinica[8].

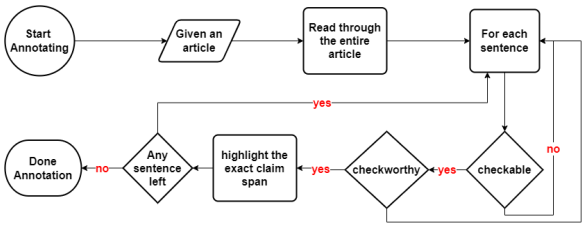

Figure 3: Annotating Process

To ensure data integrity and limit subjectivity, we adopted a crowdsourcing approach. A diverse group of 114 annotators, aged between 16 and 24 years, comprising high school, undergraduate, and master's students, contributed to our dataset. Each rumor was annotated by three randomly assigned annotators to minimize personal bias and secure a broad representation. We quantified the inter-annotator agreement using the Fleiss Kappa score(Fleiss, 1971) at the token level, which indicated a moderate agreement of 0.41.

Each rumor was annotated by three individuals, resulting in three sets of highlighted spans per rumor. A majority vote at the character level was used to generate the final labeled claim span for each rumor. Our labels, assigned at the character level, utilized the markers B, I, or O to denote the beginning of a claim span, inside a claim span, or outside any claim span, respectively.

3.3 Dataset Statistics

In line with (Sundriyal et al., 2022), we compiled statistics about our dataset (Table 1). Our dataset stands out due to the average rumor length and the average span length, which surpasses those in most existing claim detection datasets. This discrepancy primarily arises because many of these datasets(Sundriyal et al., 2022; Alam et al., 2021; Gupta et al., 2021) use tweets, which are inherently length-limited. This difference poses an added challenge but also contributes to the realism of our dataset, reflecting real-world claim detection scenarios. Our dataset comprises 619 articles with 1611 check-worthy claims.

[8] Annotation platform developed by Academia Sinica

Table 1: Dataset Statistic comparing to CURT (Sundriyal et al., 2022)

	Ours	CURT
#Rumors	619	7555
Avg. length of rumors	458.75	27.34
Avg. length of spans	54.12	10.89
# span per rumor	2.62	1.25
#Single-span rumors	247	6039
#Multiple-span rumors	372	1483

4 Model

In this section, the model architecture for span-level claim detection is presented. We utilize the structure proposed in (Li et al., 2021), which views span detection as a sequence labeling problem. Despite the paper addressing a distinct task, the input and output format aligns with our needs, facilitating the application of their architecture to our dataset.

The architecture, illustrated in Figure 4, consists of an encoder and a classification layer. The encoder translates each token in input texts into a representation vector, while the classification layer classifies each token's representation into B, I, or O labels.

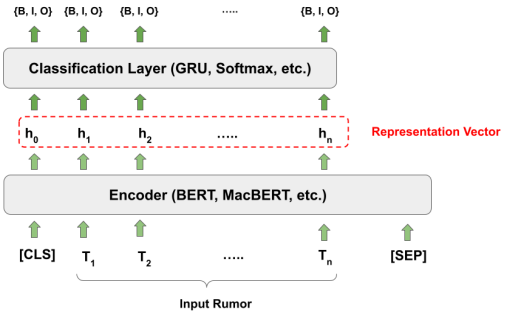

Figure 4: The model architecture used for span-level claim detection.

This architecture offers flexibility, allowing for different implementations. For instance, in (Li et al., 2021), BERT and SpanBERT are used as encoders, and FFN (Feed Forward Neural Network), CRF (Conditional Random Field), and RNN (Recurrent Neural Network) are tested at the classification layers. In our experiments, we explore BERT(Devlin et al., 2019), MacBERT(Cui et al.,

Table 2: Experiment Results. The F1, P, R on the left is macro-level score (average of claim and non-claim class). Always B+I represent the method that predict all the tokens as claim. Always O represent the method that predict all the tokens as non-claim.

Model name	F1	P	R	F1		Precision		Recall		Accuracy
				B+I	O	B+I	O	B+I	O	
Always B+1	24.34	66.09	50	48.69	0	32.18	100	100	0	32.18
Always O	40.41	83.91	50	0	80.83	100	67.82	0	100	67.82
BERT + Softmax	62.17	66.16	67.78	58.6	65.74	45.72	**86.61**	**82.19**	53.36	62.64
MacBERT + Softmax	62.71	66.16	67.94	**58.67**	66.76	46.14	86.19	81.04	54.84	63.27
RoBERTa + Softmax	66.61	66.59	68.3	57.88	75.35	51.56	81.62	66.42	70.18	68.97
BERT + CRF	**67.35**	**68.12**	67.42	55.59	**79.11**	**57.16**	79.09	55.33	**79.52**	**71.74**
MacBERT + CRF	65.38	66.74	67.27	56.29	74.47	52.36	81.13	64.26	70.21	68.3
RoBERTa + CRF	66.45	67.18	**68.49**	57.89	75.01	51.95	82.27	67.25	69.73	68.93
GPT-3.5-trubo	60.10	59.88	60.63	45.92	74.28	42.96	76.8	49.32	71.93	65.14
GPT-4	66.3	65.91	68.00	55.82	76.78	49.45	82.37	64.09	71.91	69.56

2020), and RoBERTa(Liu et al., 2019) as encoders, and FFN and CRF as classification layers. FFN projects each token representation vector into the dimension of the label set size via a fully connected layer before applying softmax to convert it into a probability distribution among labels. Conversely, CRF(Lafferty et al., 2001), a traditional method for sequence labeling tasks, considers the entire sequence while labeling tokens.

Capitalizing on the rise of GPT, we have also tried to employ it for this task through zero-shot learning. The prompt we use and the post-processing to convert GPT's response to labels can be referred to in Appendix A.2.

5 Experiments

In this section, we showing the experiment result of various combinations of encoders and classification layers. We allocated 70%, 15%, and 15% of the dataset to training, validation, and testing sets, respectively. For evaluation, we followed the metric used in (Sundriyal et al., 2022), calculating the precision (P), recall (R), and F1 score at the token level. We merge B and I when computing these scores on classes because we only concerned with the scores corresponding to claim and non-claim, making it unnecessary to distinguish between B and I since they both correspond to claim.

Each result from the BERT-based encoder + classifier in Table 2 represents the average of three runs. For each run, we trained the model for ten epochs using the AdamW optimizer with a learning rate of 1e-5 on a single RTX3090. For the GPT result, we only obtained the test set result once, since we set the temperature to 0 to maximally reduce the output's randomness. The combination of BERT as the encoder and CRF as the classification layer yielded the best macro F1 score. While the superiority of CRF over Softmax in sequence labeling tasks is well-known, the unexpected finding that BERT as the encoder produces better results than MacBERT and RoBERTa could be due to the relatively small size of our dataset. On the other hand, we observed that GPT-4 performs on par with most of the fine-tuned models, and even outperforms some. This was unexpected, considering it was achieved using zero-shot learning, further investigation would be beneficial to understand the reasons behind this performance.

While the performance of GPT-4 is on par with that of the fine-tuned BERT-based model, the cost associated with its use makes it an important consideration. Furthermore, in terms of flexibility, when using GPT through zero-shot learning, users cannot effectively alter the model's predictions, as these depend on the prior knowledge encoded in the pre-training data. In contrast, when fine-tuning a model, users can adjust its predictions by revising labels or adding new data instances.

6 Conclusion

Our work has laid the groundwork for managing claim detection in Traditional Chinese by introducing a novel dataset. As this was primarily a data construction endeavor, we recognize that the model's performance can be significantly improved. This dataset not only serves as a valuable resource for future studies but also provides a foundation for more advanced approaches to enhance claim detection performance. We envision that this contribution will foster further research in this field, driving the development of more robust, automated fact-checking technologies for Traditional Chinese.

Acknowledgements

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. In *Advances in Information Retrieval*, pages 499–507, Cham. Springer International Publishing.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. 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.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Fung, Kathryn Conger, Ahmed ELsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, and Heng Ji. 2022. NewsClaims: A new benchmark for claim detection from news with attribute knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6002–6018, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shreya Gupta, Parantak Singh, Megha Sundriyal, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. LESA: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3178–3188, Online. Association for Computational Linguistics.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1803–1812, New York, NY, USA. Association for Computing Machinery.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Manling Li, Revanth Gangi Reddy, Ziqi Wang, Yi-shyuan Chiang, Tuan Lai, Pengfei Yu, Zixuan Zhang, and Heng Ji. 2022. COVID-19 claim radar: A structured claim extraction and tracking system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 135–144, Dublin, Ireland. Association for Computational Linguistics.

Xiangju Li, Wei Gao, Shi Feng, Yifei Zhang, and Daling Wang. 2021. Boundary detection with BERT for span-level emotion cause analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 676–682, Online. 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.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Empowering the fact-checkers! automatic identification of claim spans on Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7701–7715, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

A.1 Preprocessing of Long Articles

Given the length limit imposed by the BERT base encoder we use, it's necessary to preprocess articles in our dataset that exceed this limit - accounting for 26.66% of all articles. To address this, we divided such articles into sections, each consisting of no more than 510 tokens. This ensures an overall length of 512 tokens, considering the additional CLS and SEP tokens added later. During training and prediction, each slice was separately fed into BERT as an individual data instance.

A.2 GPT Usage

For each GPT-based model, we engage a methodology known as "role prompting." In this approach, we set the role of the system by using the phrase "*Imagine you are a rigorous fact-checker.*" This essentially instructs the GPT model to function as a meticulous fact-checker. Additionally, we prepend the phrase "*Please select the parts in the message below that are both 'verifiable' and 'checkworthy' and list them as separate sentences:*" before the article. This prompt guides the GPT to output checkworthy claims. Figure 5 provides an example of this application. Regarding hyperparameters, we set the temperature to 0 for the most deterministic output.

```
openai.ChatCompletion.create(
 model="gpt-3.5-turbo",
 temperature=0,
 messages=[
     {"role": "system", "content": "想像你是一個嚴謹的事實查核人員"},
     {"role": "user", "content": "請將下面訊息中\"可以查核\"且\"值得查核\"的部分挑選出來, 並以文中的句子條列:\n"
                  "通告各位好友, 我見已確定:你若有辦敬老卡每人每個月乘坐交通工具補助480元沒用完, "
                  "可在月底前去便利商店買東西把他用完以免浪費。今天才知道, 平白浪費不少!"},
 ]
)
```

Figure 5: The application of GPT. Texts highlighted in red are the prompts instructing GPT, while texts highlighted in blue represent the input article.

The responses generated by GPT models are in natural language, as shown in Figure 6. However, our evaluation metric requires the exact span of the claim within the input article. To reconcile this difference, we implemented a two-step post-processing approach.

Firstly, we separate the predicted claims from the GPT response. We achieve this by leveraging the model's tendency to initiate each claim on a new line - thereby splitting the GPT response using '\n'. This step is necessary as it aids in isolating individual claims for further processing.

Secondly, we handle situations where the GPT response includes words not present in the original article, as demonstrated in Figure 6. For this, we employed the Longest Common Sequence (LCS) algorithm, a method used to find the most extensive sequence common to both the predicted claim and the input article. This step is crucial as it helps in mapping the predicted claims back to the original text.

By implementing these two steps, we effectively convert each GPT-predicted claim into token-level labels, which enables the direct application of our evaluation metric to the claims.

A.3 Result Analysis

Our in-depth analysis of various models' results reveals intriguing patterns and provides valuable insights into each model's performance.

In the example provided in Table 3, despite our system's role prompt being heavily focused on 'rigorous' claims, GPT3.5-turbo interprets all sentences as claims. In contrast, GPT4 and RoBERTa+Softmax identify words with emotional connotations as claims. Only BERT+CRF pro-

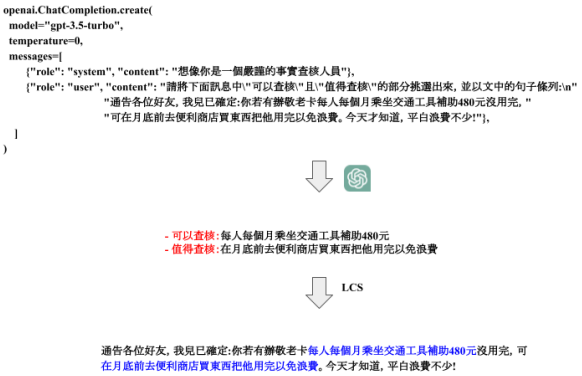

Figure 6: GPT response. Each claim predicted by GPT is displayed on a single line. The text highlighted in red indicates parts not originating from the article, while the text highlighted in blue represents the token-level result after conversion by the LCS.

duces a response that aligns closely with the actual claim. The difference in outcomes could be attributed to the advantage of BERT+CRF that learns to align with the human annotators' conception of claims during the fine-tuning process.

The example in Table 4 presents a different picture, where GPT3.5-turbo marks a significantly fewer number of claims, overlooking vital information. The other models identify the key locations and events as claims, with RoBERTa+Softmax and BERT+CRF presenting more succinct results. Again, the ability to align with human annotators' perspectives could explain why the fine-tuned models (RoBERTa+Softmax and BERT+CRF) perform better in this case.

While this analysis showcases the strength of fine-tuned models, it is important to acknowledge scenarios where GPT3.5-turbo or GPT4 outperforms them. For example, GPT-based models, due to their language generation capabilities, often perform better in scenarios involving narrative or storytelling style claims.

In conclusion, understanding the strength and weaknesses of each model is crucial. It aids in choosing the right tool for specific tasks and illuminating potential improvements for future work. The fine-tuned models' ability to better align with human annotators' perspectives on what constitutes a claim emerges as a significant factor in their performance.

A.4 Limitations

While this study makes significant strides towards addressing the issue of claim detection in traditional Chinese texts, it is not without limitations.

Table 3: Example of prediction of different models

Ground Truth	日本6月4日送台灣的AZ疫苗是2021年5月已經過期了,菜政腐就是要讓台灣人滅族,些日畜後代在跪舔日蝗,菜菜子下台！！
GPT3.5-turbo	日本6月4日送台灣的AZ疫苗是2021年5月已經過期了,菜政腐就是要讓台灣人滅族,些日畜後代在跪舔日蝗,菜菜子下台！！
GPT4	日本6月4日送台灣的AZ疫苗是2021年5月已經過期了,菜政腐就是要讓台灣人滅族,些日畜後代在跪舔日蝗,菜菜子下台！！
RoBERTa + Softmax	日本6月4日送台灣的AZ疫苗是2021年5月已經過期了,菜政腐就是要讓台灣人滅族,些日畜後代在跪舔日蝗,菜菜子下台！！
BERT + CRF	日本6月4日送台灣的AZ疫苗是2021年5月已經過期了,菜政腐就是要讓台灣人滅族,些日畜後代在跪舔日蝗,菜菜子下台！！

Firstly, the effectiveness of our models is intrinsically tied to the quality and representativeness of our dataset. Although the dataset is extensive, it may not comprehensively capture all nuances of traditional Chinese or the diversity of misinformation. Its performance with rare or unseen contexts remains uncertain. Moreover, because the annotation process depends on human annotators, potential biases or errors might influence the reliability of the dataset and, consequently, the model's training.

Secondly, we used pre-trained models like BERT and GPT, fine-tuned to our specific task. These models are trained on large, diverse corpora, and while they demonstrate a broad understanding of language, they may not be fully attuned to the specifics of claim detection or the nuances of misinformation in different contexts.

Thirdly, while we found GPT-4 to be surprisingly effective as a zero-shot learning model, the financial and computational costs associated with using GPT models make them less feasible for widespread real-world applications, especially on a large scale.

Lastly, the approach of treating claims as span detection may not cover all instances of misinformation. Some false information can be conveyed subtly or implicitly, without a clearly identifiable span of text to fact-check. Our current methodol-

Table 4: Example of prediction of different models

Ground Truth	搶錢的！工商服務報告：從嘉義往太保的高鐵大道，過第一漁市場後的第一段上坡至下坡路段，昨日警察已安裝「區間測速」7-8秒，不得超速70+10公里間距，超速罰款3600元，每增加一公里，罰款多100元，累計罰款，請各位有經過此路段，請放慢速度依速限行駛，小心駕駛，以「測」安全！
GPT3.5-turbo	搶錢的！工商服務報告：從嘉義往太保的高鐵大道，過第一漁市場後的第一段上坡至下坡路段，昨日警察已安裝「區間測速」7-8秒，不得超速70+10公里間距，超速罰款3600元，每增加一公里，罰款多100元，累計罰款，請各位有經過此路段，請放慢速度依速限行駛，小心駕駛，以「測」安全！
GPT4	搶錢的！工商服務報告：從嘉義往太保的高鐵大道，過第一漁市場後的第一段上坡至下坡路段，昨日警察已安裝「區間測速」7-8秒，不得超速70+10公里間距，超速罰款3600元，每增加一公里，罰款多100元，累計罰款，請各位有經過此路段，請放慢速度依速限行駛，小心駕駛，以「測」安全！
RoBERTa + Softmax	搶錢的！工商服務報告：從嘉義往太保的高鐵大道，過第一漁市場後的第一段上坡至下坡路段，昨日警察已安裝「區間測速」7-8秒，不得超速70+10公里間距，超速罰款3600元，每增加一公里，罰款多100元，累計罰款，請各位有經過此路段，請放慢速度依速限行駛，小心駕駛，以「測」安全！
BERT + CRF	搶錢的！工商服務報告：從嘉義往太保的高鐵大道，過第一漁市場後的第一段上坡至下坡路段，昨日警察已安裝「區間測速」7-8秒，不得超速70+10公里間距，超速罰款3600元，每增加一公里，罰款多100元，累計罰款，請各位有經過此路段，請放慢速度依速限行駛，小心駕駛，以「測」安全！！

ogy may struggle to detect such implicit misinformation.

Future work should aim to address these limitations, for instance, by further diversifying the dataset, improving the model's contextual understanding, exploring cost-effective alternatives for real-world applications, and refining the methodology to detect more subtle forms of misinformation.

A.5 Application

Our research's practical application extends well beyond the realms of academia. Alongside developing a novel dataset and training sophisticated models, we have constructed a user-friendly web platform leveraging these trained models for real-time claim detection. As shown in Figure 7, the operation of this platform is simple and intuitive.

Users — who could be journalists, researchers, or everyday citizens — can effortlessly input the text they wish to analyze for potential misinformation into the top input box. By clicking the "SUBMIT" button, they engage our model which processes the text and presents the results in the bottom box, distinctly highlighting the spans predicted as claims.

Importantly, we're not just limiting this tool to a laboratory context. We're in the process of launching this service on a public platform. We believe that this transition will expand its reach, serving a wider audience and thereby playing a vital role in the fight against misinformation. Further, as more users engage with the platform, we will be able to gather more data regarding potential claims for verification, thereby continually improving the model's performance and reliability. Thus, our research is not just confined to theory; it translates into a practical tool, demonstrating our commitment to aiding society in mitigating the pernicious effects of misinformation.

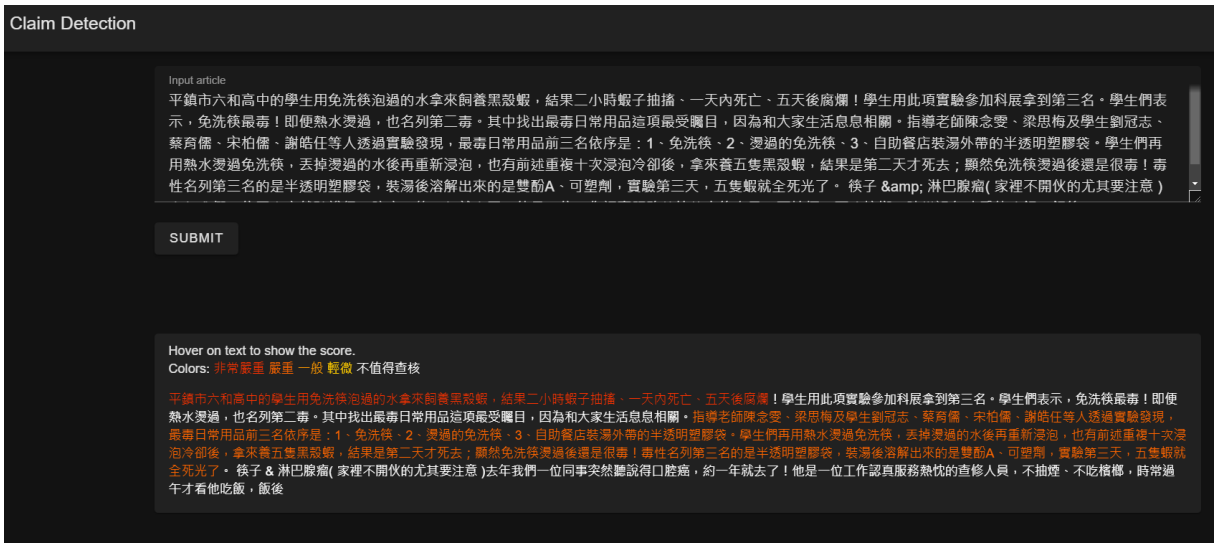

Figure 7: The platform we developed. We use four colors to highlight the claims, each specifying a different level of severity. The severity ranges from low to high, indicated by the color scheme from yellow to red.