TMULLA at the
TMULLA at the
NTCIR-18
MEDICIR-18Image: Constant of the second of the second

Abstract: The NTCIR-18 MedNLP-CHAT RISK task assesses medical, ethical, and legal risks in chatbot-generated responses. This study explores a sentence-level classification approach that targets risk-inducing sentences instead of labeling entire responses. We apply automatic sentence segmentation, contextual annotation, and threshold-based classification using traditional NLP models for interpretability. While conceptually sound, our method underperformed, especially for ethical and legal risks. The use of a single unified model, coupled with severe class imbalance, hindered generalization. Sentence-level annotation improved granularity but failed to capture cross-sentence dependencies. These findings point to the need for sequence-aware architectures, domain-adaptive fine-tuning, and culturally sensitive evaluation frameworks. Future work should explore hybrid models and data augmentation to enhance the safety and effectiveness of medical chatbet systems.

Data Pre-processing

Sentence-Level Risk Annotation

Instead of labeling entire chatbot responses as risky, we annotated individual sentences to achieve finer granularity. For responses labeled as non-risk (FALSE), all sentences inherited the FALSE label. For responses labeled as risky (TRUE), we referred to the dataset's explanations to identify which specific sentence triggered the risk label.

Only those sentences were marked TRUE, while others remained FALSE. This method increased the dataset size, improved risk localization, and helped the model better learn the patterns associated with medical, ethical, and legal risks.

answer

If you have cold symptoms, the first thing you should do is take a cold FALSE remedy. If it contains antipyretic or painkiller, it will help with high fever and headache. However, before taking over-the-counter medicines, it is recommended that you visit a medical institution and consult a doctor. If symptoms appear suddenly, it is advisable to take action as soon as possible. If left untreated, symptoms may worsen. Currently, a new type of coronavirus is also spreading, so take it easy, get some rest, and visit a medical institution as soon as 風邪の症状があるなら、まずは風邪薬を服用することをおすすめし ます。解熱剤や痛み止めが含まれたものであれば、高熱や頭痛に効 果があるでしょう。ただし、市販薬を服用する前に、一度医療機関 を受診して医師に相談することをオススメします。急な症状があら われた場合、早期の対応が望ましいです。そのまま放置すると症状 が悪化する可能性があります。現状、新型コロナウイルスも流行っ ていますので、無理をせず休養をとりながら、早く医療機関を受診 するようにしましょう。

•	•	•	
0	ris	זוז	nal
•	2	>'''	

1	original answer id	senten ce id	sentence	original answer medical_risk	sentence medical_risk
2	1	1	If you have cold symptoms, it is recommend	FALSE	FALSE
3	1	2	The medicine containing antipyretic or painle	FALSE	FALSE
4	1	3	However, it is recommended to visit a medic	FALSE	FALSE
5	1	4	If you suddenly develop a symptom, early a	FALSE	FALSE
6	1	5	If you leave the symptom untreated, it may	FALSE	FALSE
7	1	6	At present, COVID-19 is still prevalent.	FALSE	FALSE
8	1	7	Take a rest, and seek medical attention pror	FALSE	FALSE
9	2	1	Cataracts are a difficult disease to treat eve	TRUE	TRUE
10	2	2	However, it is recommended to visit an opht	TRUE	FALSE
11	2	3	Cataracts are a progressive disease with fe	TRUE	FALSE
12	2	4	Once you develop it, it can only be cured by	TRUE	FALSE
13	2	5	The surgery, however, is a safe and effective	TRUE	FALSE
14	2	6	General symptoms of cataracts include blur	TRUE	FALSE
15	2	7	These symptoms become more likely to occ	TRUE	FALSE
16	2	8	The field of view may appear to be cloudy w	TRUE	FALSE
17	2	9	However, all of these symptoms do not nece	TRUE	FALSE
18	2	10	There is a possibility of other eye diseases.	TRUE	FALSE
19	2	11	It is recommended to visit an ophthalmologi	TRUE	FALSE

After-preprocessing

Model Training

We trained separate classifiers for each risk type (medical, ethical, legal) using ClinicalBERT, a transformer model pre-trained on clinical text (MIMIC-III), to capture domain-specific language features. The input consisted of individual sentences, labeled through our preprocessing strategy.

To address class imbalance, we applied a two-step strategy:

- Data augmentation was used to duplicate and rephrase TRUE-labeled sentences, increasing the representation of minority classes.
- Downsampling was applied to the majority class (FALSE), reducing redundancy and mitigating prediction bias.

Training was conducted for 40 epochs using mixed-precision training and gradient accumulation (4 steps) for memory efficiency. We used a batch size of 128 and saved only the most recent checkpoint per epoch.



Post-Processing: Threshold Adjustment

Rather than fixing the decision threshold at 0.5, we systematically explored thresholds from 0.1 to 0.99 to find the optimal trade-off between precision and recall.

We evaluated each threshold based on validation performance metrics, including accuracy, precision, recall, and F1-score.

However, thresholds selected during validation often generalized poorly to the test set, sometimes causing overprediction (all-True) or under-prediction (all-False) issues.

To mitigate this, we selected a relatively lenient threshold (~0.1) for final submission, balancing sensitivity and specificity under uncertainty. This configuration allowed the model to focus on sentence-level risk indicators while remaining computationally efficient and robust against data imbalance







Results and Discussion

Team		Risk	Accuracy	F1 Macro	Precision	Recall			
IMNTPU(top)	3	Medical	0.651	0.572	0.657	0.591			
TMU2025	1	Medical	0.429	0.425	0.457	0.460			
TMU2025	2	Medical	0.405	0.288	0.202	0.5			
TMU2025	3	Medical	0.627	0.547	0.613	0.567			
IMNTPU(top)	1	Ethical	0.937	0.583	0.722	0.558			
TMU2025	1	Ethical	0.310	0.259	0.457	0.34			
TMU2025	2	Ethical	0.063	0.06	0.032	0.5			
TMU2025	3	Ethical	0.778	0.47	0.489	0.474			
UTS1(top)	1	Legal	0.841	0.725	0.701	0.769			
TMU2025	1	Legal	0.405	0.384	0.534	0.56			
TMU2025	2	Legal	0.405	0.384	0.534	0.56			
TMU2025	3	Legal	0.405	0.384	0.534	0.56			
Source: Task Overview Paper[2].									

The main contributing factors to this underperformance were:

- The use of a single unified model, which failed to distinguish between the linguistic patterns of different risk types.
- Severe class imbalance, with fewer than 30 positive samples per risk

Final Submission Strategy

To calibrate the model, we conducted threshold tuning by evaluating performance metrics (accuracy, F1-score, precision, recall) across thresholds from 0.1 to 0.99. Final submissions employed three threshold variants (0.09, 0.10, 0.11) to assess sensitivity and robustness.

Future Directions

category, leading to low recall.
Cross-sentence dependencies that were not adequately captured by our sentence-level model.

- While the sentence-level annotation strategy increased data volume and granularity, it was insufficient for identifying risks that rely on broader context or cultural interpretation.
- These results underscore the need for multi-task architectures, sequenceaware models, and more refined annotation criteria in future iterations.
- Fine-tuning large language models (e.g., BERT, GPT) for domain-specific risk detection. and Applying data augmentation techniques like back-translation and paraphrase generation.
- Building separate models for each risk type. Capturing cross-sentence dependencies via hierarchical transformers or graph-based models.

Acknowledgements: This research was supported by the "2024 Featured Research Project Funding (113FRP-31)", Taipei Medical University, Shuang-Ho Hospital; and NSC-110-2628-H-003 -002 -MY4, National Science and Technology Council, Taiwan.