

HpiVaxVigil at #SMM4H-HeaRD 2025: Detecting Adverse Vaccine Events on Reddit using GPT-4o Chain-of-Thought Reasoning and Fine-Tuned PLMs with Stratified Cross-Validation

Akhyar Ahmed¹, Esther-Maria Antao¹

¹Digital Global Public Health
Hasso Plattner Institute für Digital Engineering gGmbH
Prof.-Dr.-Helmert-Straße 2-3
14482 Potsdam, Germany
akhyar.ahmed@hpi.de, esther.antao@hpi.de

Abstract

Detecting vaccine adverse events in noisy social media texts is both challenging and crucial to timely pharmacovigilance. In this paper, we detail our participation in Task 6 of the Workshop #SMM4H (Social Media Mining for Health) 2025, specifically focusing on potentially identifying personally experienced vaccine adverse reactions from unstructured social media posts. We present an enhanced hybrid methodology that combines the sophisticated chain-of-thought (CoT) reasoning of GPT-4o and the enhancement of domain-specific knowledge with pre-trained language models (PLM) such as BERTweet-large and DeBERTa-v3-base. By systematically employing stratified cross-validation, robust regularization (mixout and layer-wise learning rate decay), and conditional text enhancement, our model significantly outperforms the previous benchmarks, achieving an F1 score of 0.96 on both validation and test sets, underscoring the effectiveness of integrating human-like reasoning capabilities with traditional classification models.

Code — GitHub Repository

Datasets — Dataset Folder

Introduction

In the digital age, social media platforms have become powerful spaces where individuals share personal experiences, concerns, and subjective opinions about health, often in real time. Among these discussions, conversations about vaccines are particularly prominent, ranging from personal accounts of side effects to broader expressions of vaccine skepticism and public fear. In addition, these data are inherently noisy and unstructured, and they present a unique opportunity for public health surveillance. During the COVID-19 pandemic, a new vaccine was introduced using technology that had not been widely deployed prior to the global crisis. Attitudes towards the COVID-19 vaccine shifted over time (Betsch 2021; Lazarus et al. 2024). This change was influenced by the evaluation of scientific knowledge, fluctuating public perceptions, the spread of misinformation, and reports of adverse vaccine reactions (Lazarus et al. 2024). Analyzing social media content, particularly public perceptions and opinions on sensitive topics such as vaccination,

through approaches such as the emerging practice of social listening (Boender et al. 2023) can help public health professionals more effectively address misinformation and strengthen health communication strategies for greater impact. It is important to note that public attitudes and perceptions on social media may reveal trends or provide hints. However, they should not be considered scientific evidence for the presence or absence of adverse vaccine reactions. Further investigation is always necessary.

Researchers and public health authorities can find trends in large-scale interactions using natural language processing and data mining. This involves finding new adverse reaction reports, measuring mood and disinformation patterns, and understanding public concerns that traditional reporting systems may miss. Real-time analytics can improve pharmacovigilance, targeted communication, and public health policy responsiveness and transparency.

In contrast, earlier research has investigated many approaches for the detection of vaccine adverse events (VAE) based on social media. Combining topic modeling with classification methods, the VAEM-Mine method extracted VAE from Twitter data (Khademi Habibabadi et al. 2022). Consequently, the bidirectional encoder representations from transformer-based models (BERT) have been used for the detection of adverse events, thus demonstrating the efficiency of transformer topologies in this field (Dong et al. 2024). However, these approaches can struggle with issues such as class imbalance and the need for domain-specific knowledge integration.

In participating in the #SMM4H 2025 workshop (Klein et al. 2025), we present an advanced hybrid pipeline that combines a large language model (LLM) with pre-trained language models (PLM) to effectively address the detection of vaccine adverse event from social media texts. The social media data specifically highlights public perceptions of the Shingles vaccine and its potential effects. Inspired by a recent study (Khademi et al. 2024), which evaluated separate applications of an LLM based on GPT and a PLM similar to the BERT, our approach synergistically leverages their complementary strengths within a unified framework. We use GPT-4o (OpenAI 2024) with a meticulously designed CoT reasoning protocol to annotate Reddit posts, subsequently injecting domain-specific knowledge derived from compre-

hensive exploratory data analysis. This enriched dataset provides contextualized insights that substantially improve the performance of the PLM classifier. Specifically, our conditional text enhancement strategy ensures that posts identified as containing adverse reactions are augmented with domain-specific clarifications, while non-reaction posts remain unmodified to preserve authenticity and reduce noise. Furthermore, we integrate focal loss to effectively address class imbalance, apply rigorous stratified cross-validation, implement robust ensemble learning, and optimize decision thresholds to maximize precision-recall trade-off. Our refined hybrid methodology significantly outperforms the current benchmark (F1 score of 0.95) established by (Khademi et al. 2024), achieving an outstanding F1 score of 0.96 on unseen Reddit vaccine reaction data.

Experiments

This section describes the experimental design, data processing techniques, and model optimization strategies applied to evaluate our methodology for detecting vaccine adverse events from social media texts. To delineate our contributions, we present detailed descriptions of our preprocessing approach, dataset construction, model architectures, and the comprehensive evaluation pipeline, including heuristic labeling, baseline, and enhanced pipelines.

Data Preprocessing

The preprocessing step involved a careful transformation of raw social media posts to enhance clarity and highlight relevant information about vaccine reactions. Initially, every post was excluded from the number digits, lowercase, and punctuation removed (Bird, Klein, and Loper 2009). Subsequently, the texts were tokenized using NLTK’s tokenizer, followed by stopword elimination to retain informative words. Nonetheless, caution was used in the deletion of stopwords to prevent unintentional alteration of semantics, especially with negation terms such as “no” potentially reversing intended meanings (e.g., “no side effects” becoming “side effects”). For data augmentation we require sets of domain-specific vaccine-related keywords. These domain-specific vaccine-related terms were consistently identified using fuzzy keyword matching via RapidFuzz (Bachmann 2022), that addresses user-generated typographical variations. The identified terms related to symptoms, specific vaccines, and vaccination impacts were systematically compiled into domain-specific tags and appended to the relevant social media posts. Finally, we applied a *conditional text enhancement* strategy (See Appendix 1 for better understanding), for each post that GPT-4o’s CoT reasoning labeled as describing a personally experienced adverse reaction (OpenAI 2024) (See Appendix 3 for CoT prompt), we prefixed the normalized text with the corresponding domain-specific tags. The posts that were not flagged remained unmodified. This conditional augmentation enriched our dataset with contextual cues precisely where they were most needed.

Dataset Construction

A PyTorch dataset created through pre-processing was designed to facilitate model training and evaluation. Each so-

cial media post was converted into fixed 256 token length numerical tensors using the Hugging Face AutoTokenizer (Wolf et al. 2020). Subsequently, the model training and inference processes requires standardized input formatting to ensure efficient batch processing. Importantly, we maintained separate training, validation, and test sets consistent with the dataset splits used in the reference paper (Khademi et al. 2024) to facilitate a fair and direct comparison.

Model Architectures

This research used two state-of-the-art transformer-based PLM’s. The first model, `vinai/bertweet-large` (Nguyen, Vu, and Nguyen 2020), is explicitly trained on Twitter data, making it very adept at analyzing informal and chaotic social media text. The second model, `microsoft/deberta-v3-base` (He, Gao, and Chen 2023), is distinguished for its strong representation learning and enhanced contextual embedding proficiency. Both models were augmented by including a binary classification head to accurately predict the presence or absence of adverse vaccination responses.

Pipeline Definitions

To thoroughly examine the effectiveness of our method, we clearly defined and implemented three distinct experimental pipelines:

- **Heuristic Label Pipeline:** In this approach, GPT-4o (OpenAI 2024) with CoT reasoning independently labeled posts as containing adverse reactions or not, providing heuristic labels without further augmentation or PLM classification.
- **Baseline Pipeline:** The PLMs were trained directly on preprocessed textual data without any conditional enhancements.
- **Enhanced Pipeline:** This integrated pipeline used GPT-4o’s (OpenAI 2024) CoT reasoning labels combined with conditional text enhancement, augmenting posts flagged as containing adverse reactions with domain-specific contextual tags, thereby enriching the input representation for subsequent fine-tuning of PLM classifiers.

Cross-Validation and Hyperparameter Optimization

To rigorously evaluate and improve our models, we performed a 5-fold stratified cross-validation (cv), maintaining consistent class distributions in all folds and ensuring reliable evaluation metrics (Pedregosa et al. 2011). We split the original training data in a 80% training subset and a 20% validation subset for stratified cross-validation thereby maintaining class ratios in every fold. We explored various hyperparameters, explicitly testing learning rates ($2e - 5$, $3e - 5$, and $5e - 5$), batch sizes (8, 16, 32), epochs (20 per fold), mixout regularization (with probabilities 0.2 and 0.3), and layer-wise learning rate decay (0.8 and 0.9). Each fold was independently trained using the AdamW optimizer with a cosine decay learning rate schedule and an initial warm-up phase of 500 steps (Loshchilov and Hutter 2016). Techniques such as gradient accumulation and mixed precision

(FP16) training allowed practical training within computational constraints. Early stopping criteria (patience of two epochs) were employed to prevent overfitting. Detailed logs were maintained throughout the process, ensuring transparency and reproducibility.

Model Training

The hyperparameter configuration that yields the highest average F1 score in all folds from the above CV was selected for final training and inferences (See Appendix 2 for optimal hyperparameters). Using this optimal configuration, we retrained the best-performing model in the entire training set and validated its performance in the independent validation set.

Inference

During the inference phase, predictions were generated from the best-performing model trained with the optimal hyperparameter configurations for the test set. We optimized the decision threshold on the validation set to achieve the ideal precision-recall equilibrium (Fawcett 2006). The model logs were converted to predictions and the results were evaluated on the basis of precision, recall, and the F1 score.

Results

The Enhanced Pipeline consistently outperforms the Baseline and Heuristic-only methods, as shown in Table 1. This approach integrates PLM fine-tuning with GPT-4o chain-of-thought labeling and conditional domain-specific augmentation. The model `vinai/bertweet-large` (Nguyen, Vu, and Nguyen 2020) demonstrates the effectiveness of contextual tags and LLM reasoning, achieving a F1 score 0.96 on both validation and test datasets. This result reflects a strong balance between recall (0.97 to 0.96) and consistent precision (0.95). Likewise, `microsoft/deberta-v3-base` (He, Gao, and Chen 2023) benefits from the same augmentation method, achieving F1 scores of 0.95 (validation) and 0.94 (test), thereby confirming that our conditional enhancement is applicable across different architectures. In comparison, the Baseline Pipeline (PLMs alone) achieves a F1 score 0.95, indicating a consistent increase of 0.01 to 0.02 due to augmentation. The GPT-4o (OpenAI 2024) Heuristic Label independently achieves F1 scores of approximately 0.90 on validation and 0.89 on test sets accordingly, indicating that while CoT reasoning provides valuable initial annotations.

Conclusion

This study presents a hybrid pipeline integrating conditional domain-specific augmentation with fine-tuned pre-trained language models (PLMs) and GPT-4o’s chain-of-thought (CoT) reasoning. Evaluated on Reddit vaccination reaction data, our Enhanced Pipeline—built on `vinai/bertweet-large` (Nguyen, Vu, and Nguyen 2020) exceeded the 0.95 benchmark set by Khademi et al. (Khademi et al. 2024) with an unprecedented F1 score of

Table 1: Performance comparison of pipelines on validation (Valid) and test (Test) sets. Precision, recall, and F1 scores for three pipelines—enhanced, baseline, and GPT-4o heuristic—across `vinai/bertweet-large` (Nguyen, Vu, and Nguyen 2020), and `microsoft/deberta-v3-base` (He, Gao, and Chen 2023).

Model	Dataset	Precision	Recall	F ₁
<code>vinai/bertweet-large</code> (Enhanced)	Valid	0.95	0.97	0.96
	Test	0.95	0.96	0.96
<code>microsoft/deberta-v3-base</code> (Enhanced)	Valid	0.94	0.95	0.95
	Test	0.93	0.95	0.94
<code>vinai/bertweet-large</code> (Baseline)	Valid	0.94	0.95	0.95
	Test	0.93	0.96	0.95
<code>microsoft/deberta-v3-base</code> (Baseline)	Valid	0.94	0.96	0.95
	Test	0.95	0.95	0.95
GPT-4o (Heuristic Label)	Valid	0.89	0.91	0.90
	Test	0.88	0.91	0.89

0.96 on both validation and test sets. Especially on vaccine reaction, this improvement shows that adding LLM-generated annotations and domain tags coupled with strong regularization and threshold adjustment greatly improves the categorization of adverse event detection.

Beyond improving performance, our analysis yields three main conclusions: (a) strong training strategies with robust regularizations (mixout, layer-wise learning rate decay, and threshold optimization) are fundamental to reducing overfitting in noisy social media settings, (b) LLM-driven CoT annotations serve as high-quality heuristics for downstream classification, and (c) conditional text augmentation effectively concentrates contextual cues where they matter most.

Despite these developments, our method depends on API-based LLM inference and only utilizes English Reddit data. Future study could investigate (a) scaling to other social media platforms and languages, (b) incorporating multi-modal signals such as: images embedded in posts, emoji usage patterns, and user metadata—to capture richer context, and (c) lightweight on-device adaptation to lower computing costs. In our opinion, these strategies will improve pharmacovigilance systems and increase the relevance of hybrid LLM-PLM models to gather additional public health intelligence. While this use case highlights public perceptions of the Shingles vaccine, it also demonstrates the potential to expand such models for detecting trends in public attitudes across various health topics, particularly vaccinations. These insights can help public health professionals better understand and address the needs of the general public.

References

- Bachmann, M. 2022. RapidFuzz Documentation. <https://maxbachmann.github.io/RapidFuzz/>.
- Betsch, C. 2021. Ergebnisse aus dem COVID-19 Snapshot Monitoring COSMO: Die psychologische Lage 2021. <https://projekte.uni-erfurt.de/cosmo2020/web/summary/54-55/>. [Accessed 28 December 2024].
- Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media.
- Boender, T.; Schneider, P.; Houareau, C.; Wehrli, S.; Purnat, T.; Ishizumi, A.; Wilhelm, E.; Voegeli, C.; Wieler, L.; and Leuker, C. 2023. Establishing Infodemic Management in Germany: A Framework for Social Listening and Integrated Analysis to Report Infodemic Insights at the National Public Health Institute. *JMIR Infodemiology*, 3: e43646.
- Dong, F.; Guo, W.; Liu, J.; Patterson, T. A.; and Hong, H. 2024. BERT-based language model for accurate drug adverse event extraction from social media: implementation, evaluation, and contributions to pharmacovigilance practices. *Front Public Health*, 12: 1392180.
- Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8): 861–874.
- He, P.; Gao, J.; and Chen, W. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv:2111.09543.
- Khademi, S.; Palmer, C.; Dimaguila, G. L.; Javed, M.; and Buttery, J. 2024. Exploring Large Language Models for Detecting Online Vaccine Reactions. In *Health. Innovation. Community: It Starts With Us*, volume 318 of *Studies in Health Technology and Informatics*, 30–35. IOS Press. Ebook.
- Khademi Habibabadi, S.; Delir Haghighi, P.; Burstein, F.; and Buttery, J. 2022. Vaccine Adverse Event Mining of Twitter Conversations: 2-Phase Classification Study. *JMIR Med Inform*, 10(6): e34305.
- Klein, A. Z.; Dasgupta, T.; Flores Amaro, I.; Jana, S.; Khademi, S.; Lopez-Garcia, G.; Onishi, T.; Powell, J.; Raithel, L.; Rajwal, S.; Roller, R.; Sarker, A.; Sinha, M.; Thomas, P.; Tutubalina, E.; Xu, D.; Zweigenbaum, P.; and Gonzalez-Hernandez, G. 2025. Overview of the 10th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HearD) Shared Tasks at ICWSM 2025. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.
- Lazarus, J. V.; White, T. M.; Wyka, K.; Ratzan, S. C.; Rabin, K.; Larson, H. J.; Martinon-Torres, F.; Kuchar, E.; Abdool Karim, S. S.; Giles-Vernick, T.; Müller, S.; Batista, C.; Myburgh, N.; Kampmann, B.; and El-Mohandes, A. 2024. Influence of COVID-19 on trust in routine immunization, health information sources and pandemic preparedness in 23 countries in 2023. *Nature Medicine*, 30(6): 1559–1563.
- Loshchilov, I.; and Hutter, F. 2016. SGDR: Stochastic Gradient Descent with Warm Restarts. In *ICLR Workshop on Optimization for Deep Learning*.
- Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A Pre-trained Language Model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- OpenAI. 2024. GPT-4o Technical Report. Technical report, OpenAI.
- Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.
- Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; and Rush, A. M. 2020. HuggingFace's Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Appendix

Listing 1: Example of Baseline vs. Enhanced Text Processing

```
1 Raw Text: "1st shingles shot over 72
  hours ago and still fatigued: Second
  shot sucks too..."
2 Baseline Text: "1 st shingles shot over
  72 hours ago and still fatigued :
  second shot sucks too ."
3 Enhanced Text: "<KNOW_FATIGUE> <
  KNOW_SHINGLES> <KNOW_SHOT> </s> 1 st
  shingles shot over 72 hours ago and
  still fatigued : second shot sucks
  too ."
```

Listing 2: Optimal hyperparameter configuration for each model

```
1 vinai/bertweet-large:
2 [
3   learning_rate: 2e-5, epoch: 20,
4     batch_size: 16, mixout_p: 0.3,
5     llrd_decay: 0.9, early_stopping: 2
6 ]
7 microsoft/deberta-v3-base:
8 [
9   learning_rate: 3e-5, epoch: 20,
10     batch_size: 16, mixout_p: 0.3,
11     llrd_decay: 0.9, early_stopping: 2
12 ]
```

Listing 3: Chain-of-Thought Prompt Used for GPT-4o Screening

```
1 You are an expert in detecting adverse
  reactions related to vaccinations
2 from social media posts. Your task is to
  analyze posts that describe
3 personal experiences with vaccines,
  focusing on adverse events or vaccine
4 failures. These posts may refer to
  vaccines for shingles,
```

5 covid-19/coronavirus, HPV, influenza, or
6 other immunizations. Use a
7 detailed chain-of-thought reasoning
8 process to answer the following
9 questions step-by-step, considering both
10 the specific keywords provided
11 and their synonyms or related phrases.
12 Q1: Does the text indicate that the post
13 is discussing a vaccine?
14 Consider general terms such as '
15 vaccination', 'immunization', 'jab',
16 'shot', 'booster', as well as specific
17 vaccine names such as 'shingrix',
18 'zostavax', 'covid vaccine', 'pfizer', '
19 moderna', 'johnson & johnson',
20 'astrazeneca', 'flu shot', 'hvp vaccine
21 ', and 'chicken pox' (for
22 shingles), along with their synonyms or
23 related phrases.
24 Q2: Does the text describe any symptoms
25 or reactions? Look for mentions
26 of symptoms such as 'injection site pain
27 ', 'pain', 'body ache',
28 'migraine', 'blister', 'allergic
29 reaction', 'fever', 'headache', 'rash
30 ',
31 'skin symptoms', 'eye symptoms', 'chills
32 ', 'nausea', 'dizziness',
33 'swelling', 'redness', 'itching', '
34 urticaria', 'anaphylaxis', 'bumps',
35 'herpes zoster', or 'shingles outbreak',
36 and any synonyms or similar
37 expressions.
38 Q3: Are there clear indications that
39 these symptoms occurred after
40 receiving the vaccine? Search for
41 temporal cues such as 'after', 'post
42 ',
43 or other phrases linking the onset of
44 symptoms to the vaccination event.
45 Q4: Does the text mention any effects,
46 outcomes, changes, or responses
47 resulting from the vaccine? Look for
48 references to seeking medical
49 attention (e.g., doctor, hospital,
50 emergency, appointment), disruptions
51 to
52 daily life (e.g., overslept, missed work
53 , absenteeism, downtime,
54 recovery), use of medication to manage
55 side effects, sleep
56 disturbances, or descriptions of
57 discomfort, stress, or being unwell.
58 Even if the tone is optimistic, if there
59 is any indication that the
60 symptoms affected daily functioning or
61 required intervention, answer
62 'Yes'.
63 Q5: Does the text mention any specific
64 vaccines, vaccine brands, or
65 related terms, as well as any associated
66 effects, outcomes, changes, or
67 responses? Consider all general and
68 specific vaccine-related terms

38 mentioned above, and determine whether
39 the post is describing a reaction
40 from personal experience or reporting
41 someone else's reaction.
42 Based on your answers to Q1 through Q5,
43 provide a brief explanation for
44 each question. Then, if the text clearly
45 describes a personally
46 experienced adverse reaction to a
47 vaccine, output a final label of '1'.
48 Otherwise, output '0'. Please ensure you
49 label the text based on your
50 complete analysis of all the chain-of-
51 thought answers.
52 Text: "<your input text here>"
53 Chain-of-thought reasoning and final
54 label: