Enhancing Vaccine Reaction Detection from Social Media Using Optimized Transformer Fine-Tuning

Abdusalam F. A. Nwesri[1] and Mai M. Elbaabaa[1] and Nabila A. S. Shinbir[2]

[1] University of Tripoli, Tripoli - Libya
[2] College of Science and Technology, Tripoli - Libya
a.nwesri@uot.edu.ly
DOI: https://doi.org/10.5281/zenodo.18111761

Abstract. This paper describes the system developed by the University of Tripoli (UoT) team for Task 6 of the 10th Social Media Mining for Health (#SMM4H) Shared Tasks, which focused on detecting personally experienced vaccine reactions in social media posts. We fine-tuned the CardiffNLP Twitter-RoBERTa-Large model using optimized training settings and data preprocessing strategies. Our best submission achieved an F1-score of 0.945 on the test set, outperforming the average system and nearing the benchmark score of 0.946. We describe our training pipeline, evaluation metrics, and results, and compare our system with both large language models and benchmark transformer-based models.

Keywords: Vaccine Reaction Detection, LLMs, Social Media Mining for Health

1 Introduction

Social media platforms such as Twitter, Reddit, and Facebook have become valuable sources for real-time public health monitoring. These platforms allow individuals to share personal experiences with medical treatments, including vaccines, making them a rich source of health-related data [1,2]. Mining this data can aid in understanding vaccine uptake, hesitancy, and adverse reactions at a large scale and in near real-time, which is critical for public health surveillance and rapid response.

The Social Media Mining for Health (#SMM4H) workshop series has established itself as a leading venue for research at the intersection of natural language processing (NLP), machine learning (ML), and public health. Its 10th edition introduced the Health Real-World Data (HeaRD) track, encouraging research on extracting meaningful health signals from web-based textual data [3]. Within this track, Task 6 focused specifically on detecting personally experienced reactions to the shingles vaccine, a problem with direct implications for vaccine safety monitoring and post-marketing surveillance.

Participants were provided with labeled datasets for training and validation, alongside an unlabeled test set. The core challenge was to build models that could robustly identify vaccine adverse event mentions (VAEMs) in informal and noisy social media

posts. Such classification tasks are inherently difficult due to the variability in user language, spelling errors, sarcasm, and lack of context in short texts like tweets.

In this paper, we present the system developed by the UoT team for Task 6. Our approach leverages the CardiffNLP Twitter-RoBERTa-Large model [4], fine-tuned using optimized hyperparameters and targeted preprocessing strategies to enhance performance. Our system achieved an F1-score of 0.945 on the official test set, nearly matching the benchmark score of 0.946. We detail our training pipeline, evaluate the performance of several transformer-based models, and discuss our findings in comparison with existing approaches from prior SMM4H tasks.

2 Related Work

While early research on monitoring vaccine conversations on social media relied on manual online Browse, the field has progressed significantly with the integration of ML and NLP. This shift has allowed studies to concentrate on automatically detecting public sentiment toward vaccines [5].

For instance, Khademi Habibabadi *et al*. [6] pioneered gaining early insights into vaccine safety by extracting adverse event mentions from Twitter conversations. Their approach, which combined topic modeling and classification, successfully identified 8,992 VAEMs from over 811,000 vaccine-related tweets, achieving an F1 score of 0.91 with their ML classifiers.

Similarly, Portelli *et al.* [7] developed a sophisticated tool for assessing public opinion on COVID-19 vaccines using Twitter data. This system leveraged NLP models for sentiment analysis, fine-tuning a RoBERTa model to achieve a 72.6% macro-averaged recall [8]. Their participation in the SMM4H Shared Task further showcased their expertise, where they achieved top results with a deep-learning architecture combining SpanBERT [9] and Conditional Random Fields (CRFs) [10].

The detection of VAEMs through social media text has consistently been a central research theme within the SMM4H forum, as evidenced by the comprehensive overviews of the SMM4H 2022, 2023, and 2024 shared tasks [11,12,13,14], which detail past system performances, challenges, and ensemble modeling strategies.

In this paper, we aim to improve performance by employing focused fine-tuning and robust preprocessing techniques.

3 The Task

Shared task 6 on the 10th Social Media Mining for Health (#SMM4H) focuses on binary classification to identify Reddit posts that specifically mention personal adverse reactions to shingles vaccines. This involves distinguishing such posts from general vaccine-related discussions. Participants will receive training and validation datasets to build their models, which will then be assessed on a separate test set. The effectiveness of each system will be measured using the F1-score.

3.1 Data set

We used the official dataset provided by the organizers, which included 2521 training posts and 786 validation posts. Each post was labeled 0 or 1, with 0 meaning that the post has no Vaccine Event Adverse Mentions (VAEM) and 1 meaning that the post has VAEM. The organizer later released an unseen test dataset containing 8113 unclassified text posts that should be classified by the participants' classification techniques and uploaded to the task official website for automatic evaluation. Table 1 shows the details of the training dataset.

Table 1. Details of the Task Dataset

Dataset	VAEMs (1)	No VAEMs (0)	Total
Training	1149	1372	2521
Validation	366	420	786
Test	N/A	N/A	8113

3.2 Models used

Several models have been fine-tuned to test their effectiveness in classifying the training posts. Models have been carefully chosen based on their relevance and performance on similar tasks in the literature namely: The cardiffnlp/twitter-roberta-base and its large version cardiffnlp/twitter-roberta-large-2022-154m, the google-bert/bert-base-uncased model and its large version google-bert/bert-large-uncased, the LuizNeves/DeBERTa-v3-large-vaccine model, and the abigailp/vaccinated models.

The twitter-roberta-base model is trained on nearly 58M tweets on top of the original RoBERTa-base checkpoint [15]. The model performs well on several tasks. The RoBERTa-large model [4] is the latest updated version of Roberta-base. It is trained on 154M tweets filtered from 220M tweets covering the period between January 2018 and December 2022 incorporating more recent vocabulary and topics, including COVID-19, vaccines, and political trends. The model is considered state-of-the-art when fine-tuned for sentiment analysis on Twitter data [16]. It was developed by the Cardiff NLP group and is part of their suite of models designed to handle the linguistic nuances and informal language often found in social media content.

The LuizNeves/DeBERTa-v3-large-vaccine model is a fine-tuned transformer model based on DeBERTa-v3-large [17], specifically trained for vaccine-related sentiment classification and stance detection. It was developed by Luiz Neves, a researcher focused on biomedical NLP, particularly in the context of public health and social media analysis. The DeBERTa-v3-large model was superior to other models in several tasks [18].

The google-bert/bert-base-uncased and its large version have also been considered since the BERT model has been used in language understanding tasks and was proven to be superior to its predecessor models [19].

Another model which is trained on vaccine data is the abigailp/vaccinated[1]. This model is a fine-tuned version of bert-base-uncased on an unknown dataset. It achieves an F1 score of 0.90. Since the model is trained on vaccine data it is included in our evaluation.

Models were iterated on different fine-tuning strategies, first freezing all layers of the transformer model and gradually unfreezing the final two layers and pooler to balance generalization with task-specific learning. We tuned learning rates and warm-up steps to reduce instability often observed in small or imbalanced datasets. In addition to using weighted loss to counter class imbalance, we relied on epoch-level evaluation and early stopping to maintain a balance between learning and overfitting. Each training run was evaluated using multiple metrics, with the best model selected based on lowest validation loss and highest validation F1. These controlled procedures enabled us to achieve high precision and recall in our best submission.

3.3 Evaluation Measures

The main metric used to evaluate teams' submissions in the task was the F1-score which is calculated using the precision and recall measures as:

$$F1 = \frac{2\times(Precision \times Recall)}{Precision+Recall} \quad (1)$$

Where

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives} \quad (2)$$

and

$$Recall = \frac{True\ Positives}{True\ Positives+False\ Negaives} \quad (3)$$

4 Experiments

Our experimentation was driven by a clear goal: to improve the benchmark results by tailoring training settings specifically to social media data

Several well-known models have been fine-tuned using the labeled training and the validation datasets provided as part of the task.

All models have been fine-tuned using HuggingFace's Trainer API. while training a model, we froze most layers while unfroze the final two encoder layers and pooler. Training hyperparameters are set after careful fine-tuning. We reported the best model based on the F1-socre performance. We ensured consistent label encoding (0: No Reaction, 1: Reaction) and tokenized the text using the model's tokenizer with a maximum sequence length of 128.

[1] https://huggingface.co/abigailp/vaccinated

Table 2 presents the results of fine-tuning the selected models on identifying vaccine adverse mentions in tweets. The models were evaluated on both the training and test datasets.

Across all metrics F1-score (F1), precision (P), and recall (R). twitter-roberta-large achieved the highest performance, with an F1-score of 0.957 on the training set and 0.945 on the test set. Its strong recall on the test set (0.983) indicates exceptional sensitivity in detecting vaccine reaction mentions.

The DeBERTa-v3-large-vaccine model ranked second, achieving F1-scores of 0.935 (training) and 0.926 (test). Although slightly behind the top model, it maintained balanced precision and recall, reflecting strong generalization.

The bert-large-uncased model also performed well, achieving 0.941 on the training set and 0.923 on the test set. However, its performance was consistently lower than the two domain-adapted models, showing the limitations of models not pre-trained on social media text.

Smaller models; twitter-roberta-base, bert-base-uncased, and vaccinated; showed substantially lower F1-scores, with test-set results ranging from 0.850 to 0.868. These models exhibited reduced precision and recall, consistent with their smaller capacity and less specialized pre-training.

These results underline the effectiveness of using large, transformer-based models fine-tuned on relevant domains for health-related social media mining. Meanwhile, smaller base models, although computationally efficient, show limitations in both accuracy and generalization.

Table 2. Results of fine-tuning models on the training and test datasets.

	Training Dataset Results			Testset Dataset Results		
Model	F1	P	R	F1	P	R
twitter-roberta-large	0.957	0.944	0.970	0.945	0.911	0.983
DeBERTa-v3-large-vaccine	0.935	0.899	0.974	0.926	0.892	0.962
bert-large-uncased	0.941	0.944	0.938	0.923	0.918	0.928
twitter-roberta-base	0.865	0.839	0.893	0.850	0.802	0.904
bert-base-uncased	0.863	0.843	0.884	0.864	0.818	0.914
vaccinated	0.866	0.854	0.878	0.868	0.846	0.890

4.1 Official Results

After extensive hyperparameter tuning of the cardiffnlp/twitter-roberta-large-2022-154m model, two official runs were submitted for evaluation (Table 3). In the first submission, the learning rate was set to 2e-4, the training and evaluation batch sizes were both set to 8, and the model was trained for 10 epochs. This configuration achieved an F1-score of 0.910, with precision of 0.916 and recall of 0.904, providing a strong baseline but indicating that recall could be further improved.

In the second submission, we refined the hyperparameters by reducing the learning rate to 2e-6, maintaining batch sizes at 8, introducing warm-up decay (0.05) and a warm-up ratio of 1000, and increasing the number of epochs to 16. These adjustments produced a substantial improvement, yielding an F1-score of 0.945, preserving precision at 0.916, and significantly increasing recall to 0.976. This marked gain demonstrates the effectiveness of the optimized training strategy. Both runs were trained solely on the original dataset provided by the task organizers.

Table 3. Official submitted runs: results on the unseen dataset

	F1	P	R
Roberta-large Run 1	0.910	0.916	0.904
Roberta-lage Run 2	**0.945**	**0.916**	**0.976**

When compared with all participating team submissions (Table 4), our best system outperformed both the mean F1-score (0.938) and the median F1-score (0.944), while also exceeding the median recall (0.972). Notably, although the benchmark system achieved a slightly higher F1-score of 0.946 using the same underlying model, our system delivered higher recall, indicating superior sensitivity in identifying vaccine reaction mentions. Overall, these results underscore the robustness of our domain-specific optimization approach and its competitiveness within the evaluation setting.

Table 4. Average teams' performance

Statistic	F1	P	R
Mean	0.938	0.916	0.961
Median	0.944	0.922	0.972

Our best submission (F1 = 0.945) exceeded both the mean (0.938) and median (0.944) F1-scores reported across all team submissions to Task 6. It also achieved notably high recall (0.976), indicating strong sensitivity in detecting vaccine reaction mentions. While the benchmark system achieved a slightly higher F1-score of 0.946 using the same base model, our system's recall surpassed the benchmark's reported average, suggesting superior performance in identifying relevant examples. These results highlight the robustness of our optimization strategy in a competitive evaluation setting.

5 Discussion

The experimental results offer several critical insights into effectively detecting personally experienced VAEMs from informal social media data. Our findings confirm the significance of domain-specific model pre-training and underscore the value of precision-focused hyperparameter tuning.

The CardiffNLP Twitter-RoBERTa-Large model consistently achieved the highest F1-scores across all experiments, demonstrating its suitability for this classification task. This superior performance, compared to the general-purpose BERT and even the specialized DeBERTa models, is primarily attributed to its foundational training on a massive corpus of Twitter data (154 million tweets). Social media language is characterized by noise, specialized jargon, abbreviations, and informal phrasing (e.g., "shingles shot," "arm hurts," "felt rough"). A model pre-trained on this linguistic domain possesses a crucial advantage in encoding these nuances, leading to better feature extraction and stronger generalization ability on the test set. This confirms the necessity of using social media-native transformers when targeting public health surveillance tasks on platforms like Reddit and Twitter.

The competitive success of our final system (Run 2, F1: 0.945) was achieved through a dedicated optimization process. The significant performance gap between Run 1 (F1: 0.910) and Run 2 (F1: 0.945) highlights that model selection alone is insufficient; the fine-tuning recipe is equally critical. By adjusting the learning rate to a lower value (2e-6), introducing warm-up decay, and increasing epochs, we facilitated a more stable convergence, enabling the model to learn subtle classification boundaries without overfitting.

When benchmarked against all participating systems (Table 4), our submission exceeded both the mean (0.938) and median (0.944) F1-scores. Most notably, our system achieved a high Recall of 0.976. In the context of pharmacovigilance and real-time public health monitoring, Recall (Sensitivity) is often the paramount metric, as minimizing False Negatives (missed adverse events) is necessary to ensure the timely detection of emerging safety signals. Our system's high sensitivity makes it a reliable tool for initial filtering and alerting in a real-world surveillance pipeline.

6 Conclusion

This study presented the UoT system developed for Task 6 of the #SMM4H-HeaRD 2025 Shared Task, which involved binary classification of social media posts mentioning personal shingles vaccine reactions. The system leveraged the CardiffNLP Twitter-RoBERTa-Large model and a carefully optimized fine-tuning pipeline to produce strong and competitive results. Our main contributions include:

- Demonstrating the effectiveness of large, social media–specific transformer models for health-related text classification.
- Achieving a competitive F1-score of 0.945, surpassing both the mean and median scores of all task participants.
- Attaining a high recall (0.976), ensuring strong sensitivity for public health monitoring applications where missed cases carry high risk.

Future work will focus on improving precision to reduce false positives, thereby lowering the manual verification burden on downstream reviewers. We also plan to explore automated annotation techniques for LLM-generated text, aiming to develop a fully autonomous and scalable augmentation pipeline.

References

1. Emilie Karafillakis and et al. Monitoring covid-19 vaccine conversations on twit- ter: A comparison of automated sentiment analysis and manual annotation. In *Vaccine*, 2021.
2. Amin Khademi and et al. Extracting adverse events from covid-19 vaccine con- versations on twitter. In *Proceedings of the International Conference on Social Media Mining for Health*, 2022.
3. Ari Z. Klein, Tirthankar Dasgupta, Ivan Flores Amaro, Sudeshna Jana, Sedigh Khademi, Guillermo Lopez-Garcia, Takeshi Onishi, Jeanne Powell, Lisa Raithel, Swati Rajwal, Roland Roller, Abeed Sarker, Manjira Sinha, Philippe Thomas, Elena Tutubalina, Dongfang Xu, Pierre Zweigenbaum, and Graciela Gonzalez- Hernandez. Overview of the 10th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ICWSM 2025. In *Work- shop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press, 2025.
4. Daniel Loureiro, Kiamehr Rezaee, Talayeh Riahi, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. Tweet insights: A visualization platform to extract temporal insights from twitter. *arXiv preprint arXiv:2308.02142*, 2023.
5. Emilie Karafillakis, Sam Martin, Clarissa Simas, Kate Olsson, Judit Takacs, Sara Dada, and Heidi Jane Larson. Methods for social media monitoring related to vac- cination: Systematic scoping review. *JMIR Public Health Surveill*, 7(2):e17149, Feb 2021.
6. Sedigheh Khademi Habibabadi, Pari Delir Haghighi, Frada Burstein, and Jim Buttery. Vaccine adverse event mining of twitter conversations: 2-phase clas- sification study. *JMIR Med Inform*, 10(6):e34305, Jun 2022.
7. Beatrice Portelli, Simone Scaboro, Roberto Tonino, Emmanuele Chersoni, En- rico Santus, and Giuseppe Serra. Monitoring user opinions and side effects on covid-19 vaccines in the twittersphere: Infodemiology study of tweets. *Journal of medical Internet research*, 24(5):e35115, May 2022.
8. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
9. BeatricePortelli,DanielePassab`ı,EdoardoLenzi,GiuseppeSerra,EnricoSantus, and Emmanuele Chersoni. Improving adverse drug event extraction with spanbert on different text typologies, 2021.
10. Charles Sutton, Khashayar Rohanimanesh, and Andrew McCallum. Dynamic conditional random fields: factorized probabilistic models for labeling and seg- menting sequence data. In *Proceedings of the Twenty-First International Con-*
11. Davy Weissenbacher, Juan Banda, Vera Davydova, Darryl Estrada Zavala, Luis Gasco Sa´nchez, Yao Ge, Yuting Guo, Ari Klein, Martin Krallinger, Mathias Led- din, Arjun Magge, Raul Rodriguez-Esteban, Abeed Sarker, Lucia Schmidt, Elena Tutubalina, and Graciela Gonzalez-Hernandez. Overview of the seventh social media mining for health applications (#SMM4H) shared tasks at COLING 2022. In Graciela Gonzalez-Hernandez and Davy Weissenbacher, editors, *Proceedings of the Seventh Workshop on Social Media Mining for Health Applications, Work- shop & Shared Task*, pages 221–241, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics.
12. Ari Z Klein, Juan M Banda, Yuting Guo, Ana Lucia Schmidt, Dongfang Xu, Ivan Flores Amaro, Raul Rodriguez-Esteban, Abeed Sarker, and Graciela Gonzalez-Hernandez. Overview of the 8th social media mining for health appli- cations (smm4h) shared tasks at the amia 2023 annual symposium. *Journal of the American Medical Informatics Association*, 31:991—-996, 2024.

13. Dongfang Xu, Guillermo Garcia, Lisa Raithel, Philippe Thomas, Roland Roller, Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, Pierre Zweigenbaum, Karen O'Connor, Sai Samineni, Sophia Hernandez, Yao Ge, Swati Rajwal, Sudeshna Das, Abeed Sarker, Ari Klein, Ana Schmidt, Vishakha Sharma, Raul Rodriguez- Esteban, Juan Banda, Ivan Amaro, Davy Weissenbacher, and Graciela Gonzalez- Hernandez. Overview of the 9th social media mining for health applications (#SMM4H) shared tasks at ACL 2024 – large language models and generaliz- ability for social media NLP. In Dongfang Xu and Graciela Gonzalez-Hernandez, editors, *Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 183–195, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
14. AzmineToushikWasiandSheikhRahman.DILABat#SMM4H2024:RoBERTa ensemble for identifying children's medical disorders in English tweets. In Dong- fang Xu and Graciela Gonzalez-Hernandez, editors, *Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 10–12, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
15. Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet clas- sification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November 2020. Association for Computational Linguistics.
16. Sedigh Khademi, Christopher Palmer, Gerardo Luis Dimaguila, Muhammad Javed, and Jim Buttery. Exploring Large Language Models for Detecting Online Vaccine Reactions. In *Proceedings of HIC 2024 - Health. Innovation. Commu- nity: It Starts With Us*, volume 318, pages 30–35, 2024.
17. Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543, 2021.
18. Mae ̈l Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and Andre ́ Freitas. SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In Atul Kr. Ojha, A. Seza Dog ̆ruo ̈z, Giovanni Da San Mar- tino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceed- ings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada, July 2023. Association for Computational Linguistics.
19. JacobDevlin,Ming-WeiChang,KentonLee,andKristinaToutanova.BERT:pre- training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

تحسين اكتشاف ردود الفعل تجاه اللقاحات من خلال وسائل التواصل الاجتماعي باستخدام الضبط الدقيق الأمثل للمحولات

عبدالسلام النويصري[1] ، مي البعباع[1]، نبيلة المبروك شنبر[2]

[1] جامعة طرابلس، طرابلس - ليبيا
[2] كلية العلوم والتقنية ، طرابلس - ليبيا
a.nwesri@uot.edu.ly
DOI: https://doi.org/10.5281/zenodo.18111761

ملخص. تصف هذه الورقة البحثية النظام الذي طوره فريق جامعة طرابلس للمهمة السادسة من المهام المشتركة العاشرة لتعدين وسائل التواصل الاجتماعي من أجل الصحة (#SMM4H)، والتي ركزت على رصد ردود الفعل الشخصية تجاه اللقاح في منشورات وسائل التواصل الاجتماعي. قمنا بتحسين نموذج CardiffNLP Twitter-RoBERTa-Large باستخدام إعدادات تدريب مُحسّنة واستراتيجيات معالجة مسبقة للبيانات. حقق أفضل نموذج قدمناه 0.945 بمقياس F1 على بيانات الاختبار، متفوقًا على المتوسط العام لاداء جميع الأنظمة المشاركة ومقتربًا من درجة المَعيار المرجعي البالغة 0.946. نوضح مسار التدريب ومقاييس التقييم والنتائج، ونقارن نظامنا بنماذج اللغات الكبيرة والنماذج القائمة على مُحوّلات المعايير المرجعية.

الكلمات المفتاحية: الكشف عن ردود الفعل التحسسية للقاحات، النماذج اللغوية الضخمة، استخراج البيانات من وسائل التواصل الاجتماعي لأغراض صحية.