Gharyan University Journal of Engineering Science (GUJES)
ISSN (3105-4560)
Website: http://journals.gu.edu.ly
email: gujes@gu.edu.ly

Enhancing Vaccine Adverse Event Detection on Social Media through LLM-Driven Synthetic Data Augmentation

Abdusalam F. Ahmad Nwesri[1], Mai M. Elbaabaa[2], Nabila Almabrok S Shinbir[3]. Hassan Ali Ebrahem[4] , and Marwa N. Solla[5]

[1] University of Tripoli, Tripoli – Libya, a.nwesri@uot.edu.ly
[2] University of Tripoli, Tripoli – Libya, m.elbaabaa@uot.edu.ly
[3] College of Science and Technology, Tripoli – Libya, shinbir@tcst.edu.ly
[4] University of Tripoli, Tripoli – Libya, H.Ebrahem@uot.edu.ly
[5] University of Tripoli, Tripoli – Libya, M.Solla@uot.edu.ly

*Corresponding author: a.nwesri@uot.edu.ly

Received: 03 March 2026	Accepted: 07 March 2026	Published: 10 March 2026

Abstract

This paper evaluates the performance impact of synthetic data augmentation on the detection of personally experienced vaccine reactions in social media posts. Our study uses the UoT team's submission for Task 6 of the 10th Social Media Mining for Health (#SMM4H) Shared Tasks as a foundation. By establishing a baseline through the fine-tuning six Large Language Models (LLMs), we analyze how augmenting the training set with synthetically generated examples influences classification metrics. Our experiment shows that synthetic augmentation leads to substantial performance improvements across all models with an additional benefit to small models.

Keywords: LLMs, Vaccine Adverse Event Reporting, Synthetic Augmentation

1 Introduction

Social media platforms like X (formerly Twitter), Reddit, and Facebook have transformed into vital "digital listening posts" 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]. Unlike traditional clinical trials or passive surveillance systems like the Vaccine Adverse Event Reporting System (VAERS), which often suffer from reporting lags and under-reporting of mild symptoms, social media provides an unfiltered, immediate view of patient experiences [3]. These platforms allow individuals to describe vaccine reactions in their own words, capturing "real-world" data that includes nuanced, non-technical descriptions of side effects often missed in formal clinical settings [4, 5]. Mining this data is critical for infoveillance—the early detection of emerging safety signals and the monitoring of public sentiment toward vaccination.

However, utilizing social media for health surveillance presents a significant "data scarcity" challenge. While general discussions about vaccines are abundant, specific reports of personally experienced adverse reactions are relatively rare, leading to a severe class imbalance [6]. In many health-related datasets, the "positive" class (actual reports of side effects) constitutes only a small fraction of the total data, creating a "long-tail" distribution where specific rare symptoms are underrepresented [5]. Standard machine learning models, including high-performing transformers like RoBERTa, often struggle to generalize from these limited samples, frequently biasing toward the majority "negative" class and resulting in high false-negative rates [6].

To bridge this gap, synthetic data augmentation has emerged as a transformative solution. Recent studies have demonstrated that using LLMs to generate synthetic training examples can effectively expand the diversity of rare posts without the prohibitive cost of manual labeling [5,6]. By leveraging the contextual understanding of models like GPT-4 or DeepSeek, researchers can create realistic, semantically varied reports that mimic the informal language of social media while covering a broader range of reaction descriptions [6]. For instance, augmentation has been shown to improve the detection of rare symptoms by providing the "linguistic novelty" required for models to learn more robust decision boundaries [5].

This paper evaluates the impact of such augmentation on the #SMM4H Task 6 dataset [7]. We establish a baseline using fine-tuned six language models and demonstrate how the strategic inclusion of synthetic data expands the representation of rare vaccine reactions, ultimately enhancing the systems’ sensitivity.

1 Related Work

The detection of vaccine adverse events (VAEs) from social media has emerged as a critical area within computational pharmacovigilance, offering a real-time complement to traditional surveillance systems that often experience significant reporting delays [3]. Despite the promise of Large Language Models (LLMs) in this domain, their performance is frequently constrained by the long-tail distribution of symptom data, where rare but clinically significant adverse events are sparsely represented in training corpora [8]. This imbalance limits model sensitivity to infrequent events and motivates the adoption of synthetic data augmentation strategies to enhance representation during fine-tuning. Chen and Zhang [9] proposed ENDA, a synthetic augmentation technique that generates additional training samples through controlled manipulation of numeric attributes while preserving label validity. Their results demonstrated notable improvements in BERT-based classification, particularly in early COVID-19 discourse analysis, showing that synthetic expansion can improve trend detection and temporal analysis under limited data conditions. Addressing the long-tail problem more directly, Kim and Nakashole [8] implemented a targeted augmentation framework using an autoregressive model to generate standardized symptom terminology and diverse patient-report variations, thereby increasing training coverage for rare adverse events and outperforming strong baselines in symptom detection and normalization. Complementing augmentation-focused approaches, Khademi Habibabadi et al. [3] emphasized data quality by extracting nearly 9,000 high-precision vaccine adverse event mentions from over 800,000 social media posts, achieving an F1-score of 0.91 without reliance on exhaustive keyword lists; their methodology provides a reliable "gold standard" dataset that supports downstream fine-tuning and augmentation efforts. Meanwhile, Scaboro et al. [10] highlighted the importance of linguistic nuance in adverse drug event extraction, particularly the handling of negation and speculation, and demonstrated through BERT ensemble benchmarking that models must be trained on data reflecting real-world expression variability to achieve deployment-level robustness. Finally, Feng et al. [11], in a large-scale review of 100 studies on health misinformation detection, identified class imbalance as a persistent challenge and underscored the role of embedding-based deep learning models, ensemble strategies, domain-specific fine-tuning (e.g., BioBERT, RoBERTa), and synthetic augmentation in achieving high performance (F1-scores above 0.90) under complex and evolving health communication scenarios. Collectively, this body of work establishes synthetic augmentation and targeted fine-tuning as essential strategies for improving LLM-based detection of vaccine-related adverse events in noisy, imbalanced social media environments.

2 The Baseline Experiment

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 [7]. 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.

2.1 The Dataset

The official dataset provided by the organizers 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 test set which was later released contains 8113 unclassified text posts that should be classified by the participants' classification techniques. Table 1 shows the details of the training dataset.

Table 1. Details of the Task Dataset

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

2.2 Models used

Several models have been fine-tuned by the UoT team to test their effectiveness in classifying the training posts [12]. The twitter-roberta-base model is trained on nearly 58M tweets on top of the original RoBERTa-base checkpoint [13]. The model performs well on several tasks. The RoBERTa-large model [14] 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 [15]. 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 [16], 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 [17].

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 [18].

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. Learning rates and warm-up steps were tuned to reduce instability often observed in small or imbalanced datasets. In addition to using weighted

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

loss to counter class imbalance, epoch-level evaluation and early stopping were employed 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.

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. The Baseline: fine-tuning models on the original unbalanced training and test dataset.

	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

3 Data Augmentation

To assess the influence of dataset size and the intrinsic imbalance within the training data, we incorporated additional posts generated by large language models, specifically ChatGPT [2] and Deepseek[3].

A series of prompts (presented in Table 3) were formulated to guide the generation of relevant content, which was subsequently manually annotated as either "related" or "non-related" based on the prompt's context.

Four participants were asked to use these prompts to retrieve related results from both ChatGPT and Deepseek AI engines. The collected statements were manually reviewed to assure their correctness.

This augmentation led to the inclusion of 451 automatically generated and manually annotated posts, yielding an augmented training dataset comprising 3,380 posts, with a balanced distribution of 1,486 posts for each class (label 0 and label 1), and a validation dataset comprising of 467 posts containing mentions of adverse vaccine events and 467 posts with no mentions of adverse vaccine events.

Table 3. Prompts used to collect vaccine related posts

Prompt	Label assigned
give 100 more detailed stories about different experiences people have reported after receiving the shingles vaccine. must include vaccine general name (shingles) must exceed 100 words	1
give me 100 examples of experiencing shingles vaccine reactions	1
give 100 detailed stories about different experiences people have reported after receiving vaccines other than shingles. must include vaccine name and must exceed 100 words	0

All models have been re-fine-tuned using the new augmented datasets.

Table 4 shows that fine-tuning models on the augmented dataset led to substantial performance improvements compared with the baseline results, which used only the original data.

The results indicate that synthetic augmentation and class balancing have a differential impact depending on model capacity. While the highest-capacity model (twitter-roberta-large) exhibits virtually no change in test performance, suggesting saturation and inherent robustness

[2] https://chatgpt.com
[3] https://deepseek.com

to class imbalance, most other models benefit from the augmented balanced dataset. In particular, DeBERTa-v3-large-vaccine shows a clear improvement in F1 driven primarily by increased precision, indicating better class discrimination after balancing. The most substantial gains are observed in base-sized models, where F1 improvements of up to 4.4% are achieved, accompanied by consistent increases in both precision and recall. These findings suggest that synthetic augmentation effectively mitigates imbalance-induced bias, enhances minority class sensitivity, and improves generalization stability, especially for lower-capacity architectures that are more sensitive to skewed data distributions.

Table 4. Results of fine-tuning models on the augmented training set and the testset.

	Training Dataset Results			Testset Dataset Results		
Model	F1	P	R	F1	P	R
twitter-roberta-large	0.960	0.937	0.983	0.945	0.910	0.983
DeBERTa-v3-large-vaccine	0.935	0.899	0.974	0.937	0.923	0.952
bert-large-uncased	0.947	0.927	0.967	0.923	0.899	0.948
twitter-roberta-base	0.904	0.883	0.926	0.894	0.853	0.938
bert-base-uncased	0.896	0.887	0.904	0.897	0.871	0.924
vaccinated	0.896	0.883	0.910	0.879	0.848	0.911

4 Discussion

The experimental findings provide several important insights into the design of robust systems for detecting personally experienced Vaccine Adverse Event Mentions (VAEMs) in social media data. First, the results confirm that transformer-based architectures, when carefully fine-tuned, are highly effective for this task. Large, domain-adapted models such as twitter-roberta-large and DeBERTa-v3-large-vaccine demonstrated strong baseline performance, highlighting the importance of pre-training on social media or vaccine-related corpora. Their ability to capture informal language patterns, lexical variability, and contextual nuance is central to accurate VAEM detection.

Second, the study clearly illustrates the impact of class imbalance and dataset size on model behavior. Although the original dataset imbalance was moderate, smaller and base-sized models showed clear limitations in precision–recall trade-offs, indicating sensitivity to skewed class distributions. The introduction of 451 synthetically generated and manually validated posts improved both dataset balance and linguistic diversity. As a result, most models exhibited consistent gains in F1-score, precision, and recall after re-training on the augmented corpus. The performance improvement was especially pronounced in base models, where gains of up to 4.4% on the test set were observed. This suggests that lower-capacity architectures benefit more substantially from augmentation, as they rely more heavily on sufficient class representation to form stable decision boundaries.

In contrast, the largest model (twitter-roberta-large) showed minimal change in test performance following augmentation. This behavior suggests a form of performance saturation, where the model's representational capacity and domain-specific pre-training already enable robust handling of moderate imbalance. Therefore, synthetic augmentation appears to function primarily as a compensatory mechanism for models with limited capacity or less specialized pre-training rather than universally boosting all architectures.

Importantly, the results validate LLM-based synthetic augmentation as a practical and scalable solution for low-resource health NLP tasks. By leveraging generative models such as ChatGPT and Deepseek to produce realistic, semantically coherent vaccine-related narratives, we reduced dependence on costly manual data collection while preserving annotation quality through manual verification. The balanced augmented dataset enhanced minority-class sensitivity without sacrificing generalization, indicating that controlled synthetic generation can meaningfully expand the effective training distribution.

From an applied perspective, these findings have implications for real-time pharmacovigilance systems. In operational settings where rapid deployment and computational efficiency are required, base-sized transformer models enhanced through targeted augmentation may offer an optimal trade-off between accuracy and resource consumption. However, careful prompt design, quality control, and human validation remain essential to prevent distributional drift or synthetic bias.

Overall, the study demonstrates that performance improvements in VAEM detection are not solely dependent on increasing model size, but can be systematically achieved through strategic data-centric interventions. Synthetic augmentation, when combined with disciplined fine-tuning and validation procedures, represents a robust approach for mitigating imbalance, enhancing generalization, and improving reliability in social media–based vaccine safety monitoring.

5 Conclusion

This study investigated the impact of Large Language Model (LLM)–driven synthetic data augmentation on detecting personally experienced Vaccine Adverse Event Mentions (VAEMs) in social media posts within the #SMM4H Task 6 framework. By first establishing strong baselines through systematic fine-tuning of six transformer-based architectures, we demonstrated that domain-adapted large models—particularly twitter-roberta-large and

DeBERTa-v3-large-vaccine—achieve high classification performance even under moderate class imbalance. However, our results also confirmed that smaller and base-sized models remain more sensitive to skewed distributions and limited training data.

The introduction of 451 synthetically generated and manually validated posts substantially improved dataset balance and overall representational diversity. Empirical results show that augmentation yields measurable performance gains across most models, with the most significant improvements observed in base architectures, where F1-score increases of up to 4.4% were achieved on the test set. These improvements were accompanied by consistent gains in both precision and recall, indicating enhanced minority-class sensitivity and reduced bias toward the majority class. In contrast, the highest-capacity model exhibited minimal change after augmentation, suggesting that larger architectures may already possess sufficient representational robustness to mitigate moderate imbalance effects.

Collectively, the findings confirm that LLM-driven synthetic augmentation is an effective, scalable strategy for strengthening vaccine reaction detection in low-resource and imbalanced social media settings. Beyond raw performance gains, augmentation enhances decision boundary stability and improves generalization, particularly for computationally efficient models that are more suitable for real-world deployment.

References

[1] Guellil, I., Berrachedi, Y., Chenni, N. *et al.* Detecting Adverse Drug Events in Social Media: A Brief Literature Review. *SN COMPUT. SCI.* **7**, 199 (2026). https://doi.org/10.1007/s42979-026-04752-9

[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] 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.

[4] Abeed Sarker et al. (2016). Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. Drug Safety. 39. 10.1007/s40264-015-0379-4.

[5] Bosung Kim and Ndapa Nakashole. 2022. Data Augmentation for Rare Symptoms in Vaccine Side-Effect Detection. In Proceedings of the 21st Workshop on Biomedical Language Processing, pages 310–315, Dublin, Ireland. Association for Computational Linguistics.

[6] Ahmet Okan Arık, Gizem Parlayandemir, Serra Çelik (2026), LLM-based data augmentation for text classification on imbalanced datasets: A case study on fake news detection, Egyptian Informatics Journal, Volume 33, 2026,100886, ISSN 1110-8665, https://doi.org/10.1016/j.eij.2026.100886.

[7] 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.

[8] Bosung Kim and Ndapa Nakashole. 2022. Data Augmentation for Rare Symptoms in Vaccine Side-Effect Detection. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 310–315, Dublin, Ireland. Association for Computational Linguistics.

[9] Yuan Chen, Zhisheng Zhang, An easy numeric data augmentation method for early-stage COVID-19 tweets exploration of participatory dynamics of public attention and news coverage, Information Processing & Management, Volume 59, Issue 6, 2022, 103073, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2022.103073.

[10] Simone Scaboro, Beatrice Portelli, and Giuseppe Serra, Detection of Adverse Drug Events from Social Media Texts - Research Project Overview77-86, in proceedings of HC@AIxIA 2022: 1st AIxIA Workshop on Artificial Intelligence For Healthcare, November 30, 2022, Udine, IT

[11] Feng X, Luo J, Yang Y, El Baz D, Shi L. Health Misinformation Detection: Approaches, Challenges and Opportunities. Inquiry. 2025 Jan-Dec;62:469580251384784. doi: 10.1177/00469580251384784. Epub 2025 Nov 4. PMID: 41189452; PMCID: PMC12589804.

[12] Abdelsalam Nwesri, Mai Elbaabaa, Nabila Shinbir, Enhancing Vaccine Reaction Detection from Social Media Using Optimized Transformer Fine-Tuning, **Libyan Journal of** InformaticsVolume 03**,** No**.** 02**,** December. 202**5.**

[13] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November 2020. Association for Computational Linguistics.

[14] 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.

[15] 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.

[16] 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.

[17] 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 Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada, July 2023. Association for Computational Linguistics.

[18] JacobDevlin,Ming-WeiChang,KentonLee,andKristinaToutanova.BERT:pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.