

NU_Health_Miners at #SMM4H-HeaRD 2025: Bagging Methods for Detecting Adverse Vaccine Events

Bhavana Rajan Nair, Deahan Yu

Northeastern University
{rajannair.b, d.yu}@northeastern.edu

Abstract

This paper details our classification model developed for the 10th Social Media Mining for Health Research and Applications Workshop (SMM4H 2025), addressing Task 6 focused on a binary classification task to distinguish Reddit posts that contain mentions of adverse vaccine events. Our objective was to enhance the detection of posts discussing adverse reactions to herpes zoster (shingles) vaccines from other vaccine-related discussions. To do this, we used a pre-trained language model, RoBERTa, in various sizes and various training methods. As we observe unstable fluctuations in performance metrics during training, we implement the ensemble approach of bagging that combines predictions from different models. Although our best-performing model achieved F1 scores of 0.96 on the validation set and 0.94 on the test set, the experiment indicates that the bagging approach contributes to improved generalizability.

1. Introduction

Patient-reported experiences from online discussions provide valuable unstructured data on lived experiences, barriers to care, and treatment effectiveness. Among such platforms, Reddit provides a unique environment for candid, community-driven discourse on personal experiences with medications and vaccines. Since these discussions are often anonymous, they allow individuals to share their perspectives openly without fear of judgment. This potential is especially critical in pharmacovigilance efforts aimed at identifying adverse drug reactions. These efforts rely on the timely and accurate detection of harmful side effects from medications, often using data from diverse and informal sources like social media.

Task 6 of the 10th Social Media Mining for Health (SMM4H 2025) workshop presents a focused challenge to accurately classify Reddit posts that mention personal adverse reactions to herpes zoster (shingles) vaccines (Klein et al. 2025). To address this, we employed RoBERTa, a robust transformer-based language model

pre-trained on a large corpus of English text (Liu et al. 2019). Recognizing the limitations posed by dataset imbalance and linguistic variability, we experimented with various RoBERTa model sizes and fine-tuning strategies. Observing fluctuations in individual model performance, we adopted an ensemble bagging strategy to stabilize predictions and improve generalization.

2. Related Work

The systematic review of social media based surveillance systems for healthcare (Gupta and Katarya 2020) examined the potential of machine learning techniques for early detection and public health monitoring on social media. One example is the collection of adverse events following the use of drugs and vaccines (Klein et al. 2025). (Sarker and Gonzalez 2015) highlighted the challenges of extracting adverse drug reactions from social media due to informal language, misspellings, and idiomatic expressions. They proposed various machine learning approaches to address these challenges. Building on this, (Breden and Moore 2020) came up with a domain-specific preprocessing pipeline combined with BERT ensembling to detect adverse drug reactions from Twitter data. More recently, a study published in Vaccines (Lian et al. 2022) used machine learning and natural language processing to identify COVID-19 adverse vaccine adverse events (AVEs) from Twitter posts. Their findings highlight the growing role of social media in pharmacovigilance and the feasibility of using such platforms for large-scale AVEs detection.

Task 6 of the 10th SMM4H 2025 is related to the benchmark paper (Khademi et al. 2024), which explored the application of large language models, specifically GPT-3.5 and GPT-4, to detect personal mentions of vaccine reactions on Reddit. They evaluated various prompting strategies, including zero-shot and few-shot learning with both standard and chain-of-thought prompts. Their findings indicated that pretrained language models outperformed LLMs in classification tasks. Our study builds on their work by further analyzing and developing an NLP

approach to enhance detection methods for adverse vaccine reactions in social media.

3. System Description

3.1. Problem and Dataset

The goal of Task 6 is to develop a binary classification model that can accurately identify Reddit posts containing personal mentions of AVEs. This work aims to address these challenges by enhancing data quality through preprocessing and stabilizing model performance using ensemble techniques.

The dataset suffered from issues in data quality, particularly representative positive examples, which led to the addition of 1,016 synthetic records to the training set (Khademi et al. 2024). The class distribution of synthetic data is unknown to us. As a result, the training, validation, and test sets consist of 2,521, 786, and 629, respectively, with a balanced class distribution where positive means the post contains mentions of AVEs and negative means it does not.

Dataset	Negative	Positive	Total
Training	1,372	1,149	2,521
Validation	420	366	786
Test	338	291	629

Table 1: Data distribution on each dataset.

3.2. Base Model

We experimented with two variants of RoBERTa: RoBERTa-base and RoBERTa-large, as our primary pre-trained language models from Hugging Face. RoBERTa-base consists of 12 transformer layers with 125 million parameters, while RoBERTa-large has 24 layers and 355 million parameters, enabling it to capture deeper contextual representations. Both models were fine-tuned on our task-specific dataset to classify Reddit posts, with RoBERTa-large generally demonstrating stronger performance.

Additionally, we explored various prompting strategies for LLaMA-2 (Touvron et al. 2023), ranging from simple task instructions to few-shot learning with example inputs and outputs. We also experimented with hybrid models combining LLaMA-2 with a Linear Regression layer and with RoBERTa. However, these combinations did not outperform the standalone RoBERTa model. Our results align with the findings of (Klein et al. 2025), which further confirms the appropriateness of the model choice for this task.

3.3. Implementation Details

Initially, we experimented with task-specific fine-tuned models such as Twitter RoBERTa but found that the standard (vanilla) RoBERTa models performed better. We observed that the dataset suffered from issues in data quality, and to overcome this, we experimented with traditional machine learning models such as Support Vector Machines (SVM), Logistic Regression (LR), and Random Forest (RF) using term frequency-inverse document frequency features. Their performance plateaued well below that of transformer-based models, with F1 scores not exceeding 0.93.

We found that 5% (130) of the training data starts with quotation marks. Of those, only 8 examples belong to the negative class, and the rest in the positive class. Hence, we preprocessed our data to remove this quote, only when they appeared at the beginning of the text, and saw a slight change in the F1 score from 0.95 to 0.96.

Observing fluctuations in individual model performance, we adopted an ensemble bagging strategy to stabilize predictions and improve generalization. We believe the bagging approach would help to mitigate the change in natural class distribution due to the presence of the synthetic data. A bagging approach was also found to be effective for a similar task of identifying adverse drug reactions from Twitter (Breden and Moore 2020). The training data was split into 3 and 5 folds, and separate RoBERTa models were trained on each. In the 3-split approach, three RoBERTa models were trained on three overlapping subsets: the first 80% of the data, the middle 80% (ranging from 10% to 90%), and the last 80%. For 5-split, it was similar but with a sliding window of 60% of the data. A majority voting mechanism was then used to determine the final prediction.

While the 3-fold ensemble yielded a stable F1 score of 0.95, the 5-fold configuration showed reduced performance, achieving only 0.80, likely due to over-fragmentation of the training data. Splitting the dataset into five folds reduces the effective training size available to each model, which likely contributed to the underperformance of the 5-fold bagging ensemble compared to the 3-fold approach. The smaller training subsets resulted in weaker individual learners, thereby lowering overall ensemble performance. During model training, we found that the best F1 scores were consistently achieved with two epochs. Consequently, all results reported in the paper reflect training over two epochs. We used a batch size of 8 and a learning rate of 2×10^{-5} , while all other hyperparameters remained at their default settings.

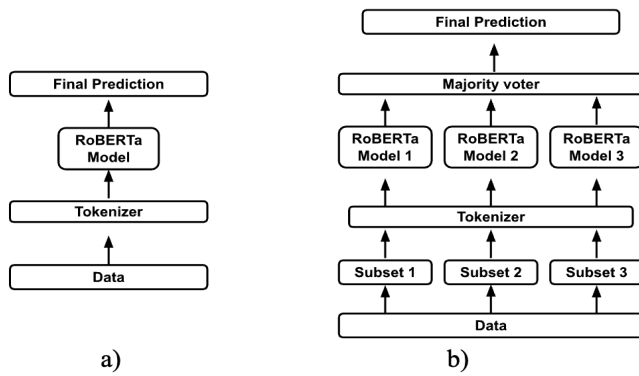


Figure 1: Model architecture, a) with pre-trained RoBERTa b) Bagged RoBERTa.

4. Evaluation

In Table 2, we can see the validation results of different model setups. We report F1, precision, and recall scores for individual RoBERTa models, as well as for the ensemble bagging approaches. The results highlight how preprocessing, model selection, and ensemble strategies impacted overall performance, particularly in handling the noisy and imbalanced dataset.

Preprocessing	Models	F1	Precision	Recall
None	LR	0.93	0.93	0.93
	SVM	0.92	0.92	0.92
	RF	0.92	0.92	0.92
	RoBERTa	0.95	0.95	0.95
Quotes removed	RoBERTa	0.96	0.96	0.96
	RoBERTa + LR	0.96	0.95	0.96
	Bagged RoBERTa (3)	0.95	0.91	0.99
	Bagged RoBERTa (5)	0.80	0.75	0.91

Table 2: Results of our models on the validation set. For preprocessing, quotes were removed only when they appeared at the beginning of the text.

Table 3 discusses the performance of our top models on the test data. Our model ranked 7th overall in terms of F1 score, achieving a competitive 0.94 and with a recall of 0.99, which is the table topper for the recall metric. In an application like collecting mentions of adverse vaccine events, missing relevant mentions can be more costly than being precise.

Models Submitted	F1	Precision	Recall
RoBERTa	0.943	0.9	0.99
Bagged RoBERTa (3)	0.94	0.905	0.98
Task Mean	0.938	0.916	0.961
Task Median	0.944	0.922	0.972

Table 3: Evaluation results by our models on the test set, together with the mean and median results of the task.

5. Discussion and Conclusion

Through careful data analysis and preprocessing, LLMs could be leveraged to generate content, where synthetic data will help address class imbalance and expand the diversity of training examples. Our analysis revealed that over-reliance on synthetic data may have hindered the model's ability to generalize, as it seemed to overfit certain patterns or phrasings typical of GPT-generated content. Having visibility into which samples are synthetic versus real in the provided dataset would give more opportunities to understand the data distribution and model accordingly.

The best configuration, which used RoBERTa, achieved an F1 score of 0.96 on the validation set and of 0.943 on the test set. However, when comparing performance across both datasets, the Bagged RoBERTa model exhibited less fluctuation across all three metrics. This observation indicates the bagging approach offers more stable and consistent performance despite its slightly lower overall scores. We plan to further develop and expand this approach in future work.

References

- Breden, A., & Moore, L. 2020. Detecting Adverse Drug Reactions from Twitter through Domain-Specific Preprocessing and BERT Ensembling. *Unpublished manuscript*, UC Berkeley School of Information.
- Gupta, A., & Katarya, R. 2020. Social media-based surveillance systems for healthcare using machine learning: A systematic review. *Journal of Biomedical Informatics*, 108: 103500. <https://doi.org/10.1016/j.jbi.2020.103500>
- Khademi, S., Palmer, C., Dimaguila, G. L., Javed, M., & Buttery, J. 2024. Exploring Large Language Models for Detecting Online Vaccine Reactions. In J. Bichel-Findlay (Ed.), *Health. Innovation. Community: It Starts With Us* (Vol. 318, pp. 30–35). IOS Press.
- 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., & Gonzalez-Hernandez, G. 2025.
- Lian, A. T., et al. 2022. Identifying COVID-19 Vaccine Adverse Events from Twitter: A Machine Learning and Natural Language Processing Approach. *Vaccines*, 10(1): 103. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8781534/>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*. <https://doi.org/10.48550/arXiv.1907.11692>

Overview of the 10th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks

at ICWSM 2025. In *Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.

Sarker, A., & Gonzalez, G. 2015. Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-Corpus Training. *Journal of Biomedical Informatics*, 53: 196–207. <https://doi.org/10.1016/j.jbi.2014.11.002>

Touvron, H., Martin, L., Stone, K., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, S., Bhosale, S., Bikel, D., Blecher, L., Clark, J., Cucurull, G., Ekin, A., Fernandes, J., Fu, Y., Goyal, N., Hossain, F., Hou, L., ... Scialom, T. 2023. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.