

# BioNLP1 at #SMM4H-HeaRD 2025: Stacked Representations for Detection of Vaccine Adverse Event Mentions in Social Media Text

Andra Păsărin<sup>1</sup>, Ana-Sabina Uban<sup>1,2</sup>

<sup>1</sup>University of Bucharest, Faculty of Mathematics and Computer Science

<sup>2</sup>Human Language Technologies Research Center  
andra.pasarin@s.unibuc.com, auban@fmi.unibuc.ro

## Abstract

This paper presents the system developed by the BioNLP1 team for Task 6 of the 10th edition of the Social Media Mining for Health (SMM4H) 2025. The task involves the binary classification problem of identifying vaccine adverse event mentions (VAEM) in shingles-related Reddit posts. Our approach leverages a stacking method that combines TF-IDF features with sentence embeddings from a fine-tuned transformer model (RoBERTa-large) to classify the posts. The proposed system achieved a high F1 score of 0.97 on the validation set and the highest F1 score of 0.96 on the test set, with scores computed on the positive class.

## Introduction

Adverse drug events (ADEs) are usually identified during clinical trials before a drug or vaccine is released to the market. However, there are cases in which new symptoms may occur post-administration. In such cases, social media can play a valuable role in detecting adverse events in real time. Extracting information from user-generated content such as posts or comments has become an important tool for identifying vaccine adverse event mentions (VAEM), as social media users often represent diverse demographic groups. Despite the potential of social media data for monitoring vaccine-related adverse events, accurately detecting such mentions using machine learning and natural language processing remains an ongoing challenge (Golder et al. 2024). This difficulty primarily stems from the noisy nature of social media text, which often includes informal language, slang, typographical errors and non-standard expressions used by non-native English speakers.

Task 6 of SMM4H 2025 workshop consists of classifying social media posts as either containing or not containing a vaccine adverse event mention (Klein et al. 2025). Transformer-based pretrained language models (PLMs) have been widely adopted for text classification tasks involving social media data. These models are often adapted through domain-specific pretraining on platforms like Twitter and Reddit to better handle the informal and noisy nature of user-generated content. Recently, large language models (LLMs) have been applied to classification tasks in both supervised and zero-shot settings, showing good performance

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

on medical and domain-specific content (Guo et al. 2024). Previous work by Khademi et al. shows that large language models (LLMs) are capable of detecting VAEMs, but cannot surpass a fine-tuned transformer-based PML on the same task (Khademi et al. 2024).

Our approach focuses on two distinct types of text representations for classification: TF-IDF features and sentence embeddings. The sentence embeddings were extracted using various fine-tuned transformer-based pretrained language models (PLMs). These representations were then used as input features for a support vector machine (SVM) and a logistic regression (LR) model.

## System Description

### Dataset

The dataset contains English Reddit posts related to shingles vaccines, collected using targeted search terms via the Reddit API. Posts were sourced from relevant subreddits and were manually labeled by experts based on predefined criteria (Khademi et al. 2024). The training, validation and test sets contain 2521, 786 and 8113 posts, respectively. Table 1 provides examples of samples for both classes.

text	label
Shingles vaccine: second dose: Good move. All I got was a sore shoulder on both. Those shots hurt.	1
Shingles shot?: 27 year old that had shingles almost two years ago, it was hell. Just get the vaccine	0

Table 1: Examples of samples

**Baseline.** The organizers provide a benchmark classifier which obtains an F1 score of 0.946 on the positive class (i.e., posts that contain mentions of vaccine adverse events) on the test set of Task 6. The model is based on a fine-tuned Twitter RoBERTa-large transformer (Khademi et al. 2024).

### Feature Extraction

**TF-IDF Features.** Term frequency-inverse document frequency transformation is used to extract lexical features from the text, capturing word and phrase frequency patterns. The vectorizer is configured with a maximum of 5000

features, applied lowercasing and included n-grams ranging from unigrams to 5-grams. Parameters were chosen empirically based on performance.

**Sentence Embeddings.** To extract contextual sentence embeddings, we fine-tuned several transformer models on the training data (Amit 2022). The transformers consist of multiple variants of BERT and RoBERTa models, used from the Hugging Face’s Transformer Library (Wolf et al. 2020): distil-BERT (Sanh et al. 2019), BERT-base (Devlin et al. 2018), RoBERTa-base, RoBERTa-large (Liu et al. 2019), twitter-RoBERTa-base (Loureiro et al. 2022), biomed-RoBERTa-base (Gururangan et al. 2020). The models were trained for 5 epochs using a learning rate of 5e-5, a batch size of 16 and weight decay of 0.01, with gradient accumulation and early stopping to improve training stability and efficiency (random state=9). Text inputs were tokenized with a maximum sequence length of 512 and padded using Hugging Face’s AutoTokenizer. All parameters of each model were left trainable, as we observed that freezing them resulted in lower performance. After training, sentence embeddings were obtained by performing mean pooling over the last hidden states of the model and weighted by the attention mask to account for padding. This method produces dense, context-sensitive representations of entire posts, capturing semantic relationships between tokens beyond the surface-level information provided by traditional lexical features. Table 6 from the supplementary material shows all transformer models used and their performance on the validation set, along with the number of trainable parameters for each model. Best results were obtained using RoBERTa-large, with an F1 score of 0.965 on the positive class.

## Ensemble Learning

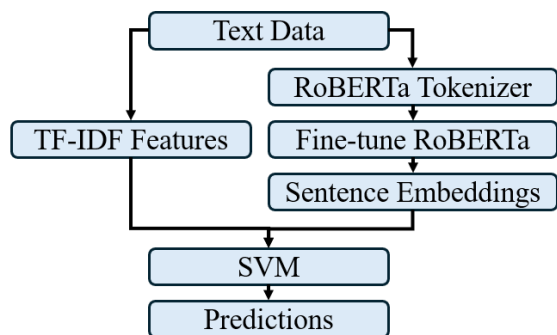


Figure 1: Architecture of the system

We combined the two complementary feature representations: TF-IDF vectors and sentence embeddings extracted from RoBERTa-large fine-tuned transformer model. The features were stacked and standardized using StandardScaler to ensure consistent scaling across dimensions. For classification, we used both logistic regression and support vector machine (SVM) with RBF kernel ( $C=3.0$ , random state=0). Logistic regression was chosen for its simplicity and interpretability, while SVM was used for its effective-

ness in handling high-dimensional feature spaces and non-linear decision boundaries. Hyperparameters were chosen empirically. The architecture of the system is illustrated in Figure 1.

## Results

### Validation Set

The results in Table 6 (supplementary material) show that RoBERTa-based models outperform their BERT counterparts in all evaluation metrics. Among them, RoBERTa-large achieves the best performance, with an F1 score of 0.965, precision of 0.957 and recall of 0.973. Interestingly, twitter-RoBERTa-base, a model pretrained on social media data, has a performance similar to the general-purpose RoBERTa-base, suggesting that domain-specific pretraining on social media does not offer additional gains over the base model, on this specific task. On the other hand, biomed-RoBERTa-base, pretrained on biomedical texts, underperforms both RoBERTa-base and twitter-RoBERTa-base, potentially indicating that biomedical domain adaptation is less effective for classifying informal and noisy social media posts.

Features	Classifier	F1	Precision	Recall
TF-IDF	LR	0.919	0.935	0.904
TF-IDF	SVM	0.892	0.895	0.888
SE	LR	0.966	0.957	<b>0.975</b>
SE	SVM	<b>0.967</b>	0.962	0.973
TF-IDF+SE	LR	0.965	0.962	0.967
TF-IDF+SE	SVM	<b>0.967</b>	<b>0.965</b>	0.970

Table 2: Performance comparison on validation set using the best sentence embedding (SE) model and the TF-IDF model

Table 2 compares the performance on validation data of different combinations of feature representations and classifiers. In particular, the simplest approach, TF-IDF features with logistic regression (LR), achieves an F1 score of 0.919, demonstrating that even basic lexical representations paired with a linear model can yield competitive results. However, models based on sentence embeddings (SE) consistently outperform standalone TF-IDF, with the best performance observed when SE is combined with TF-IDF. The highest F1 score of 0.967 is obtained by both SE + SVM and TF-IDF + SE + SVM combinations, with the latter achieving the highest precision of 0.965. Models using only SE tend to achieve better recall scores, while combined representations offer a more balanced trade-off between precision and recall. Most configurations achieve a higher recall score than precision, suggesting that the models are more likely to predict false positives than false negatives. This behavior is desirable in health-related tasks, where missing true cases may be more critical than occasional over-identification.

### Submissions and Test Set

The submissions were chosen based on the results obtained on the validation set. Both approaches use sentence embeddings extracted from RoBERTa-large fine-tuned transformer

model. The difference consists in the presence or absence of the TF-IDF features. Although the results are similar, we notice that lexical features add marginal performance improvement. Table 3 shows that both submissions surpass the baseline (0.946), mean and median scores of Task 6.

Submission	F1	Precision	Recall
RoBERTa SE	0.958	0.943	0.972
TF-IDF + RoBERTa SE	<b>0.959</b>	<b>0.946</b>	<b>0.972</b>
Task Mean	0.938	0.916	0.961
Task Median	0.944	0.922	<b>0.972</b>

Table 3: Performance comparison on test set

## Error Analysis

To better understand the model’s performance, we conducted an error analysis on our best-performing approach (TF-IDF + SE + SVM). The confusion matrix for the validation data is shown in Figure 2 in the supplementary material. The results indicate that 407 samples were correctly classified as label 0, while 13 were misclassified as false positives. Similarly, 355 samples were correctly classified as label 1, with 11 false negatives. Examples of both false positive and false negative samples can be found in Tables 4 and 5.

The false negatives (FN) generally reflect the difficulty in identifying subtle mentions of vaccine administration or adverse events. In both examples from Table 4, phrases like “got it” and “the 3rd one” implicitly refer to vaccine doses without explicitly mentioning them, making it harder for the model to detect these references and likely contributing to misclassification. The false positive (FP) cases highlight the model’s difficulty in handling ambiguous language, particularly when medical or vaccination-related terms are present without a clear indication of an adverse event. Additionally, some posts include symptoms or reactions without clearly stating whether a vaccine was administered, making it difficult for the model to distinguish between general medical experiences and vaccine-specific adverse events. These examples highlight the ambiguity in the user’s posts, particularly in distinguishing personal experiences, general statements and subtle implications.

We also performed vocabulary analysis to identify the most distinctive words associated with misclassifications. By applying the chi-squared test, we extracted the most discriminative features between FP and TP cases, and separately between FN and TN cases. Across all configurations, we can see that the words that cause FPs are medical and treatment-related terms (“5mg”, “kesimpta”, “doc”), which may trigger incorrect associations with adverse events. Notably, “anxiety” appears in both TF-IDF+SE and SE configurations, reinforcing the idea that models are prone to falsely associating psychological distress with adverse events. In contrast, words that cause FNs include actual vaccine names (“shingrex”, “hpv”) and symptom-related terms (“bloating”, “lightheaded”), which might be too implicit or context-dependent for the model to detect. The presence of “reassurance” in the SE configuration further suggests that models struggle to identify adverse events when the language is

## False Negatives

A roofer with shingles. How ironic...: Yes, it’s been in the same area for me each time. I’m the wrong person to ask about the vaccine: I got it 4 years ago, and it made my outbreaks worse and more frequent.

My sister’s mother-in-law had shingles triggered, along with clots after 1 and 2 that paralyzed half her body temporarily. Oddly enough she went for the 3rd one anyway, so far no adverse reactions. The main issue is no one will talk about it.

Table 4: Examples of false negative samples

## False Positives

Turned 50 and got my first shingles vaccine yesterday evening: My wife(52) had zero side effects. Sorry y’all!

Any experience with Shingrix?: Hi, folks. For those of you’ve who’ve gotten the new two-dose Shingrix vaccine, did you have any ill effects from the shot? Did it cause a flare? Did your second dose produce different effects from the first? Are your healthcare practitioners all in favor of you getting the vaccine? Thanks!

Table 5: Examples of FP and FN samples

overly optimistic. Table 7 in the supplementary material contains more examples of the words with the highest scores.

As a last step, we conducted error analysis on all 3 configurations, by comparing their predictions on the same examples. Our most simple configuration (TF-IDF) excels at avoiding false positives, but underperforms when adverse events are described using uncommon or indirect terminology, or when symptoms are scattered contextually (Examples 1 and 3). Sentence Embeddings (SE) are more capable of picking up semantic meaning, which helps in identifying nuanced expressions of adverse events (Example 3), but they are prone to false positives in emotionally charged or ambiguous contexts (Example 4). The stacked representation (TF-IDF+SE) shows the most balanced and context-aware behavior, but it may still miss subtle symptoms buried in dense language (Example 3). Table 8 in the supplementary material contains the predictions on each example.

## Conclusions

In this work, we proposed a system for identifying mentions of vaccine adverse events in shingles-related Reddit posts by using a stacked representation, combining TF-IDF features with sentence embeddings from a fine-tuned RoBERTa-large model. Our approach demonstrated strong performance on both validation and test sets, outperforming the baseline with an F1 score of 0.959. Through our experiments, we observed that models with more parameters achieved better performance. Additionally, combining lexical and contextual representations further improved classification performance. While the system performs well overall, error analysis highlighted the challenges of interpreting ambiguous or indirect references to vaccine adverse events in social media posts.

## Appendix

Model	F1	Trainable Params
distil-BERT	0.937	67M
BERT-base-uncased	0.938	110M
RoBERTa-base	0.955	125M
RoBERTa-large	<b>0.965</b>	355M
twitter-RoBERTa-base	0.955	125M
biomed-RoBERTa-base	0.947	125M

Table 6: Performance and trainable parameters comparison of fine-tuned transformers on validation set

Experiment	False Negatives	False Positives
TF-IDF+SE	gd, tsyabri, verify, reactions, reading, shingrex, hpv, particularly, lightheaded, bloating	5mg, cibinqo, loading, zero, anxiety, kesimpta, contain, send, office, strongest
TF-IDF	barre, guillain, phn, covid, vaccines, shot, lips, spot, vaccine, gypsy	major, 5mg, nervous, cibinqo, neuro, kesimpta, talk, loading, makes, doc
SE	verify, tsyabri, reactions, reading, shingrex, hpv, regularly, contracting, reassurance	5mg, loading, cibinqo, zero, wait, kesimpta, anxiety, local, summary, happening

Table 7: Top discriminative words (p-values  $< 10^{-6}$ ) for incorrect and correct predictions

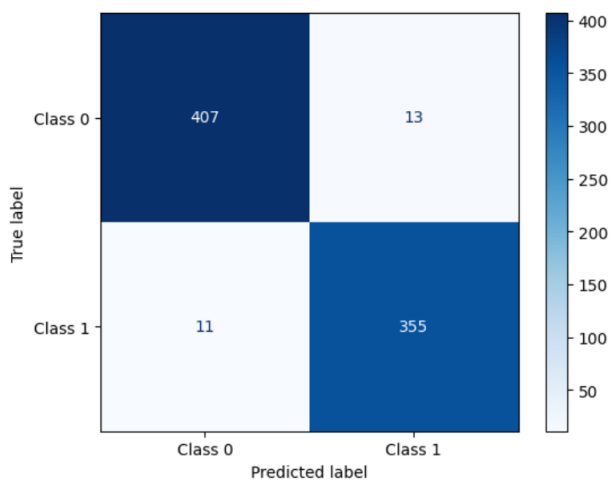


Figure 2: Confusion Matrix of predictions on validation set

Example	M1	M2	M3	G
Shingles vaccine: second dose: I haven't had a second one. Doctor told me the one using now is better than what they used to use. I didnt get sick but my arm hurt like a bitch for several days	1	0	1	1
A roofer with shingles. How ironic...: Yes, it's been in the same area for me each time. I'm the wrong person to ask about the vaccine: I got it 4 years ago, and it made my outbreaks worse and more frequent.	0	1	0	1
Initial Symptoms: Hello friends. Looking for some insight, my Mom got the shingles vax and thinks it's caused her an AI Disease, particularly GD. She has burning eyes, lips, face, fast heart rate (as in 120+), severe anxiety and depression (neither of which she ever had and now has at age 70), weight loss...were these your first symptoms when realizing you had GD? If not, would you mind sharing what prompted you to realize something was wrong and ultimately coming to the conclusion you had GD. Thanks everyone for any help you can provide...hope you're all feeling well today ??	0	0	1	1
Should I start my rinvoq: I read it could take a couple of weeks for effectiveness after the second dose, I would recommend to wait. Rinvoq has a heightened risk of shingles, it says that in the detailed summary and I didn't read that and got it in my eye before I got the vaccine. It was terrible, still have effects from it almost a year later.	0	0	1	0

Table 8: Comparison between predictions of the models on the same examples (M1 is TF-IDF+SE, M2 is TF-IDF, M3 is SE and G is the ground-truth)

## Acknowledgements

This study was partially funded by a grant of the Ministry of Research, Innovation and Digitization, CCCDI - UEFIS-CDI, project number PN-IV-P7-7.1-PTE-2024-0046, within PNCIDI IV and by "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

## References

- Amit, H. 2022. Fine-Tuning BERT for Classification: A Practical Guide. Accessed: 2025-04-10.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.
- Golder, S.; O'Connor, K.; Wang, Y.; Klein, A.; and Gonzalez Hernandez, G. 2024. The Value of Social Media Analysis for Adverse Events Detection and Pharmacovigilance: Scoping Review. *JMIR Public Health Surveill*, 10: e59167.
- Guo, Y.; Ovadje, A.; Al-Garadi, M. A.; and Sarker, A. 2024. Evaluating Large Language Models for Health-Related Text Classification Tasks with Public Social Media Data. arXiv:2403.19031.
- Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of ACL*.
- Khademi, S.; Palmer, C.; Dimaguila, G. L.; Javed, M.; and Buttery, J. 2024. Exploring Large Language Models for Detecting Online Vaccine Reactions. *Studies in Health Technology and Informatics*, 318: 30–35.
- 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.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.
- Loureiro, D.; Barbieri, F.; Neves, L.; Espinosa Anke, L.; and Camacho-collados, J. 2022. TimeLMs: Diachronic Language Models from Twitter. In Basile, V.; Kozareva, Z.; and Stajner, S., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 251–260. Dublin, Ireland: Association for Computational Linguistics.
- Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.
- Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771.