BrynMawrNLP @ SMM4H-HeaRD 2025: Stacking RoBERTA with Naive Bayes to Identify Personal Shingles Vaccine Reactions on Reddit

Joon Luther Adam Poliak

Bryn Mawr College
{jluther,apoliak}@brynmawr.edu

Abstract

We present the BrynMawr NLP System for the 10th Social Media Mining for Health Research and Applications Workshop (SMM4H 2025), addressing Task 6 - distinguishing Reddit posts that contain personal mentions of adverse reactions to herpes zoster (shingles) vaccines from other vaccine-related discussions. Our system is based on a simple logistic regression model trained on the predictions from a fine-tuned RoBERTa model and Naive Bayes model. We explore stacking a fine-tuned RoBERTa model with traditional non-neural models to improve upon just fine-tuning a pre-trained RoBERTA model. Our experiments demonstrate that statistical, non-neural models perform well on this dataset but their performance do not compare with those of a fine-tuned Transformer.

Introduction

While Large Language Models are the state-of-the-art for most NLP tasks (Mars 2022; Zubiaga 2024), fine-tuning off-the-shelf Transformer-based contextualized language models is a common baseline for sentence-classication problems (Ruder 2021). In fact, Khademi et al. noticed that fine-tuning a RoBERTa model trained on Twitter (Loureiro et al. 2023) to predict whether a Reddit post described a personal health mention related to Shingles vaccine outperformed LLMs like GPT3 and GPT4 in both zero-shot and few-shot prompting settings.

When developing our system for Task 6 (Klein et al. 2025) of the 2025 Social Media Mining for Health Applications, we immediately noticed that “simple”, non-neural approaches can perform very well on this task. Therefore, we explored combining predictions from a fine-tuned RoBERTA model and a Naive Bayes model. Training both on all available data, i.e. the released training and development set, resulted in an F1 score of 93.2, Precision of 94.4, and Recall of 92.1 on the shared task’s evaluation set. Interestingly, just fine-tuning a RoBERTa model on the same dataset resulted in a less-conservative approach that earned a higher F1 of 95.0, lower Precision of 92.8, and higher Recall of 97.2.

Task & Dataset Description

Task 6 focuses on classifying Reddit posts to determine whether they contain information about personal adverse effects from the Shingles vaccine. The goal is to distinguish these specific posts from more general discussions about vaccines, shingles, or side effects. The dataset contains 3307 example Reddit posts, 1515 examples mentioning personal Shingles vaccine side effects and 1792 examples not. The dataset has been partitioned into 2521 examples for training and 786 for validation.

System Description

We evaluate the performance of machine learning classifiers on the dataset using 5-fold cross validation. We compare logistic regression, Naive Bayes, and Decision Trees using unigram, bigram, and trigram features, using sklearn’s python implementions (Pedregosa et al. 2011).

We begin by comparing how well logistic regression, naive bayes, and decision tree perform when using uni-gram, uni-gram & bi-gram features, and uni-gram, bi-gram, and tri-gram features. With uni- and bi-gram features, Logistic Regression had the highest average F1 score across 5 folds of 90.74. Table 1 demonstrates that Logistic Regression models slightly outperform Naive Bayes models and drastically outperform Decision Trees. Therefore, we perform hyperparameter tuning for Logistic Regression and Naive Bayes.

Model	Feature Set	Precision	Recall	F1
Logistic Regression	Unigram Features	90.06	90.25	90.15
Logistic Regression	Bigram Features	91.83	89.69	90.74
Logistic Regression	Trigram Features	91.92	89.37	90.62
Naive Bayes	Unigram Features	86.58	91.81	89.11
Naive Bayes	Bigram Features	87.19	93.73	90.32
Naive Bayes	Trigram Features	85.14	95.13	89.84
Decision Tree	Unigram Features	83.51	82.70	83.09
Decision Tree	Bigram Features	82.87	81.73	82.25
Decision Tree	Trigram Features	83.79	82.74	83.23

Table 1: ML Model Performance with uni-, bi-, and tri-gram features. Scores are averaged across 5 folds. There is a clear difference in how well LR and NB perform compared to Decision Trees.

For logistic regression, we explore various values for C, penalty, and solver, while for Naive Bayes, we test different alpha and fit prior values. Using a combination of ni-, bi-, and tri-gram features with hyperparameters C=10, penalty=l2, and solver=lbfgs achieved the highest average F1 score of **90.77** among the logistic regression models. For Naive Bayes, the best performing model achieved an average F1 of **90.40** with a combination of uni-, bi-, and tri-gram features with an α of 0.1 and using the best fit prior.

After evaluating the ML models and performing hyperparameter tuning, we compare their performance with the RoBERTa classifier (Zhuang et al. 2021). Since the statistical models noticeable underperform the RoBERTa model reported by Khademi et al., we similarly fine-tune a RoBERTa model using 5-fold cross-validation on all available data. This model achieves an F1 score of 95.18, with 92.10 precision and 98.48 recall, demonstrating its superiority over the other methods.

Although RoBERTa outperforms traditional ML methods, we want to explore whether improvement is possible with an ensemble of the classifiers examined. We train three simple logistic regression classifiers based on the predicted probabilities from the fine-tuned RoBERTa model, the logistic regression model, and the naive Bayes model. We test combinations of Naive Bayes with RoBERTa with Naive Bayes, Roberta with Logistc Regression, and RoBERTa with both Naive Bayes and Logistic Regression. As shown in Table 2, the Naive Bayes with RoBERTa stacked model achieved the highest F1 score. leading us to select this model for our submission.

Model Combination	F1	Precision	Recall
RoBERTA	95.18	92.10	98.48
LR + RoBERTa	95.55	93.00	98.22
NB + RoBERTa	**95.79**	93.41	98.28
LR + NB + RoBERTa	95.58	93.39	97.89

Table 2: Performance Metrics for Stacked Models

Results

To maximize the training data, we fine tuned both the RoBERTa and Naive Bayes models on the combined training and development sets. Using their predicted class probabilities, we trained our stacked model with the previously identified optimal parameters and threshold. This submission achieved an F1 score of 93.2 with 94.4 precision and 92.1 recall. For comparison, we also submitted a standalone RoBERTa model trained on the same combined dataset. This submission achieved an F1 score of 95.0, with 92.8 precision and 97.2 recall, placing our submission in 5th place.

Analysis

To analyze the Logistic Regression model, we show the top 10 features from the that are most strongly associated with the positive and negative classes (Table 3). Feature weights were aggregated by summing their values across 5 cross-validation folds to obtain an overall weight. The terms most associated with a positive classification are related to vaccines, e.g. “arm”, “shot”, ‘vaccine”, and symptoms, e.g. “sore” and “fever”. The features most associated with the negative class seem to be stylistic, e.g, “:” and “? :”.

Top Positive Features		Top Negative Features	
Feature	**Weight**	**Feature**	**Weight**
second	7.6816	:	-4.3821
vaccine	7.0912	? :	-3.4459
arm	5.9677	you	-3.2938
shot	5.8587	years	-3.2540
days	5.5547	dose : same	-2.7086
both	5.5541	n’t	-2.6055
after	5.3971	what	-2.3354
sore	5.2603	zoster	-2.3232
fever	3.9379	get	-1.8140
2nd	2.9134	: same	-1.2765

Table 3: Overall Top 10 Positive and Negative Features for Logistic Regression

To analyze the Naive Bayes model, we examine the log probabilities of each feature given the positive and negative classes. Features that are strongly associated with one class will show a large difference between these log probabilities. We show the 10 features with the greatest positive and greatest negative log probability differences (Table 4). Features like “& johnson vaccine,” “johnson vaccine,” and “astrazeneca” are strongly associated a positive classification, while features such as “colonoscopy” “well , i” and “first colonoscopy” are strongly associated with a negative classification.

Top Positive Features	Top Negative Features
& johnson vaccine	colonoscopy .
johnson vaccine	well , i
astrazeneca	first colonoscopy
astrazeneca vaccine	i turned 50
moderna shot	my first colonoscopy
, i experienced	, i turned
after getting vaccinated	got off
body aches .	just got off
else experienced this	first colonoscopy .
for a day	, scheduling my

Table 4: Overall Largest Differences in Log Probability of a Feature Given Class

Conclusion

We presented our straightforward approach of stacking a fine-tuned RobERTa model and Naive Bayes model on Task 6 of the 2025 Social Media Mining for Health Applications shared task. While not attaining state-of-the-art, these results are competitive and provide further evidence for the effectivenes of fine-tuning Transformer models for sentence classification problems. We additionally explored the benefits of combining statistical models with Transformer models. Future work might consider using different ensemble approaches to combine predictions from multiple models.

References

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*, 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.; 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.

Loureiro, D.; Rezaee, K.; Riahi, T.; Barbieri, F.; Neves, L.; Anke, L. E.; and Camacho-Collados, J. 2023. Tweet Insights: A Visualization Platform to Extract Temporal Insights from Twitter. *arXiv preprint arXiv:2308.02142*.

Mars, M. 2022. From word embeddings to pre-trained language models: A state-of-the-art walkthrough. *Applied Sciences*, 12(17): 8805.

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.

Ruder, S. 2021. Recent Advances in Language Model Fine-tuning. http://ruder.io/recent-advances-lm-fine-tuning.

Zhuang, L.; Wayne, L.; Ya, S.; and Jun, Z. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In Li, S.; Sun, M.; Liu, Y.; Wu, H.; Liu, K.; Che, W.; He, S.; and Rao, G., eds., *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, 1218–1227. Huhhot, China: Chinese Information Processing Society of China.

Zubiaga, A. 2024. Natural language processing in the era of large language models. *Frontiers in Artificial Intelligence*, Volume 6 - 2023.