Progress Report: Improving HPV Vaccine Misinformation Detection on Twitter Using a Hybrid TwHIN-BERT-LSTM Model

Yujie Pei
Department of Computer Science, University of Saskatchewan
105 Administration Place, Saskatoon, SK S7N 5A2
yup897@usask.ca

Abstract

Despite strong scientific consensus on its safety and effectiveness, HPV vaccine acceptance remains a challenge due to widespread misinformation on social media platforms like Twitter. This research aims to support public health efforts by enhancing misinformation detection and promoting vaccine confidence through data-driven strategies. A hybrid TwHIN-BERT-LSTM model is proposed in this project that combines the socially contextualized representations of tweets in TwHIN-BERT with the sequential processing capabilities of LSTM. This project hypothesizes that the integration of TwHIN-BERT and LSTM will enhance the model performance in identifying HPV vaccine-related misinformation.

The proposed model is evaluated against two baseline models, TwHIN-BERT-Misinformation-Classifier and TwHIN-BERT-Simple-Misinformation-Classifier, using accuracy, precision, recall, and F1 score as performance metrics. Preliminary results show that TwHIN-BERT-Misinformation-Classifier outperforms TwHIN-BERT-Simple-Misinformation-Classifier across all metrics. The final assessment of the hybrid model will be conducted on the completed HPVAXLIES dataset, and detailed findings will be presented in the final report.

1 Introduction

Human Papillomavirus (HPV) is the most common sexually transmitted infection worldwide, and the HPV vaccine has been rigorously tested and proven to be both safe and highly effective in preventing HPV-related cancers and other associated diseases. Despite these well-documented benefits, vaccine acceptance remains a significant challenge due to the widespread dissemination of misinformation—particularly on social media platforms such as Twitter. Misinformation can distort public perceptions, fuel vaccine hesitancy, and undermine trust in scientific research. Hence, detecting misinformation, identifying the types of misinformation being spread, and understanding how users respond to it is crucial for public health. Gaining these insights allows health organizations and policymakers to design targeted interventions that correct false narratives, reinforce scientific accurate information, and positively influence public attitudes toward HPV vaccination.

1.1 Research Motivation

A tremendous of studies have explored the use of advanced natural language processing (NLP) techniques to address the challenge of detection of misinformation on social media. This project is mainly inspired by the work of Zhang et al. [2] and Wang et al. [1].

Pre-trained language models (PLMs) are typically trained on large-scale, general-domain corpora, with only a few models explicitly designed for Twitter data. Most tweet-specified BERT-style language models, like Bert, BERTweet and RoBERTa, follow the same pre-training methodologies as general-domain PLMs and replace the training data with Tweets, and generally fail to capture social context information. To enhance language models with additional contextual information, some studies have incorporated metadata, as well as entities and relationships extracted from knowledge graphs, to enrich the pre-training corpus. However, these enhancements still fall short of fully capturing the nuanced social context in which tweets are embedded.

To address this gap, Zhang et al. [2] introduced TwHIN-BERT, a socially-enriched pre-trained language model specifically designed for multilingual tweet representations. During pre-training, TwHIN-BERT integrates both textual and social objectives and leverage the Twitter Heterogeneous Information Network (TwHIN). Compared with other PLMs, TwHIN-BERT has shown better performance in tasks such as sentiment analysis and misinformation detection, particularly in multilingual and socially contextualized settings.

Wang et al. [1] proposed a BERT-LSTM hybrid model to detect misinformation with temporal characteristics in mobile social networks. Their approach leverages capabilities of BERT in contextual understanding and the temporal sequence modeling of long short term memory (LSTM). The input to BERT-LSTM hybrid model consists of pre-processed text data, and the output is a binary classification indicating whether the text contains misinformation. Experimental results showed that the hybrid model outperforms traditional deep learning methods that rely solely on either PLMs or sequence models, by effectively capturing both linguistic and temporal dependencies.

By integrating the ideas from Zhang et al. [2] and Wang et al. [1], this project seeks to enhance the performance of misinformation detection on Twitter, with a focus on HPV vaccine-related content, by developing a hybrid TwHIN-BERT-LSTM model. Our model will leverage the advantages of both TwHIN-BERT and LSTM, and we hypothesize that the combination of these two models will lead to improved performance in identifying misinformation on Twitter, particularly in the context of HPV vaccine-related content.

1.2 Data Description

The primary dataset analyzed in this project is sourced from the VACCINELIES collection, which includes a large number of tweets related to both COVID-19 and HPV vaccines called CoVAXLIES and HPVAXLIES, respectively. For the purposes of this project, we concentrate on the annotated HPV vaccine-related tweets from the HPVAXLIES dataset. This dataset includes:

- *Tweet ID*: A unique identifier for each tweet, which can be used to retrieve the full text via Twitter's API.
- *Misinformation Targets (MisTs)*: Indicators of misinformation within the tweets concerning the HPV vaccine.
- *Stance Annotation*: Classification of the stance for a tweet towards the MisT, indicating whether the author *Accepts*, *Rejects*, or has *No Stance* towards the misinformation.
- *Taxonomy of MisTs*: A classification system for the MisTs, categorizing them into various themes and concerns to facilitate interpretation.

The dataset is divided into three collections: (1) a training collection (1860 tweets), (2) a development collection (189 tweets), and (3) a test collection (475 tweets). As described in [3], the training collection is used to train the MisT-evoking detection and stance identification system, the development collection is used to fine-tune model parameters, and the test collection is used to evaluate the model performance.

The stance of the authors towards the *MisTs* is not the central interets of this project. Instead, our main objective is to detect MisTs within the tweets. We will utilize the training collection to train our hybrid TwHIN-BERT-LSTM model for misinformation detection. The development collection will serve as a validation dataset, and the test collection will be used to assess the model performance.

2 Approach

2.1 Data Preprocessing

The full-text of tweets are not directly provided by VACCINELIES, but can be retrieved using the tweet IDs. We need to use Twitter's API to access and collect the full-text of English tweets from the training, validation, and test collections. The text data should be preprocessed to remove irrelevant information, such as URLs, hashtags, and mentions. All letters should be converted into the lowercase. Some tweets are empty because they have been deleted or are no longer available, so they are removed from the dataset.

Additionally, unlike the source data from HPVAXLIES, labels of tweets are converted into a binary format to facilitate model training, validation, and evaluation, with 1 indicating the presence of misinformation and 0 indicating the absence of misinformation.

Twitter rate limit is a problem when crawling information and retrieving tweets from Twitter's API via a free developer account. So far, only a small subset of the tweets (i.e., 133 tweets) have been collected from the training collection. Hence, experimental results demonstrated in this report are gained by training, validating, and testing models on the small subset of tweets. However, the issues caused by the Twitter rate limit will be addressed next week, and the full dataset from HPVAXLIES will be collected and used for training and evaluation for the final report.

Although the full dataset is not available at the moment, the small subset of tweets collected from the training collection (i.e., 133 tweets) will be used to show some initial results for the baseline models. Given the limited number of tweets, we randomly split the small subset of tweets into training and test sets with a ratio of 80% and 20%. The training set (i.e., 106 tweets) will be used to train the baseline models, and the test set (i.e., 27 tweets) will be used to evaluate the performance of the baseline models.

2.2 Methodology

In this project, we will evaluate TwHIN-BERT-LSTM on the HPVAXLIES dataset against the following baselines:

- *TwHIN-BERT-Misinformation-Classifier*: A fine-tuned version of Twitter/twhin-bert-large [2] for misinformation detection on unknown dataset.
- *TwHIN-BERT-Simple-Misinformation-Classifier* A fine-tuned version of Twitter/twhin-bert-large [2] without extra layer for misinformation detection on the HPVAXLIES dataset.

2.2.1 TwHIN-BERT-Misinformation-Classifier

The base model for TwHIN-BERT-Misinformation-Classifier is Twitter/twhin-bert-large, which has been introduced in Section 1. TwHIN-BERT-Misinformation-Classifier is fine-tuned on several existing datasets as presented in `https://huggingface.co/datasets/roupenminassian/twitter-misinformation`.

The training results are available in `https://huggingface.co/datasets/roupenminassian/twitter-misinformation`, and the following table shows the hyperparameters used for training the model.

It is pretty straightforward to load this classifier via the Hugging Face Transformers library and directly use this classifier to predict the misinformation in HPV vaccine-related tweets. The input to the classifier is a clean full text tweet, and the output is a binary classification (i.e., 0 or 1) indicating whether the tweet contains misinformation. The model is evaluated on the test dataset as described in Subsection 2.1. The performance of the classifier is evaluated using metrics such as accuracy, precision, recall, and F1 score.

2.3 TwHIN-BERT-Simple-Misinformation-Classifer

Instead of directly using fine-tuned TwHIN-BERT-Misinformation-Classifier from Hugging Face, we also can fine-tune the base model Twitter/twhin-bert-large on the HPVAXLIES dataset without

Hyperparameter	Value
learning rate	$2e^5$
train_batch_size	16
eval_batch_size	8
seed	42
optimizer	Adam with $\beta \in (0.9, 0.999)$ and $\epsilon = 1e^{-8}$
lr_scheduler_layer	linear
num_epochs	3

Table 1: Hyperparameters for TwHIN-BERT-Misinformation-Classifier as displayed in the Hugging Face.

adding extra layer for misinformation detection. In this project, this model is called TwHIN-BERT-Simple-Misinformation-Classifier.

After data preparation and pre-processing as described in Subsection 2.1, the training set is used to fine-tune the base model Twitter/twhin-bert-large. Here are summarized key steps for fine-tuning the base model.

First, we need to tokenize the clean full text of tweet using the AutoTokenizer from the Hugging Face Transformers library. This process involves several steps: (1) adding special tokens such as [CLS] at the beginning and [SEP] at the end of the tokenized text; (2) finding the length of longest tokenized tweet; (3) padding or truncating the tokenized text to ensure the uniform length of input data; (4) creating attention masks to distinguish between padding tokens and actual tokens; (5) embedding tokens into a high-dimensional vector space; (6) converting the tokenized text data into PyTorch tensors that serve as input for training models.

Second, dataloader is created to load the training set and validation set. It should be noted that we do not have the entire dataset at the moment, so we need to use K-fold cross-validation (i.e., 5-fold) to split the training set into training and validation datasets.

Third, pre-trained TwHIN-BERT is loaded from the Hugging Face Transformers library, and the number of labels is set to 2 (i.e., 0 or 1) for binary classification. The detailed hyperparameters for the TwHIN-BERT model are availed in paper [2]. The following table shows the hyperparameters used for fine-tuning the model:

Hyperparameter	Value
learning rate	$5e^{-5}$
train_batch_size	5
eval_batch_size	3
seed	42
optimizer	Adam with $\beta \in (0.9, 0.999)$ and $\epsilon = 1e^{-8}$
lr_scheduler_layer	linear
num_epochs	5

Table 2: Hyperparameters for fine-tuning TwHIN-BERT-Simple-Misinformation-Classifier on a small set of HPV-vaccine related tweets.

When fine-tuning the model, hyperparameters as listed in Table 2 are used to optimize the training process. In addition, when the full dataset is available, the model will be fine-tuned on the entire dataset and leveraging GPU for faster training.

Fourth, the training loop is defined regarding forward pass, loss calculation, and backward pass. The loss function used in this project is the binary cross-entropy loss. For 5-fold cross-validation, the model is trained on the training dataset and validated on the validation dataset.

Finally, model is saved as TwHIN-BERT-Simple-Misinformation-Classifier and loaded for future use. In this project, TwHIN-BERT-Simple-Misinformation-Classifier is evaluated on the test set as described in Subsection 2.1. The performance of the classifier is evaluated using the same metrics as TwHIN-BERT-Misinformation-Classifier.

2.3.1 TwHIN-BERT-LSTM

LSTM networks are designed to address the vanishing gradient problem, a common limitation in traditional recurrent neural networks (RNNs) [1]. It has been known that LSTMs employs a sophisticated gating mechanism that regulate the flow of information through network layers. Three types of gates—namely, the input gate, forget gate, and output gate—determine which information should be retained, updated, or discarded at each time step, allowing the model to capture long-range dependencies in sequential data.

The TwHIN-BERT-LSTM model is designed to harness the strengths of both TwHIN-BERT and LSTM, enabling more effective misinformation detection in tweets. The initial steps of constructing TwHIN-BERT-LSTM are identical to those used in the TwHIN-BERT-Simple-Misinformation-Classifier. First, the input data undergoes tokenization and conversion into PyTorch tensors using the AutoTokenizer from the Hugging Face Transformers library. Second, the training dataset is split into training and validation sets using K-fold cross-validation.

Ine the third step, the contextualized embeddings created from the pre-trained TwHIN-BERT model should be fed into the LSTM layer. Due to the advantages of LSTM, the output of the LSTM layer is a sequence of hidden states encoding the contextual information and temporal insights of the tweet text.

Finally, the hidden states are then passed through a fully connected layer with a sigmoid activation function to predict the presence of misinformation in the tweet. The binary cross-entropy loss function is used to calculate the loss, and the model is updated using backpropagation. The hyperparameters for the TwHIN-BERT-LSTM model are similar to those used for TwHIN-BERT-Simple-Misinformation-Classifier, as shown in Table 2.

3 Experiments

In this report, experimental results are obtained by training and evaluating the baseline models, TwHIN-BERT-Misinformation-Classifier and TwHIN-BERT-Simple-Misinformation-Classifer, on a small subset of HPV-vaccine related tweets. It is worth stressing that the issues of the Twitter rate limit and accessing data will be addressed next week, and the full dataset from HPVAXLIES will be collected and used for training and evaluation for the final report.

In this report, the small subset of tweets is randomly split into training and test sets with a ratio of 80% and 20%. For the baseline model TwHIN-BERT-Misinformation-Classifier, the model is directly loaded from the Hugging Face Transformers library and used to predict the misinformation in HPV vaccine-related tweets. The performance of the classifier is evaluated and illustrated in the following table:

Metric	Value
Accuracy	0.59259
Precision	0.61538
Recall	0.94118
F1 Score	0.74419

Table 3: Performance evaluation for TwHIN-BERT-Misinformation-Classifier.

For the baseline model TwHIN-BERT-Simple-Misinformation-Classifier, the model is fine-tuned on the training and validation dataset with 5-fold cross-validation. The loss of the classifier for epochs is calculated and illustrated in the following table:

Epoch	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
1	0.7163577816	0.7166936152	0.7152307034	0.6953794571	0.7209855423
2	0.7112625662	0.7060883746	0.7075976659	0.6978955058	0.7171074888
3	0.677092184	0.7121841382	0.705724909	0.7059453936	0.7062178675
4	0.7258692769	0.7085100973	0.7015493828	0.7154276792	0.7042601179
5	0.7166516535	0.745791593	0.7043769605	0.7209715177	0.7105007137

Table 4: Loss of TwHIN-BERT-Simple-Misinformation-Classifier for each epoch across 5 folds.

The performace of TwHIN-BERT-Simple-Misinformation-Classifier is evaluated on the test set, and the results are shown in the following table:

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Accuracy	0.5455	0.5238	0.5714	0.4762	0.3810	0.4996
Precision	0.5455	0.5238	0.5714	0.4762	0.3810	0.4996
Recall	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
F1 Score	0.7059	0.6875	0.7273	0.6452	0.5517	0.6635

Table 5: Performance of TwHIN-BERT-Simple-Misinformation-Classifier across 5 folds for each metric.

From Table 3 and Table 5, we can conclude that TwHIN-BERT-Misinformation-Classifier outperforms TwHIN-BERT-Simple-Misinformation-Classifier regarding accuracy, precision, recall, and F1 score. Furthermore, the performance of TwHIN-BERT-Simple-Misinformation-Classifier is not pretty good. The value of recall for each fold is 1 as shown in Table 5 indicating that the model is sensitive and predicts all the tweets as misinformation, which is not the case.

To be summarized, TwHIN-BERT-Simple-Misinformation-Classifier needs to be further optimized or tuned. As long as the full dataset of HPVAXLIES is available, the model will be fine-tuned on the entire dataset. Moreover, the distribution of the dataset will be checked to ensure that the model is not imbalanced and biased towards misinformation. An additional LSTM layer will be added to the fine-tuned TwHIN-BERT model to construct the TwHIN-BERT-LSTM model. The performance of TwHIN-BERT-LSTM will be evaluated on the test dataset obtained from HPVAXLIES, and the results will be illustrated in the final report. Last but not least, key mathematical formulas, algorithms, and architecture of models will be provided in the final report to help readers understand the details of the models and experiments.

References

[1] Wang, J., Wang, X. & Yu, A. (2025). Tackling misinformation in mobile social networks: A BERT-LSTM approach for enhancing digital literacy. Scientific Reports, 15(1118). https://doi.org/10.1038/s41598-025-85308-4.

[2] Zhang, X., Malkov, Y., Florez, O., Park, S., McWilliams, B., Han, J. & El-Kishky, A. (2022). TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations at Twitter. arXiv preprint arXiv:2209.07562.

[3] Weinzierl, M. & Harabagiu, S. (2022). VaccineLies: A natural language resource for learning to recognize misinformation about the COVID-19 and HPV vaccines. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 6967-6975). Marseille, France: European Language Resources Association. Available at [https://aclanthology.org/2022.lrec-1.753/].