Enhancing HPV Vaccine Misinformation Detection on Twitter: 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

1 Project Idea

The Human Papillomavirus (HPV) is the most prevalent sexually transmitted infection globally, and the HPV vaccine has been proven safe and highly effective in significantly reducing HPV-related cancers and other diseases. Despite its benefits, vaccine acceptance remains a challenge due to the widespread dissemination of misinformation, especially on social media platforms like Twitter. Identifying the types of misinformation being spread and understanding how users respond to it is crucial for public health. Such insights enable the development of targeted interventions to counter false narratives, reinforce accurate information, and positively influence attitudes toward the HPV vaccine.

Traditional approaches to misinformation detection, such as rule-based systems and machine learning techniques, face significant limitations, including limited availability of annotated vaccine misinformation datasets, poor scalability, lack of real-time adaptability, and difficulty in understanding context. Advances in natural language processing (NLP) and deep learning have provided solutions to many of these issues. This project aims to leverage these advancements by employing pre-trained language models like TwHIN-BERT—a multilingual model developed at Twitter—and long short-term memory (LSTM) networks to detect misinformation related to the HPV vaccine. TwHIN-BERT sets itself apart from other pre-trained language models by incorporating not only text-based self-supervision but also a social objective, leveraging the diverse social interactions within the Twitter heterogeneous information network (TwHIN). The LSTM network manages information using a series of gates that regulate its flow, functioning as filters to decide which information to retain and which to discard. The primary objectives are to effectively identify misinformation based solely on textual content, enhance digital literacy in an era of information overload, provide nuanced insights into misinformation dynamics, and capture the temporal characteristics of other biomedical specificity misinformation.

2 Software

- Programming Language: Python
- Libraries: spaCy for text preprocessing, tensorflow for model building, Hugging Face transformers for pre-trained models (TwHIN-BERT)
- Development Environment: Google Colab, Plato at UofS (possible??)
- Visualization Tools: Matplotlib, Seaborn
- Version Control: Git/GitHub

3 Literature Review

[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.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Table 1: Progress Report Milestone

Weeks	Tasks	Deliverables
1st	data collection & pre-processing	annotated tweets
2nd - 3rd	model architecture & baseline model development	trained models
4th	model evaluation & report progress	performance metrics
5th - 6th	model modification & evaluation	error analysis
7th - 8th	final reporting	comprehensive report & presentation

[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/].

4 Dataset Plan

Primary Source: a large set of annotated HPV vaccines-related tweets are available in VACCINELIES, including misinformation targets (MisTs), tweet IDs, annotation of the stance of each tweet author, a taxonomy of the MisTs.

Secondary Source: over 50 million filtered and pre-processed Saskatchewan tweets from December 2016 to December 22 from CEPHIL (Computational Epidemiology and Public Health Informatics Laboratory at University of Saskatchewan).

5 Evaluation Plan

Here are some possible quantitative metrics,

- F1-score
- precision
- recall
- AUC-ROC

Error analysis will identify model weaknesses in context or temporal pattern recognition.

6 Progress Report Milestone

7 Conclusion

In conclusion, this project tackles the critical challenge of detecting HPV vaccine-related misinformation on Twitter by leveraging developmemts in natural language processing and deep learning. The proposed TwHIN-BERT-LSTM hybrid model combines socially-enriched pre-training models with the temporal processing capabilities of LSTM to offer a powerful and efficient approach to identifying and analyzing misinformation. This work not only advances the detection of false medical knowledge and its spreading but also supports public health efforts by promoting accurate knowledge and enhancing digital literacy. Furthermore, the findings of this research provide a foundation for future studies addressing other biomedical misinformation, such as COVID-19 and influenza, as well as broader applications in social media analysis, including countering data poisoning attacks.