Health Fact or Fiction? A Comparison of BERT-Based Models and LLMs on Detecting Health Misinformation about COVID-19 and Measles

U. A. ALWANG, The University of Texas at Austin, USA

1 Introduction

In recent years, the spread of health misinformation online has become a major public health concern. The World Health Organization refers to the flood of health information online during a disease outbreak as an 'infodemic', an overabundance of information, accurate or not, that makes it difficult for people to find trusted sources and distinguish true and false claims [1]. As someone who frequently encounters health misinformation, I am interested in how we can use machine learning and AI to identify misinformation and limit its influence and reach.

Belief in health misinformation has deadly consequences; for example, researchers at Brown University School of Public Health have estimated that low vaccine rates and vaccine hesitancy during the COVID-19 pandemic resulted in 318,981 preventable deaths from January 2021 to April 2022 [9]. Furthermore, the emergence of measles in the United States after eradication in 2000 has been attributed to a targeted anti-vaccination campaign on social media that has in turn led to historically low childhood vaccination rates [7].

In my final project, I evaluate the performance of four different models at detecting health misinformation related to COVID-19 and measles. To see how the amount of model medical specialization affects performance, I use three BERT-based models with differing levels of medical knowledge and one LLM. I start with fine-tuning and evaluating all BERT models on the COVID-19 Rumors dataset [3], then I evaluate the LLM on the COVID-19 Rumors and Measles Rumors datasets, compare performance, and discuss model errors.

2 Related Work

There has been considerable research into using AI for automated fact-checking across different information domains. Two papers that inspired this final project are: HealthLies: Dataset and Machine Learning Models for Detecting Fake Health News [2] and ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection [6]. Both papers present a new health claims dataset and then train and evaluate several machine learning models on that dataset. HealthLies contains claims about several different diseases such as cancer, polio, HIV/AIDS, and SARS, while ANTi-Vax only contains claims related to COVID-19 vaccines. BERT-based models achieved the best results in both papers. On ANTi-Vax, BERT had 0.97 precision and 0.98 recall scores, and on HealthLies, it had 0.997 precision and 0.998 recall scores. These results indicate the strength of BERT as a base model for health misinformation detection. However, HeathLies also evaluated its best-performing BERT model on claims about diabetes, a disease outside the training set, and found that performance suffered considerably, suggesting that BERT-based models do not generalize well.

3 Methodology

3.1 Model Selection

Following related work showing that BERT-based models can accurately detect health misinformation, I selected three different BERT models to fine-tune, each with a different degree of medical specialization:

[1] https://www.who.int/health-topics/infodemic#tab=tab_1

Author's Contact Information: U. A. Alwang, ursulaalwang@gmail.com, The University of Texas at Austin, Austin, Texas, USA.

(1) **BERT Base Uncased**: the original BERT model that was trained self-supervised on a large corpus of english data. This model has no medical specific training or instruction [4].
(2) **BiomedBERT**: an extension of BERT, trained on abstracts from PubMed and full text articles from PubMedCentral [5].
(3) **BioClinialBERT**: An extension of BiomedBERT, trained on all clinical notes from MIMIC-III [1].

For the LLM, I selected Gemini 2.0 Flash-Lite [2], which is the cost-optimized version of Google's most advanced Gemini model. Even though this model is not specialized for medical use, I was curious to see how a general purpose LLM would perform, since these models are likely to be used by the general public.

3.2 Dataset Selection and Creation

For this project, I wanted to test model performance on data from both established and emerging public health events. COVID-19 Rumors is a publicly available dataset consisting of 7,179 annotated claims about COVID-19 from Google and X (formerly Twitter) from January 2020 to March 2020 that were labeled by human experts as true, false, or unverified (Fig.1). I selected the COVID-19 Rumors dataset because it has both high-quality expert-generated labels and low keyword predictability, meaning that the model has a lower likelihood of relying on certain keywords to detect misinformation in the claim [8]. Additionally, since this dataset contains claims from the wider internet and not just X, I argue that it's a more accurate representation of the variety of health misinformation people encounter online. The dataset has a distribution of 26.16% true, 51.27% false, and 22.57% unverified claims.

Claim: Tea can cure or alleviate novel coronavirus (COVID-19) infection **Label**: False

Claim: The U.S. Food and Drug Administration has authorized the use of two tests, one from the Centers for Disease Control and Prevention and one from the New York State Department of Public Health. Neither agency is currently charging patients for the test **Label**: True

Claim: Viral particles released during a sneeze can reach 10 feet **Label**: Unverified

Fig. 1. Examples of true, false, and unverified claims from the COVID-19 Rumors Dataset.

For an emerging public health event, I created a small dataset of 20 claims about measles. This dataset has 10 false claims, 7 true claims, and 3 unverified claims. All claims were sourced from real news articles and social media posts from April 2025 and were labeled by a medical doctor.

For ease of training and evaluation, I converted the false, true, and unverified labels to integers 0, 1, and 2 respectively.

3.3 Fine Tuning the BERT Models

Each BERT Model was fine-tuned on the COVID-19 Rumors training split using the HuggingFace Trainer API [3] which supports distributed training pipelines on multiple GPUs. All three models were trained with the same hyperparameters and for the same number of epochs to avoid differences in training techniques influencing the results.

[2] https://deepmind.google/technologies/gemini/flash-lite/
[3] https://huggingface.co/docs/transformers/en/main_classes/trainer

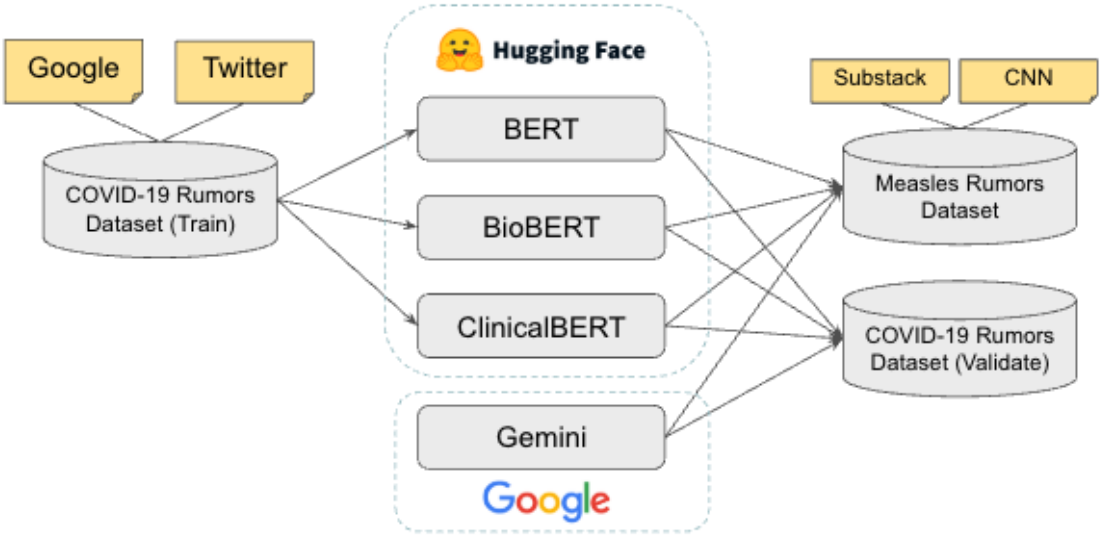

Fig. 2. Flow of training and evaluation data through all tested models

3.4 Evaluating Performance of the BERT Models

After fine-tuning, each BERT model was evaluated on the COVID-19 Rumors evaluation split and the Measles Rumors dataset, measuring loss, accuracy, precision, recall, and f1. I compared the results of all models on both datasets, including doing an in-depth examination of model errors to better understand model performance.

3.5 Evaluating Performance of Gemini 2.0 Flash-Lite

```
"You are an AI Medical Assistant trained on a vast dataset of health information. Please
evaluate the provided claim and respond with the following determination:
0 - The claim is false
1 - The claim is true
2 - I am unable to make a determination

Please only respond with a 0, 1, or 2. Do not include any other text."
```

Fig. 3. The system instructions given to Gemini.

To test the performance of Gemini, I used the Gemini API through the Google genai library [4]. I passed a prompt (Fig.3) containing information about the task as a 'system instruction' to Gemini which gives the model additional context and allows it to generate more customized responses.

[4] https://pypi.org/project/google-genai/

4 Results

4.1 Key Results

BERT achieved the highest performance on COVID-19 Rumors, with 0.86 accuracy and F1, followed closely by Biomed-BERT, and BioClinicalBERT. Gemini performed the worst on COVID-19 Rumors, with 0.60 accuracy and 0.65 F1. All three BERT-based models suffered heavy performance losses on Measles Rumors, but Gemini improved significantly to 0.84 accuracy and 0.79 F1.

These results are interesting in two respects: first, increased medical knowledge of the baseline BERT model did not improve performance in detecting medical misinformation. If anything, it seemed to slightly worsen performance. BioClinicalBERT, which has the most medical knowledge of all the BERT models tested, had the lowest performance and BERT, which had the least medical knowledge, had the highest performance. Intuitively, you would expect performance to increase with medical knowledge.

Second, none of the BERT models generalized to Measles Rumors, however, Gemini had the highest performance on this dataset by far, matching the performances of the fine-tuned BERT models on COVID-19 Rumors. Gemini also had the worst performance on COVID-19 Rumors by a wide margin. This seems to suggest that while Gemini is a good general-purpose model, it struggles with specialized tasks.

	BERT		BiomedBERT		BioClinicalBERT		Gemini	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
COVID-19 Rumors	0.86	0.86	0.84	0.83	0.82	0.80	0.60	0.64
Measles Rumors	0.52	0.44	0.57	0.56	0.57	0.46	0.84	0.79

Table 1. Performance Comparison of BERT, BiomedBERT, BioClinicalBERT, and Gemini on COVID-19 Rumors and Measles Rumors

4.2 Analysis of Errors Across Models

To better understand why increased medical knowledge did not improve performance on COVID-19 Rumors, I did further analysis on the type of errors made by each model. In general, the BERT-based models over-predicted false claims, but this problem was significantly worse for the models with medical specialization. 99% of errors made by BioClinicalBERT and 63% of errors made by BiomedBERT were on predicted false claims compared to just 47% of errors made by BERT. Increased medical specialization may make the model more cautious when it encounters claims that are health *related* but not health-specific. Many of the errors made by BiomedBERT and BioClinicalBERT involved evaluating political or social information related to public health, for example: "Vice President Mike Pence said that “the FDA [Food and Drug Administration] is approving off-label use for the hydroxychloroquine right now". A model specialized for medical use may focus on whether or not hydroxychloroquine is an effective treatment for Covid, and may overlook that the veracity of the claim relies on whether or not the FDA approved off-label use of the drug.

Gemini had the opposite problem, with 68% of errors on false claims, split almost evenly between predicting true and unverified. An analysis of errors made by Gemini reveals that the LLM has a significant recency bias. For example, the claim "A vaccine for coronavirus has been found" was false when it was written in 2020, but is now true. Likewise the claim "it reports that Brazilian president Jair Bolsonaro has tested positive for the coronavirus" was false when it was written in 2020, but Bolsonaro did test positive eventually, leading the LLM to predict true. This is a significant limitation in evaluating the effectiveness of LLMs at retrospectively identifying misinformation and helps explain why

Gemini was better at evaluating measles claims. While one can prompt the LLM to only use information prior to the date of the claim, there is no way to guarantee that it follows this guideline. Since the training datasets of large LLMs are not publicly available, we don't know if the LLM has the historical context needed to make that determination.

5 Financial Transparency and Reproducibility

All training and evaluation was done on Google Colab[5] with an A100 high RAM GPU accessed through a Google Colab Pro monthly subscription for $9.99[6]. Google Gemini was accessed through the Google Gemini API free tier. All linked datasets, libraries, and models used for this final project are publicly available, and all code used to generate any chart or example used in this paper is included in the submission file.

5.1 Conclusion

In my final project, I compared the performance of three BERT-based models with different levels of medical specialization and one LLM at identifying COVID-19 and measles misinformation. I started with fine-tuning BERT, BiomedBERT, and BioClinicalBERT on the COVID-19 Rumors dataset. Then, I evaluated all three BERT models and Gemini on the COVID-19 Rumors dataset in addition to my own Measles Rumors dataset, created from 20 claims from various internet sources about measles and labeled by a medical doctor. Overall, while all three BERT-based models achieved strong results, medical specialization did not improve model performance, with BERT achieving the highest performance on COVID-19 Rumors at 0.86 accuracy. Possibly due to recency bias, Gemini struggled on COVID-19 Rumors with an accuracy score of 0.64. All BERT-based models did not generalize to detecting measles misinformation, but Gemini achieved high performance with 0.84 accuracy. These results suggest that for identifying health misinformation, the composition of the fine-tuning dataset has a larger influence on BERT performance than the model specialization. Additionally, it highlights the limitations of using LLMs for retrospective health misinformation identification.

Acknowledgments

Thank you to Dr. L. Tracey for her help labeling the Measles Rumors dataset and to Professor Ying and the TAs for their care and attention this semester.

References

[1] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly Available Clinical BERT Embeddings. arXiv:1904.03323 [cs.CL] https://arxiv.org/abs/1904.03323

[2] Garima Chaphekar and Jorjeta G. Jetcheva. 2022. HealthLies: Dataset and Machine Learning Models for Detecting Fake Health News. In *2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService)*. 1–8. doi:10.1109/BigDataService55688.2022.00008

[3] Mingxi Cheng, Songli Wang, Xiaofeng Yan, Tianqi Yang, Wenshuo Wang, Zehao Huang, Xiongye Xiao, Shahin Nazarian, and Paul Bogdan. 2021. A COVID-19 Rumor Dataset. *Frontiers in Psychology* Volume 12 - 2021 (2021). doi:10.3389/fpsyg.2021.644801

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[5] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. arXiv:arXiv:2007.15779

[6] K Hayawi, S Shahriar, M. A. Serhani, I Taleb, and S. S. Mathew. 2022. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health* 203 (Feb 2022), 23–30. doi:10.1016/j.puhe.2021.11.022 arXiv:2021 Dec 7

[7] Benecke O and DeYoung SE. 2019. Anti-Vaccine Decision-Making and Measles Resurgence in the United States. *Glob Pediatr Health* 6 (2019), 2333794X19862949. doi:10.1177/2333794X19862949

[5] https://colab.research.google.com/
[6] https://colab.research.google.com/signup

[8] Camille Thibault, Jacob-Junqi Tian, Gabrielle Peloquin-Skulski, Taylor Lynn Curtis, James Zhou, Florence Laflamme, Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. 2025. A Guide to Misinformation Detection Data and Evaluation. arXiv:2411.05060 [cs.SI] https://arxiv.org/abs/2411.05060

[9] Ming Zhong, Tamara Glazer, Meghana Kshirsagar, Richard Johnston, Rahul Dodhia, Allen Kim, Divya Michael, Santiago Salcido, Sameer Nair-Desai, Thomas C. Tsai, Stefanie Friedhoff, William B Weeks, and Juan M. Lavista. 2023. Estimating Vaccine-Preventable COVID-19 Deaths Among Adults Under Counterfactual Vaccination Scenarios in The United States: A Modeling Study Using Observational Data. *Journal of Pharmacy and Pharmacology Research* 7 (2023), 163–16.