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

Ursula Alwang, AI in Healthcare High Risk Project
Spring 2025

Report, Code and Datasets (1 & 2)

Introduction

What is health misinformation and why is it dangerous?

- The Surgeon General defines health misinformation as “information that is false, inaccurate, or misleading according to the best available evidence at the time” (source)
- During a disease outbreak such as COVID-19 or Ebola, the flood of health information online makes it difficult for people to find trusted sources and identify misinformation
- **Belief in health misinformation is dangerous and can be deadly:** misinformation erodes trust in science and can lead people to delay or even refuse safe and effective medical care when they otherwise would not have
- **We are all vulnerable to misinformation:** willingness to believe misinformation often comes from a desire to better understand a chaotic or confusing situation. Misinformation fills the void left by a lack of trustworthy and easily accessible reliable information sources.

Goal of my high risk project

- The goal of my high risk project was to explore detecting health misinformation about COVID-19 and measles with different types of machine learning models
- Fine-tuned BERT-Based models have performed extremely well at this task in related work.
- I was curious to see if I could replicate these results or improve them using BERT-based models specialized for medical use such as BiomedBERT and BioClinicalBERT
- I was also curious to see how an LLM would perform at this task. Even though they don't have medical-specific training, LLMs have extremely powerful reasoning and large generalized knowledge bases. They are also accessible to the public, meaning that they might be the first line of defense against medical misinformation for the average person

Methodology

Dataset selection and creation

- For this project, I wanted to test model performance on data from both established and emerging public health events.
- To represent an established public health event, I selected the **COVID-19 Rumors dataset** which consists of 7,179 annotated claims about COVID-19 from January 2020 to March 2020 that were labeled as true, false, or unverified
- To represent an emerging public health event I created a small dataset of 20 claims related to measles which I call the **Measles Rumors dataset**

Example claims and labels:

Tea can cure or alleviate novel coronavirus (COVID-19) infection [False]

Viral particles released during a sneeze can reach 10 feet [Unverified]

The most effective way to prevent the spread of measles is the MMR vaccine [True]

MMR vaccine drives fuels measles outbreak [False]

Fine tuning the BERT models

- Next, I fine-tuned three BERT models on the COVID-19 Rumors dataset using the huggingface trainer API
- All models were trained with the same hyperparameters and for the same number of epochs to avoid differences in training affecting results

Overview of BERT Models used:

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
2. **BiomedBERT**: an extension of BERT, trained on abstracts from PubMed and full text articles from PubMedCentral
3. **BioClinialBERT:** An extension of BiomedBERT, trained on all clinical notes from MIMIC-III

Evaluating performance

- To test the performance of Gemini, I used the Gemini API through the google genai library and passed system instructions as well as the claim to gemini for evaluation
- Every model was evaluated on the COVID-19 Rumors evaluation split and the Measles Rumors dataset, measuring loss, accuracy, precision, recall, and f1
- I also used confusion matrices to better understand the types of errors each model was making

System instructions given to Gemini:

```
"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."
```

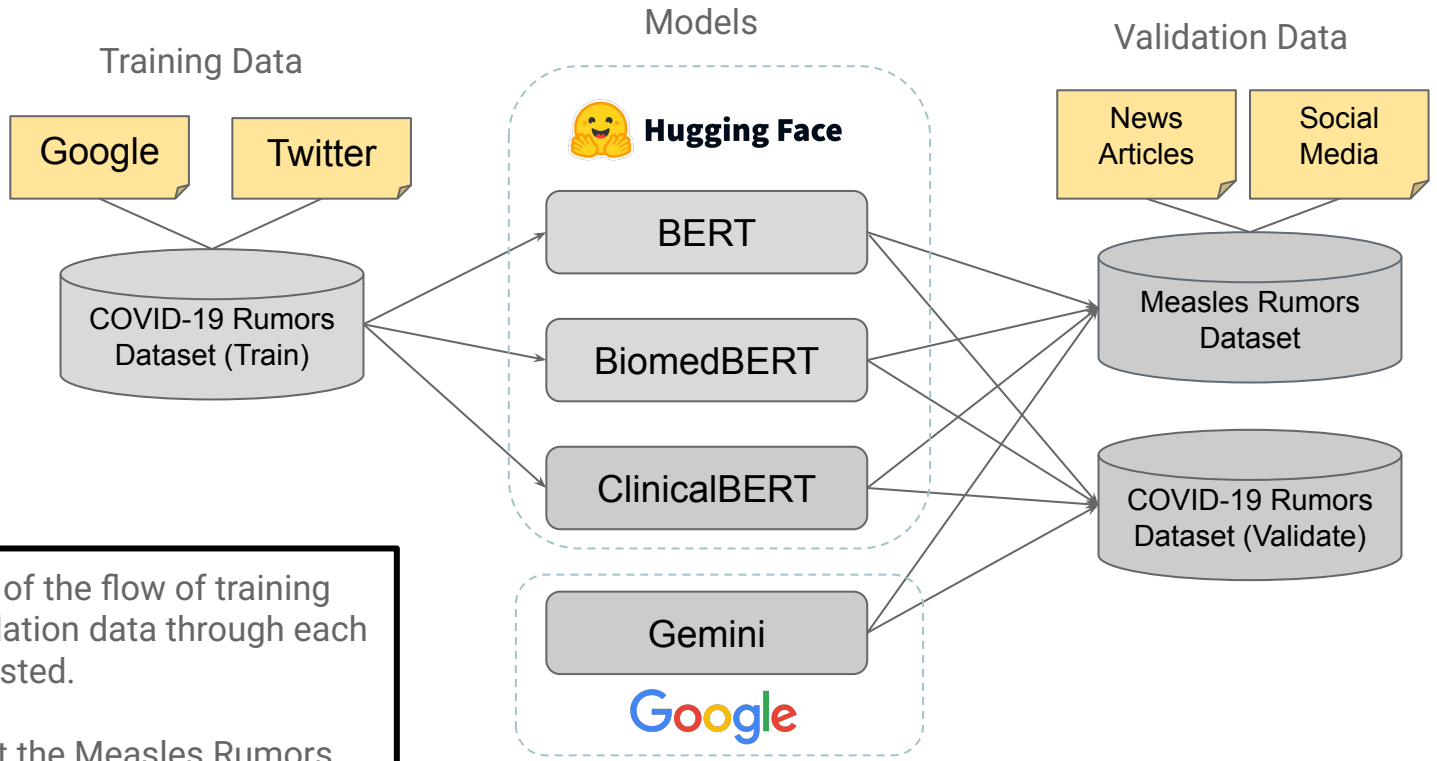

Diagram of the flow of training and validation data through each model tested.

Note that the Measles Rumors dataset was only used for evaluation and that Gemini was *not* fine-tuned on COVID-19 Rumors.

Results

Key Results

	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

- BERT achieved the highest performance on COVID-19 Rumors, with 0.86 accuracy and F1, followed closely by BiomedBERT, 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
- I think these results are interesting because intuitively, you would expect performance to increase with medical knowledge
- As seen in related work, BERT-based models did not generalize well to new datasets

Discussion of Results

Question: why didn't medical specialization improve performance?

Response: 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. Further analysis of errors reveals that BioClincialBERT and BiomedBERT significantly over-predicted false. Predicted false claims accounted for for 99% and 63% of errors made by BioClinicalBERT and BiomedBERT respectively, compared to just 47% of errors made by BERT.

Question: why did Gemini perform so poorly on COVID-19 Rumors but so well on Measles Rumors?

Response: Almost all Gemini errors were on false claims, split evenly between predicting true and unverified. Further analysis of errors reveals that Gemini has a significant recency bias and incorrectly rated many claims that were false from January to March 2020 but are now true, such as "A vaccine for coronavirus has been found" or "Brazilian president Jair Bolsonaro has tested positive for the coronavirus". Gemini performed better on Measles Rumors because it's using the same frame of reference as the dataset.

Future Directions

Further Research

Dataset composition

- My findings suggest that for identifying health misinformation with BERT, the composition of the fine-tuning dataset has a larger influence on performance than the model specialization
- Since identifying health misinformation often involves evaluating political or social claims, further research can explore how the balance of health, political, and social claims in the dataset affects performance

Retrospective Analysis with LLMs

- My project highlights the limitations and challenges of using LLMs for retrospective misinformation detection
- While it's possible to prompt the LLM to only use information from a certain time frame, there is no mechanism to limit the data it uses to generate an answer
- Further research is needed into the feasibility of 'time-framing' LLMs and the suitability of LLMs for retrospective analysis

Thank you!