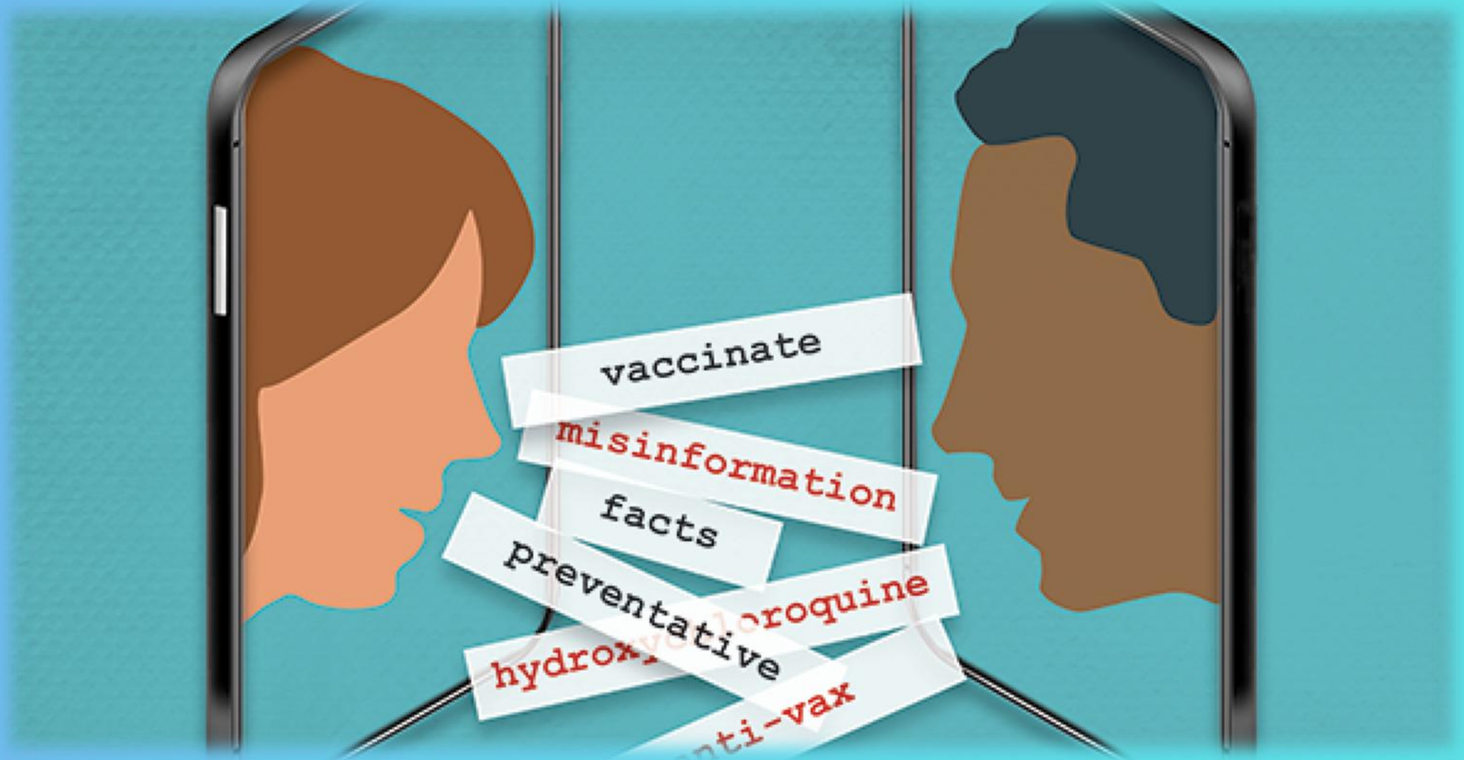
MedFakeDetect:
Medical Misinformation Detection

Student: Sara Mangistu



PROBLEM STATEMENT

Motivation

- Health-related misinformation spreads rapidly on social media distorting public understanding of medical facts.

  False or misleading claims about treatments or disease prevention often appear persuasive yet lack scientific basis. Such content undermines trust in healthcare, contributes to delayed or inappropriate treatment, and poses a real risk to individual and public health.
- **Goal:** Automatically detect and classify medical claims from online platforms as either true or false, supporting efforts to reduce public exposure to harmful misinformation.

Problem Definition

- Input: Social media post/claim
- Output: classification of the post/claim as real or fake

NLP Tasks

- Binary text classification

Why is it difficult?

- Diverse formats (short tweets vs. formal claims)
- Complex, noisy text (slang, hashtags, emotions) varies across platforms.
- claims can be partially true, misleading, or lacking context
- Requires contextual medical understanding

TRAINING AND TEST DATA

Data type and labels

- Labeled text samples: true/false labels

Datasets

- COVID19 Fake News Dataset NLP (Kaggle,🔗). social-media platforms (Twitter, Facebook, Instagram…)
- PUBHEALTH-DATASET (Kaggle,🔗). Claims related to a range of health topics
- Misinformation-Detection (Github,🔗). Detect Health Misinformation

- All datasets filtered to retain only binary-labeled examples.
- Expected dataset size (TBD): (20,007 , 2)

Post/claim	label
Lemons Kill Cancer Cells Better Than Chemotherapy	Fake
Tai Chi Reported to Ease Fibromyalgia	Real
The coronavirus outbreak is caused by 5G technology	Fake

EVALUATION

Evaluation strategy

- Train/Test split: 80% train, 20% test
- Separate evaluation per dataset. Compare model performance across domains

Evaluation Metrics

- Accuracy (Overall performance)
- Precision/Recall/F1-score
- Confusion Matrix (Error analysis)

Models

- **Baseline**: Naïve Bayes / Logistic Regression
- **Advanced**: Fine-tuned BERT
- **Domain-specific**: BioBERT