



MedTruth



Detecting Medical Misinformation On Social Media



NLP Project - Final presentation

Sara Mangistu

Michelle Zalevsky





INTRODUCTION

Problem Description & Project Objectives

Health-related misinformation spreads rapidly on social media, distorting public understanding of medical facts. This can: Undermine trust in healthcare, Lead to delayed or inappropriate treatments and pose real risks to individuals and public health

Challenges in Classification:

- Vague or emotionally charged language
- Pseudo-medical logic mimicking scientific tone
- Short, inconsistent, and noisy text content

Objectives

- Fine-tune and compare transformer-based models: BERT, BioBERT, RoBERTa
- Perform binary classification of health-related claims (true/false)
- Incorporate synthetic data augmentation to improve robustness against noisy, real-world text styles

Task Specification

Formal task specification

- **Input:** textual health-related social media claim
- **Output:** Binary label — True/False
- **Metrics:** Accuracy, Precision, Recall, F1-score, Confusion Matrix

Subtasks

- Data preparation: Merge and preprocess multiple labeled datasets
- Data generation: Synthesize 4000 diverse false health claims via GPT-4-Turbo.
- Model training: Fine-tune transformer-based classifiers (BERT, BioBERT, and RoBERTa).
- Evaluation: Compare performance across models

PRIOR ART

Source/Title	“Fact or Fiction: Verifying Scientific Claims” 2020	“Evidence-based Fact-Checking of Health-related Claims” 2021	“COVIDLIES: Detecting COVID-19 Misinformation on Social Media” 2020
Task solved	classification task: Classify scientific claims as Supported / Refuted / No Info using retrieved scientific abstracts	classification task: Classify online health claims as Support / Refute / Neutral based on evidence from scientific articles	classification task: Determine stance of tweets (Agree / Disagree / No stance) toward known COVID-19 misinformation myths
Approach/Model	RoBERTa-based Natural Language Inference (NLI) model to verify claims using retrieved abstracts	Used pre-trained transformers (BERT, SciBERT, BioBERT, T5) on claim-evidence pairs. T5 yielded best performance.	Two-stage: (1) Tweet-misconception retrieval (TF-IDF/BERTSCORE), (2) Stance classification (Agree/Disagree/No Stance) using NLI models (e.g., SBERT).
Data	1,409 scientific claims + 5,183 abstracts from S2ORC corpus (PubMed-style papers)	1,855 real-world COVID-19 claims, 738 scientific evidence passages → 14,330 annotated claim-evidence pairs.	6,761 tweet-misconception pairs labeled across 86 COVID-related claims (Agree / Disagree / No Stance).
Metrics	Accuracy, Precision/Recall, F1 (claim verification), Evidence selection F1	Accuracy, Precision, Recall, Macro F1-score for 3-way classification (SUPPORT, REFUTE, NEUTRAL).	Retrieval: Hits@1, Hits@5, MRR. Classification: Precision, Recall, F1 (per class and macro).
Results	~85% accuracy on claim verification, ~68% F1 for evidence sentence selection	T5 achieved F1-score 79.6% and accuracy 80.7%. BERT-based models (e.g., SciBERT, BioBERT) slightly lower.	Macro F1 = 50.2%, with F1 = 41.2% for Agree and 89% for No Stance. Retrieval improved from ~38% to 61.3% Hits@1 with domain adaptation.



METHODOLOGY

Data Preparation

Datasets

- PubHealth Dataset (Kaggle, [link](#))
- COVID19 Fake News Dataset (Kaggle, [link](#))
- HLR Medical Misinformation (GitHub, [link](#))
- Synthetic generation using GPT-4-Turbo

All datasets filtered to contain only a textual medical claim and a binary label (True/False).

Synthetic generation

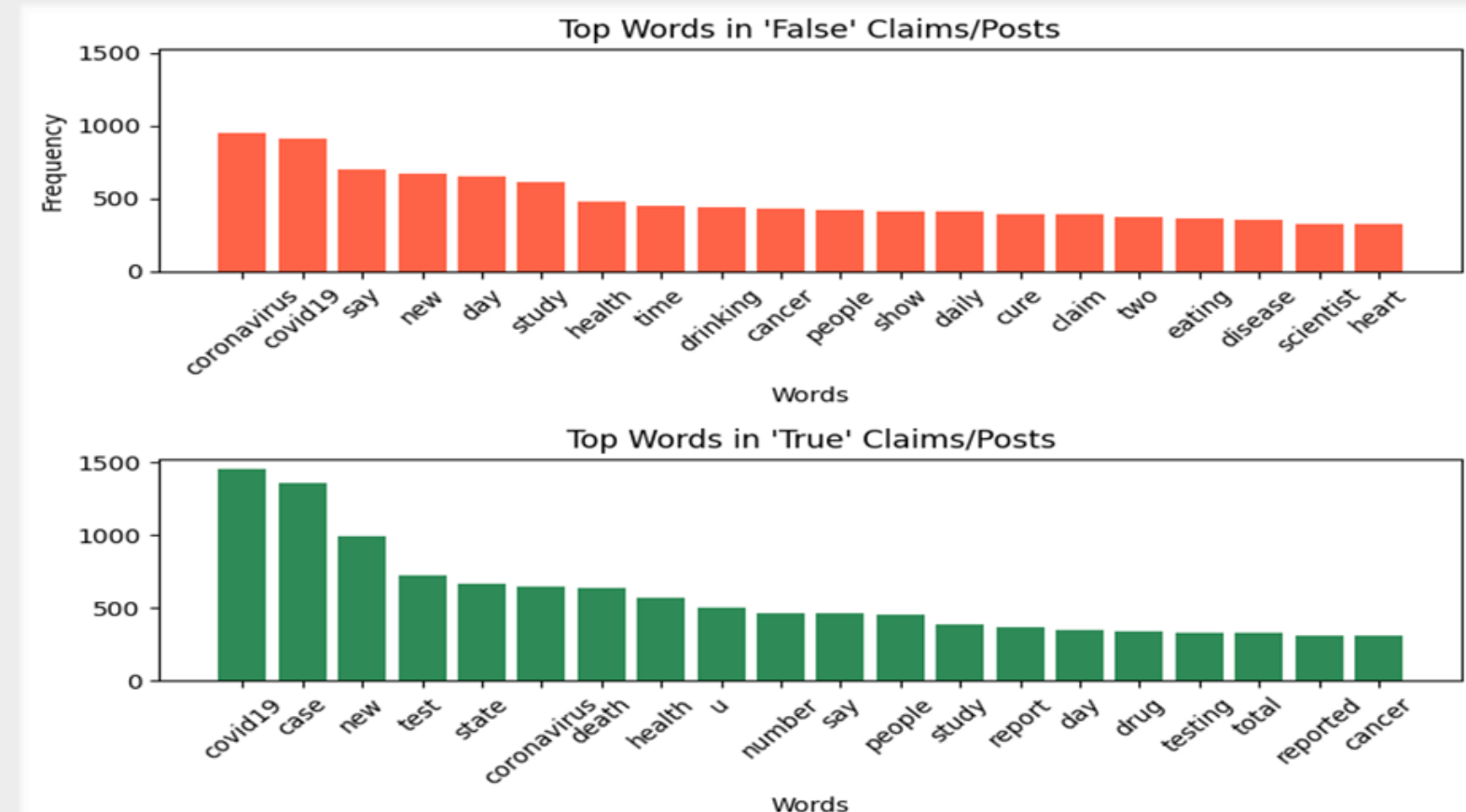
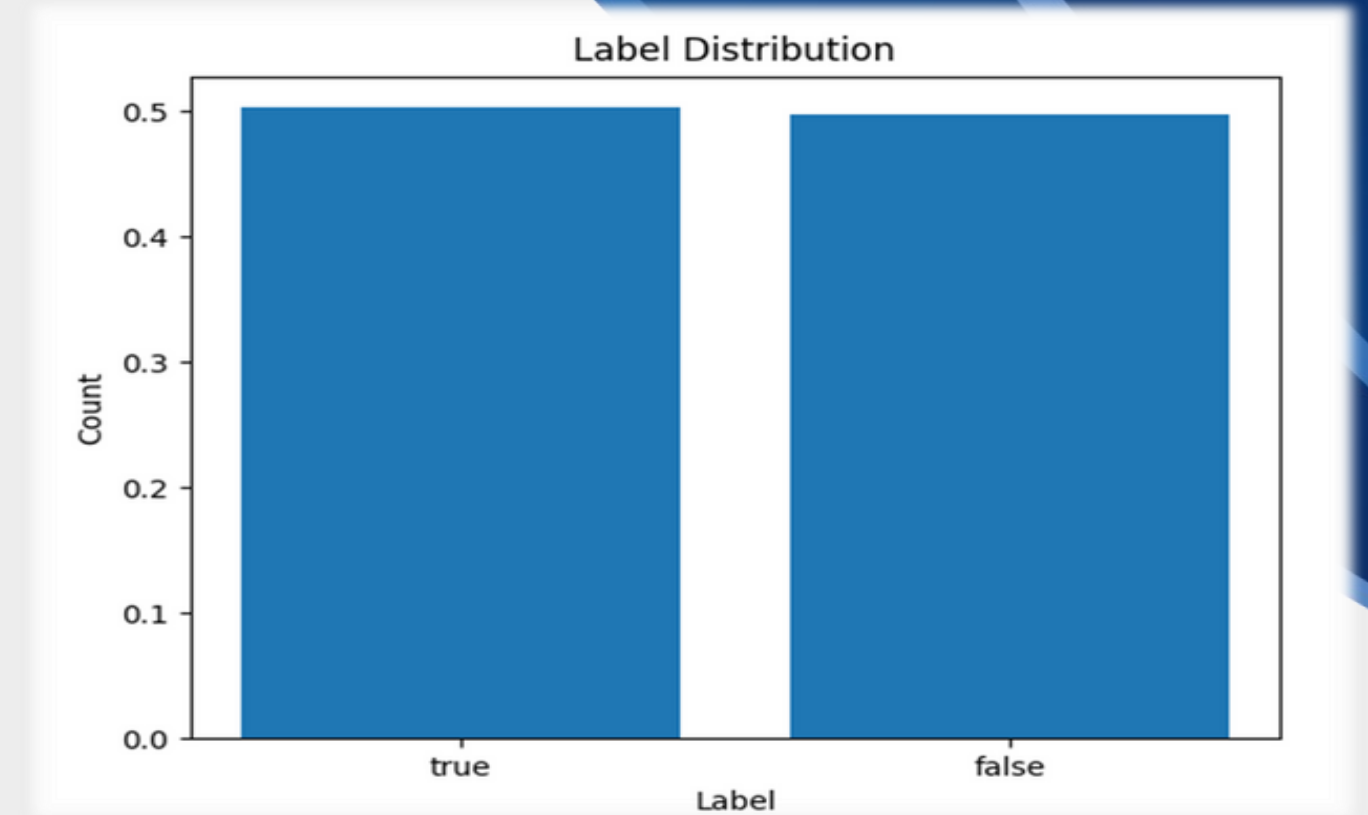
~2,300 GPT-4-Turbo-generated false claims.

Based on 4 real COVID examples used as style prompts.

evaluated on realism, deceptiveness, relevance, topical & expression diversity

Data properties / EDA

- Total ~16,000 claims
- Balanced (approx. 50% true / 50% false)
- Average length after cleaning: 12.14



Models & processing pipelines

Models

- Baseline: Logistic Regression, Naïve Bayes
- Advanced: BERT, BioBERT, RoBERTa.

Training Details

- Data split: 80/20 (train/test).
- Platform: Google Colab (CPU + GPU mix)
- Tokenizers: TF-IDF/AutoTokenizer per model

Model	Pretrained On	Cased?	Domain Aware?	Max Length	Epochs	LR	Batch Size
BERT	Books + Wikipedia	No	No	256	3	0.00002	16
RoBERTa	CommonCrawl	Yes	No	256	3	0.00002	16
BioBERT	PubMed + PMC	Yes	Yes	256	3	0.00002	16

Metrics & Evaluation

Evaluation Metrics

- Accuracy: Overall percentage of correctly classified claims.
- Precision & Recall: Computed for the "false" (misinformation) and "true" classes.
- F1 Score: mean of precision and recall, balancing false positives and false negatives.

Training:

Metrics (Accuracy, Precision, Recall, F1) computed per epoch on the validation set using Hugging Face's Trainer API

Testing:

- Final model evaluation on test set using `trainer.predict()` output.
- Extracted true vs. predicted labels (`y_true` / `y_pred`).
- Used scikit-learn for:
 - Precision, Recall, F1 (per label)
 - Confusion Matrix plots (custom function)



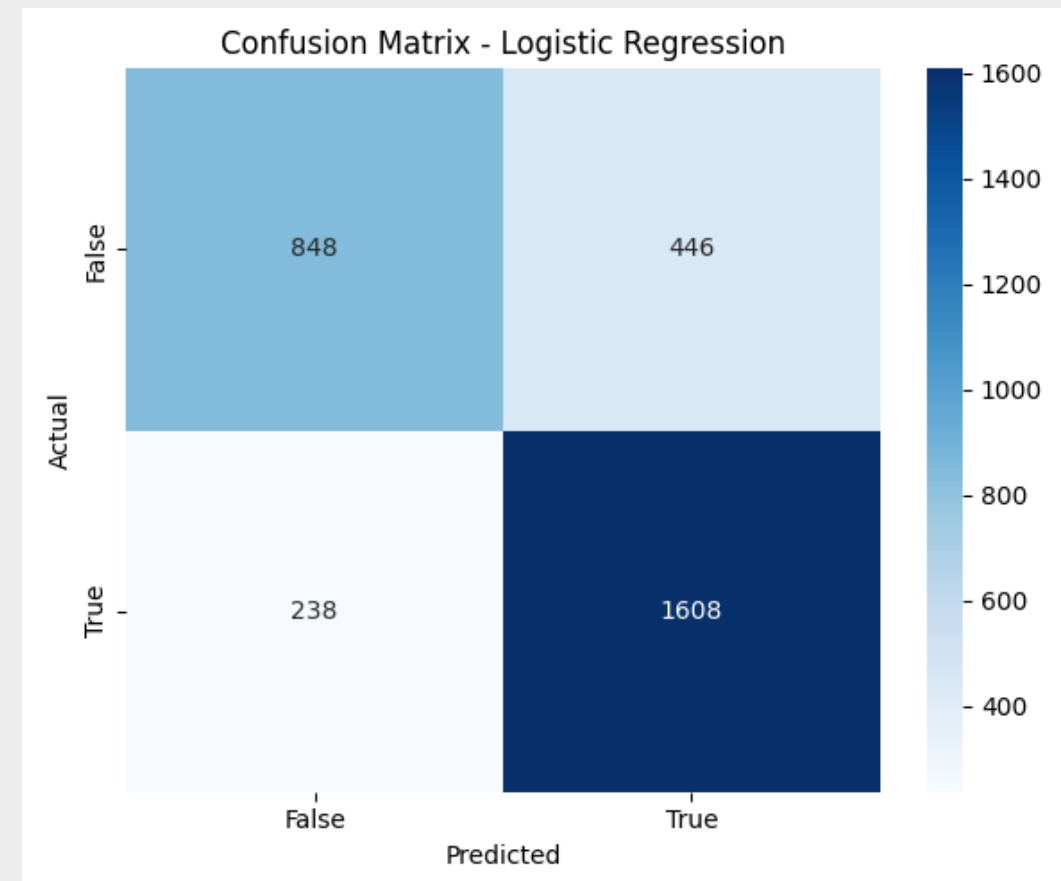
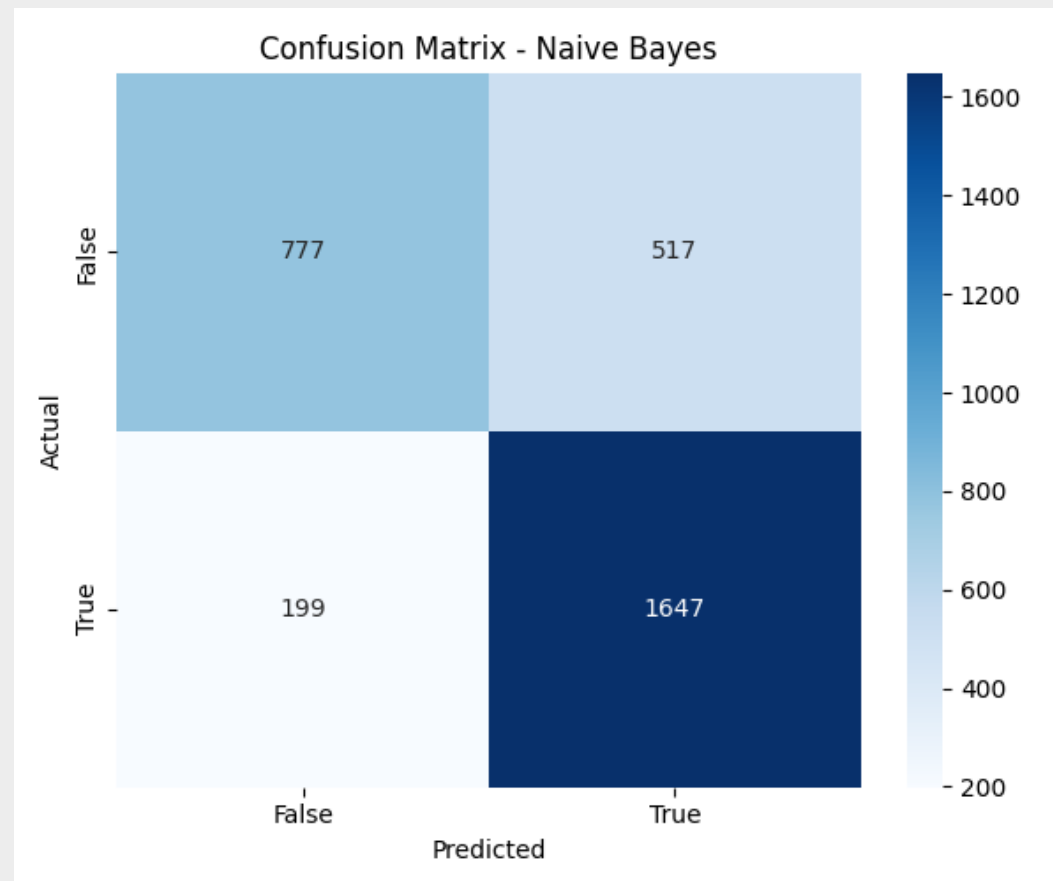
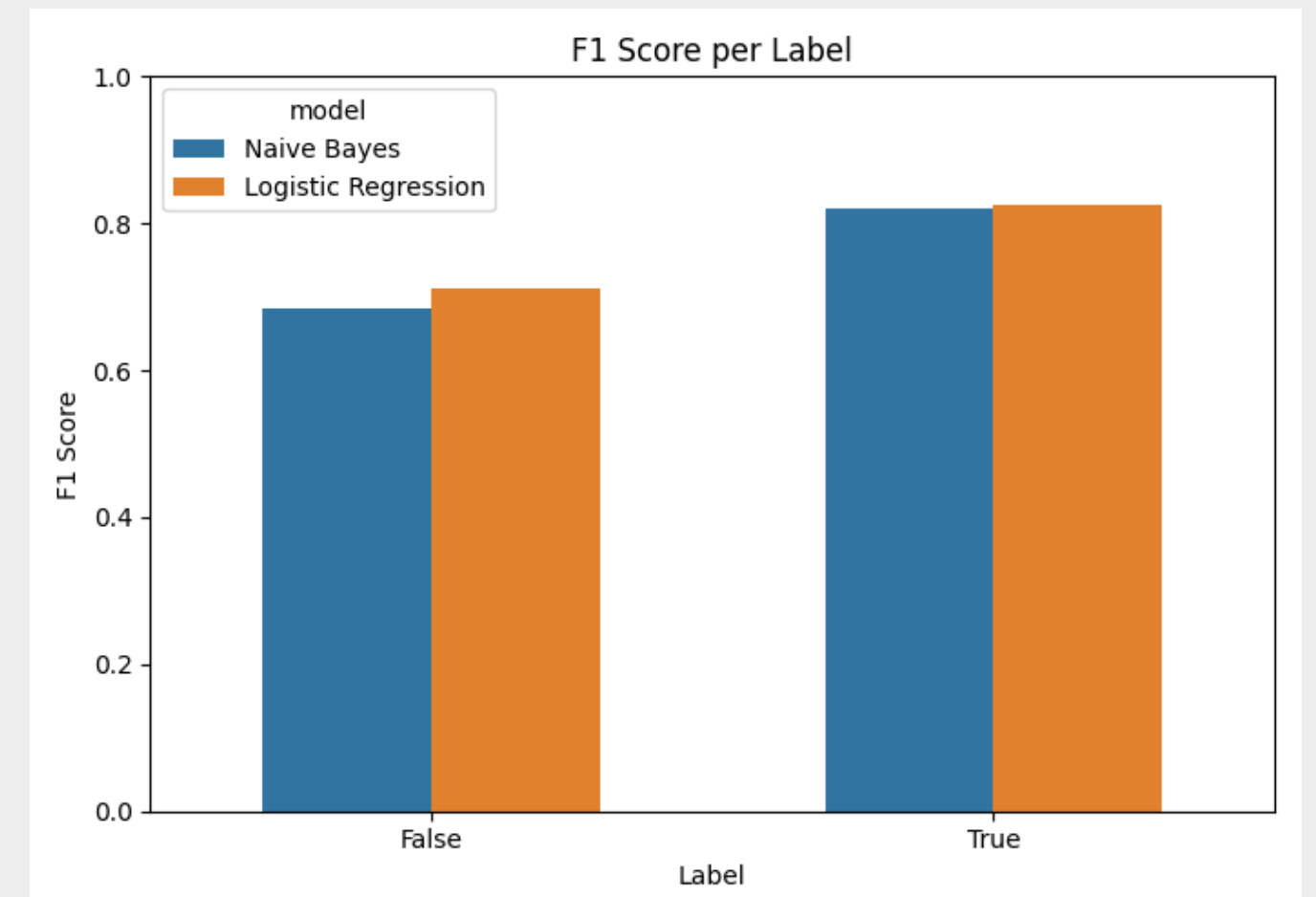
RESULTS



Baseline

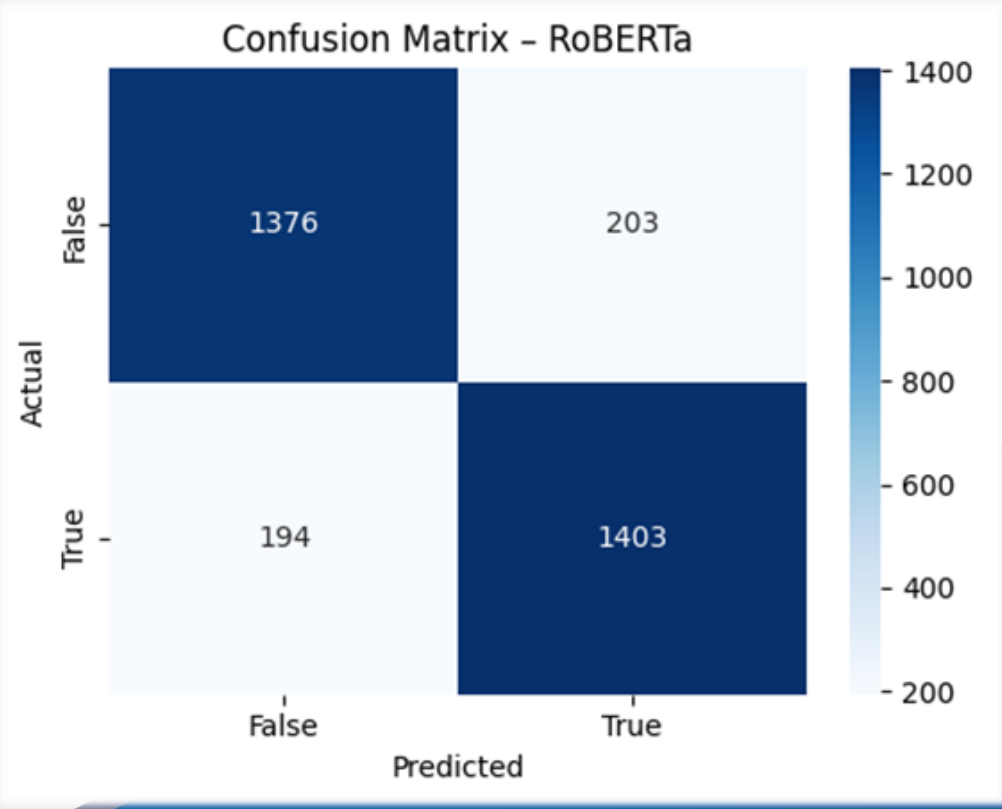
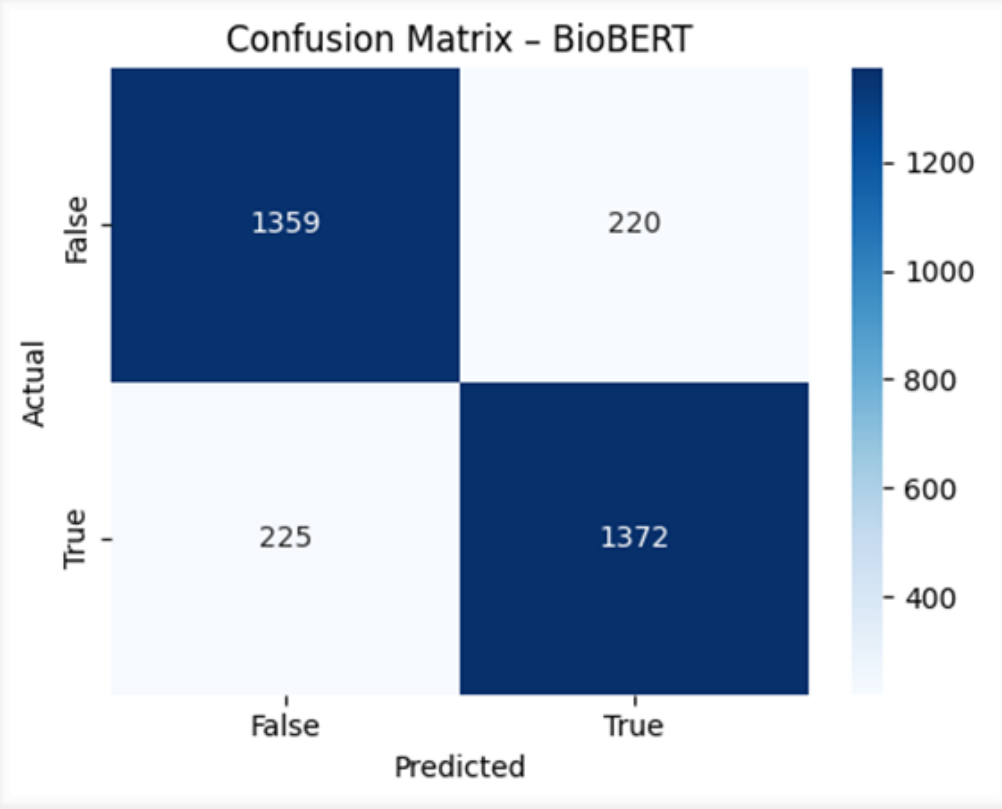
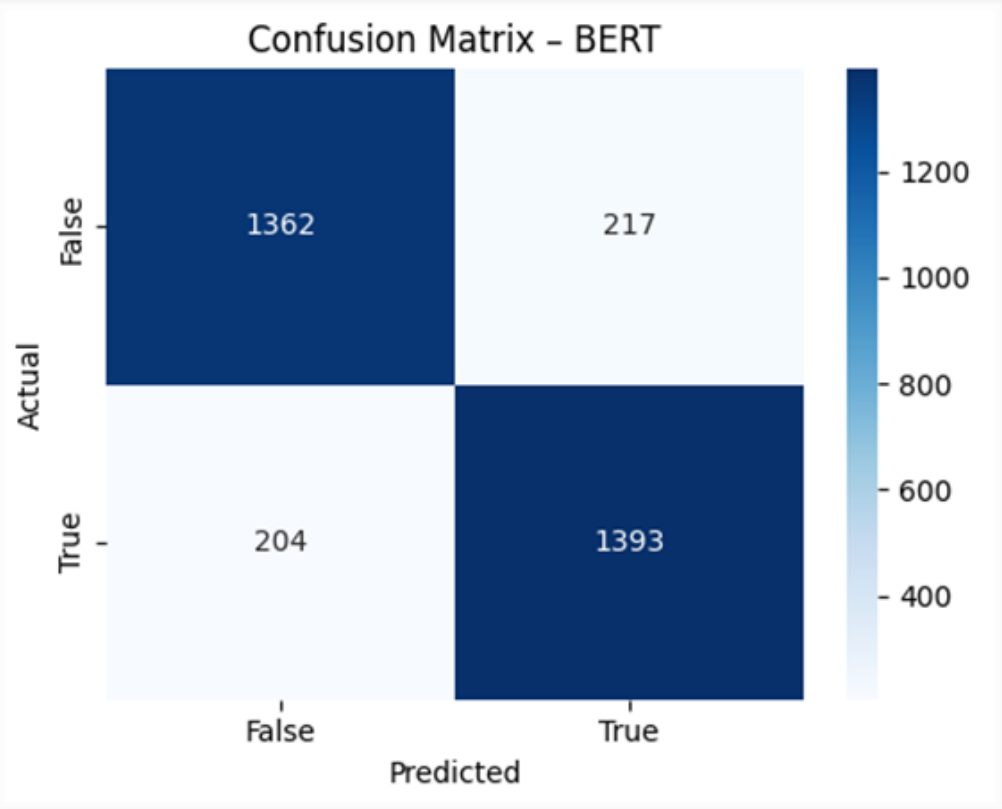
Baseline Performance

- Naive Bayes:
Accuracy: 77%, F1 (False): 0.68, F1 (True): 0.82
- Logistic Regression:
Accuracy: 76%, F1 (False): 0.71, F1 (True): 0.82



Main Results

Model	Accuracy	Precision	Recall	F1 Score	Recall (per label)	F1 Score (per label)
BERT	86.70%	86.50%	87.20%	86.90%	True: 87.2% False: 86.2%	True: 86.9% False: 86.6%
BioBERT	86%	86.20%	85.90%	86%	True: 85.9% False: 86%	True: 86% False: 85.9%
RoBERTa	87.50%	87.30%	87.80%	87.60%	True: 87.8% False: 87.1%	True: 87.6% False: 87.4%



CONCLUSION

Key Insights

- RoBERTa achieved the best overall performance (Accuracy: 87.5%, F1: 87.6%)
- BioBERT, while domain-aware, did not outperform the general models on this task.
- Synthetic data generation (GPT-4-Turbo) improved: Class balance, Style diversity and domain representation

Achieved objectives

- Compared models effectively
- Built realistic, useful synthetic dataset
- Supported decision-making with F1-focused metrics

CODE ORGANIZATION

GitHub Repository

All code, notebooks, and results are available in the public GitHub repository: [MedTruth](#)

Data Files

Sources:

- Three real-world datasets from Kaggle/GitHub
 - used fields: claim and label
- Synthetic Data:
 - ~2,300 claims generated using GPT-4-Turbo.Format: similar to real data, same fields

Code Files:

- **Baseline_models.ipynb**
 - Loads real datasets only (Corona, PubHealth, HLR)
 - Performs preprocessing, EDA, and classification with Logistic Regression and Naive Bayes
- **Synthetic_data_generation_evaluation.ipynb**
 - Fine-tunes GPT-4-Turbo using Corona dataset style
 - Generates 2,300 diverse synthetic claims and saves output to a CSV file
- **Advanced_BERTmodels.ipynb**
 - Loads real + synthetic data (combined dataset)
 - Performs data loading, preprocessing, EDA
 - Trains and evaluates BERT, BioBERT, and RoBERTa

Results and outputs

- **Evaluation metrics saved as CSVs**
 - evaluation_metrics.csv: Contains Accuracy, Precision, F1 scores
- **Visualization outputs**
 - Confusion matrices

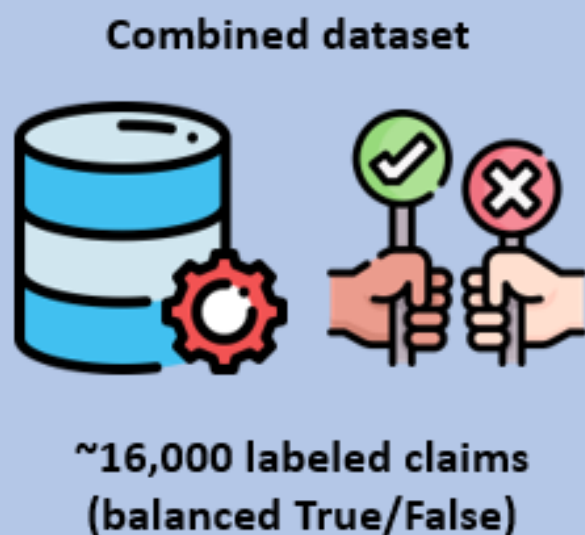


VISUAL ABSTRACT



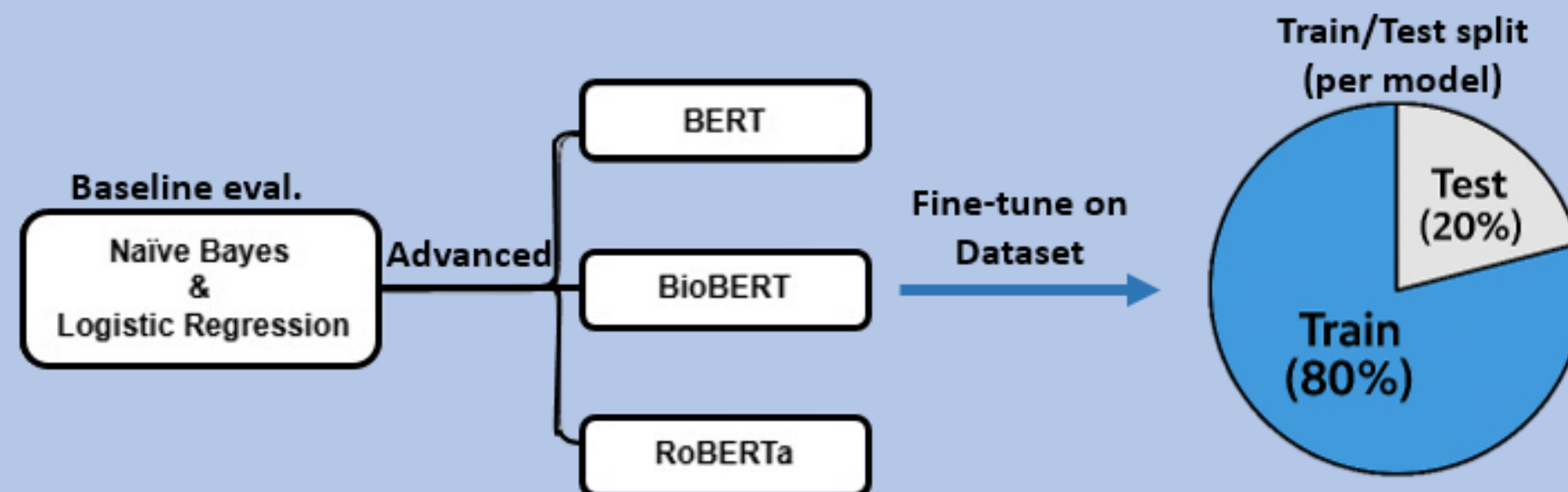
1

Data sources & generation



2

Modeling & Training

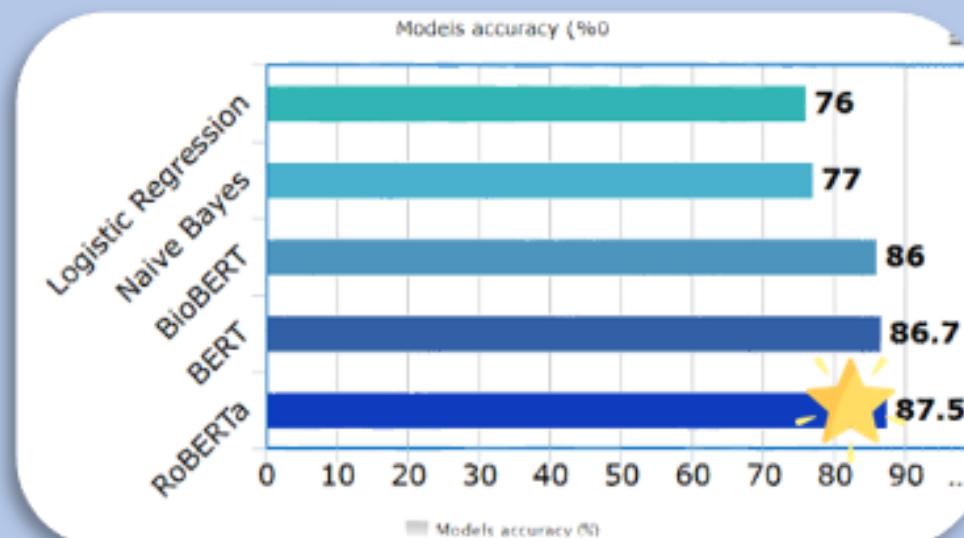


3

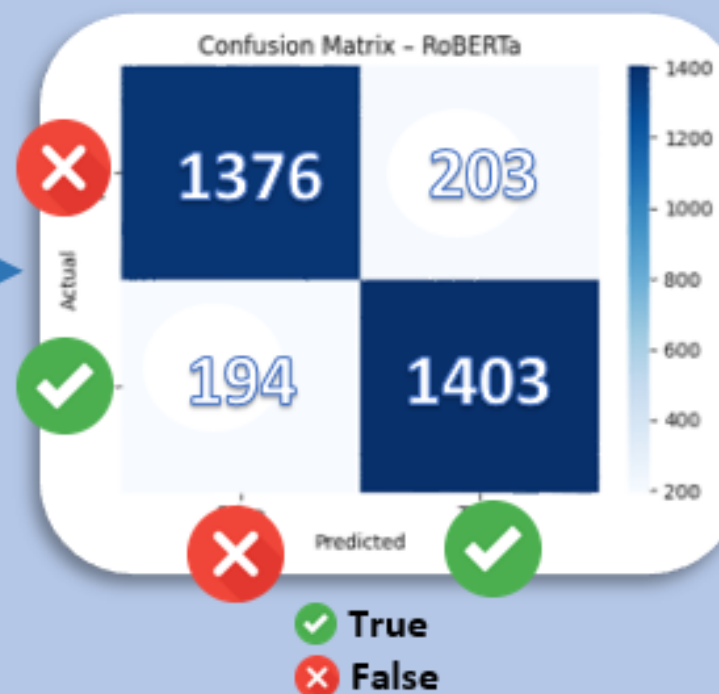
Evaluation & Results

Metrics

	Accuracy
	Precision
	Recall
	F1 Score
	Confusion Matrix



RoBERTa achieved the highest accuracy and F1-score across all models.



RoBERTa showed the best overall results

Synthetic data improved dataset balance and diversity



**THANK
YOU**