MedTruth

Detecting Medical Misinformation On Social Media

NLP Project - Interim presentation

Sara Mangistu
Michelle Zalevsky

Project Description

Project

Detect and classify medical claims on social media (Twitter, Facebook, Reddit etc.) and forums as Real or Fake.

Task

- Input: Social media posts/claims
- Output: Binary classification Fake/Real

Data

Datasets::

- COVID19 Fake News Dataset NLP (Kaggle,).
- PUBHEALTH-DATASET (Kaggle,).
- Misinformation-Detection (Github,).

Evaluation

Models:

- Baseline: Naive Bayes and Logistic Regression.
- Advanced: BERT, BioBERT, fine-tuned RoBERTa

Metrics:

- Accuracy
- Precision/Recall/F1-score,
- Confusion Matrix

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
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.

PLAN

Dataset Preparation

- Manual cleaning to ensure standardized True/False labels
- Merge Datasets

Baseline Modeling & Evaluation

- Train (80%), Test (20%) split
- Models: Naive Bayes/Logistic Regression
- Performances evaluation

Advanced Modeling & Evaluation

- Train (80%), Test(20%) split
- Models: BERT/BioBERT/fine-tuned RoBERTa
- Performances evaluation

NLP PIPLINE

Input: social media post/claim → **Preprocessing** → **Vectorization** → **Modeling** → **Classification** → **True** / **False**

Preprocessing: lowercase, remove punctuation, URLs and stopwords, apply stemming

Vectorization: TF-IDF / BERT embeddings

Modeling: Baseline / Advanced

EDA

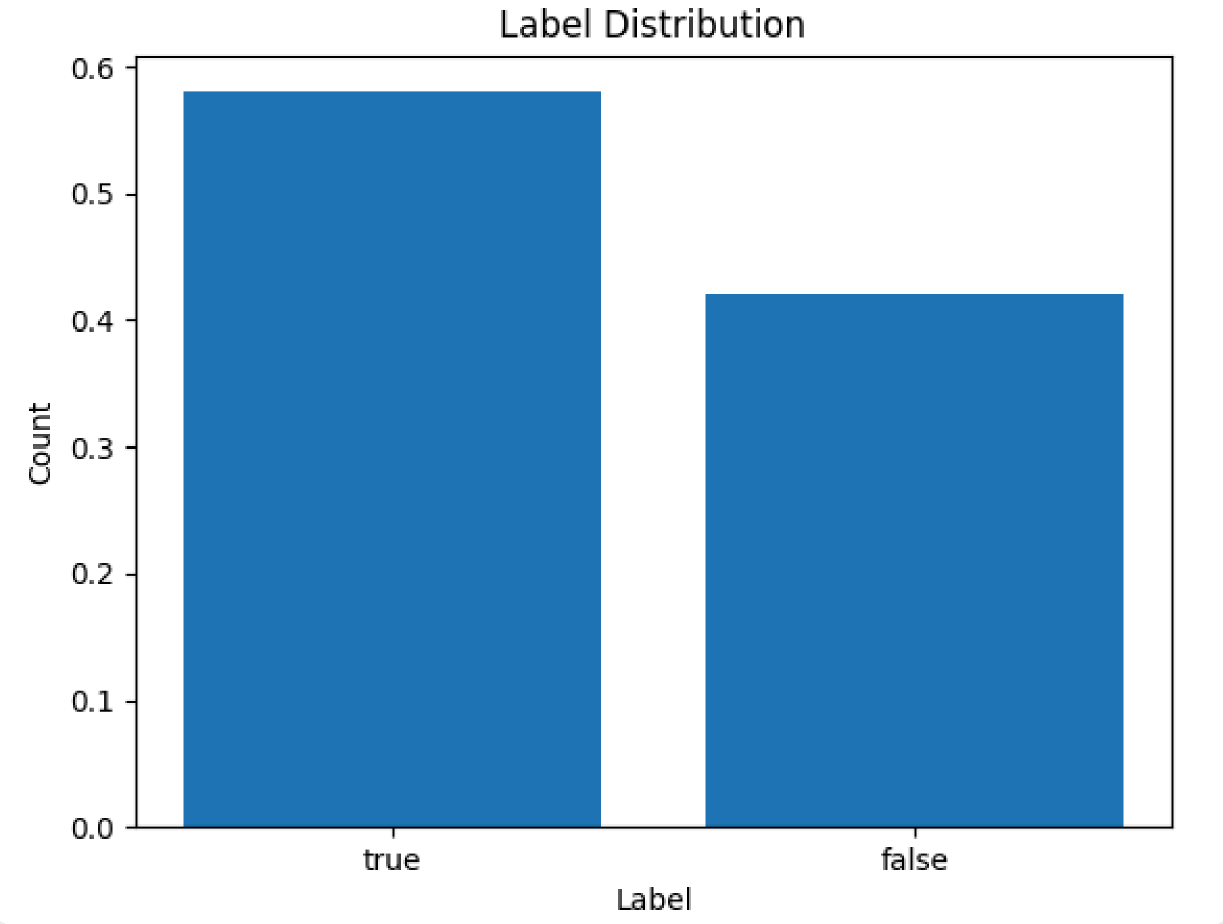
Final dataset size: (15699 , 2)

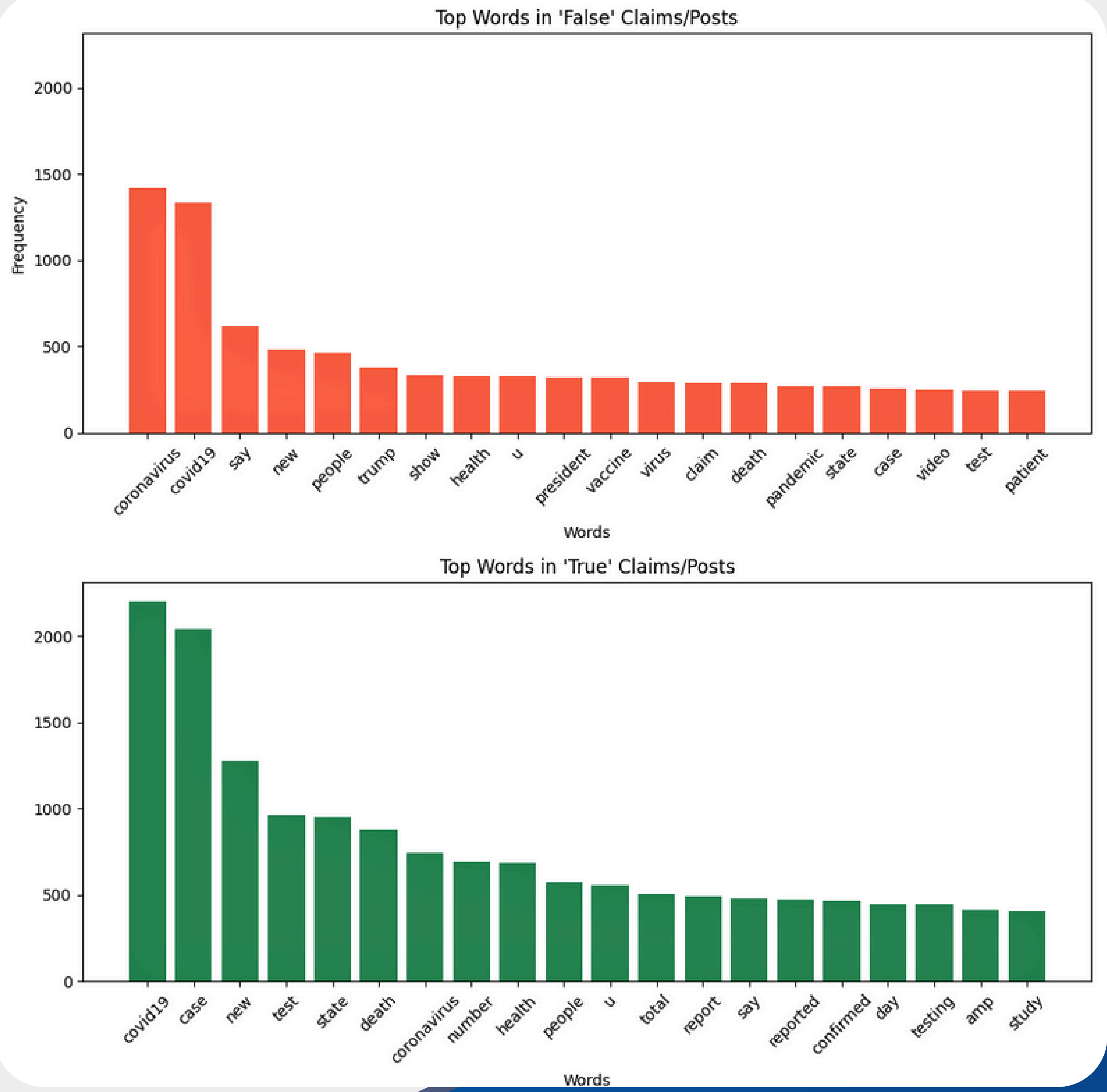
Label Distribution

- True: ~58%, False: ~42%→ Relatively balanced

Text Characteristics

- Average post/claim length: ~18
- Similar top words distribution across classes





Baseline

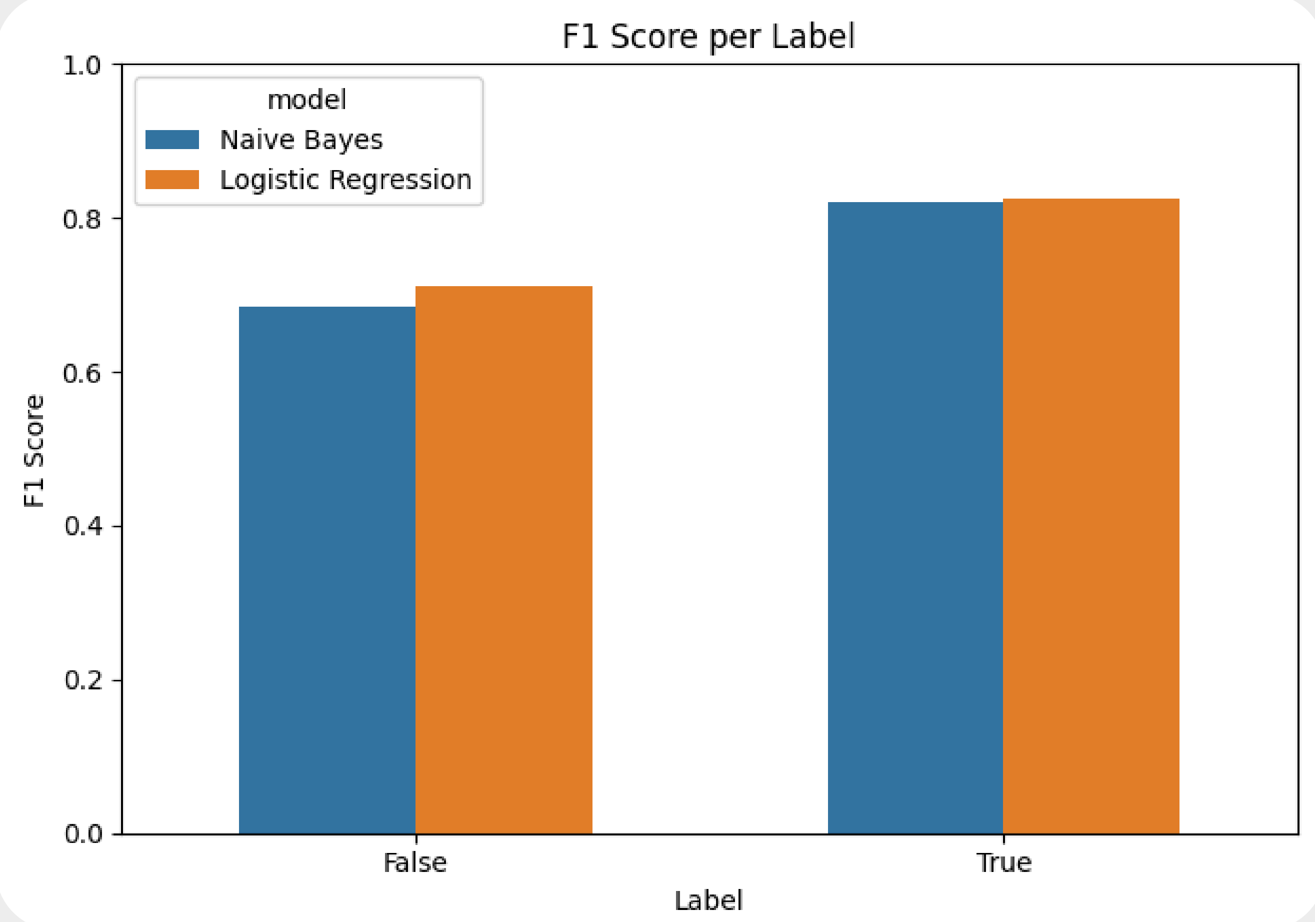
Baseline Performance

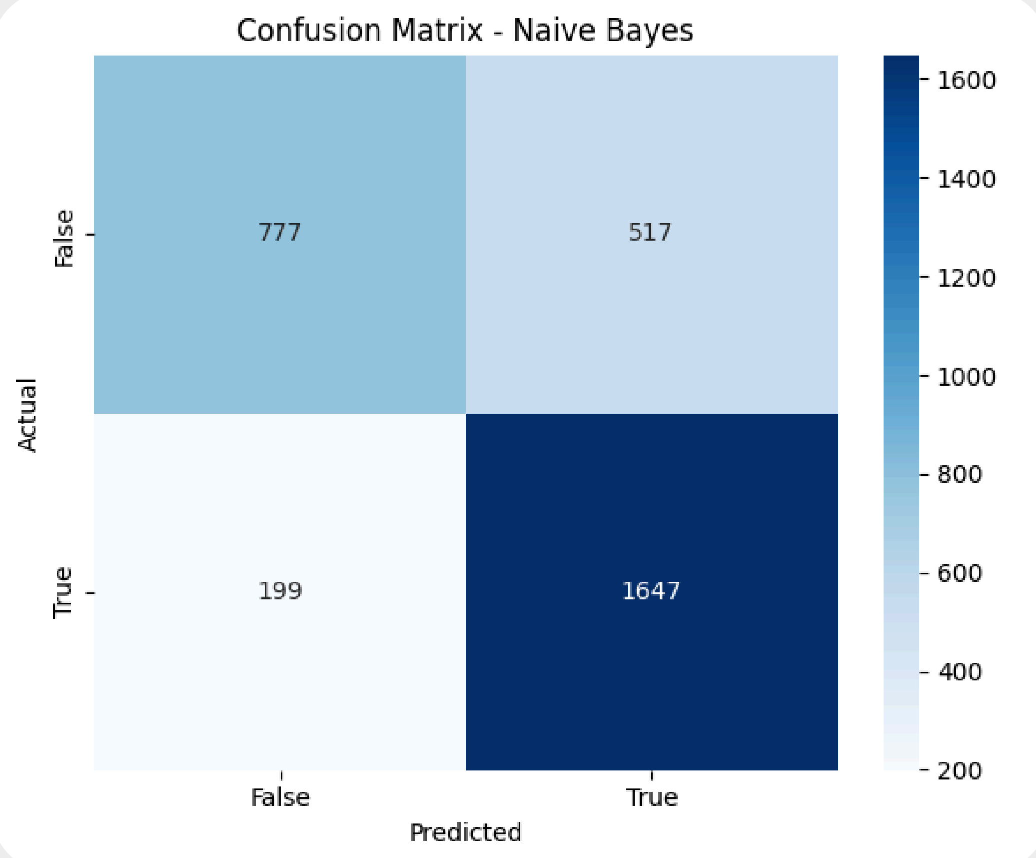
- Naive Bayes:

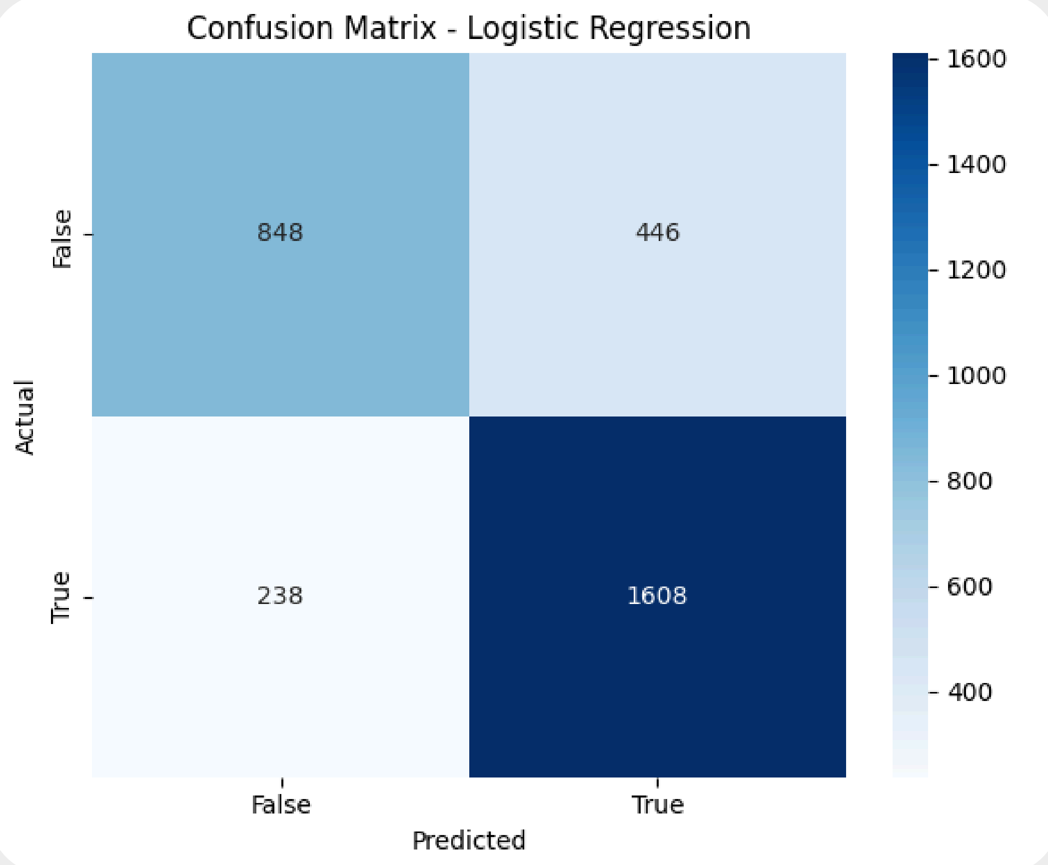
  Accuracy: 77%, F1 (False): 0.68, F1 (True): 0.82
- Logistic Regression:

  Accuracy: 78%, F1 (False): 0.71, F1 (True): 0.82

Models struggle more with the False class (lower recall)







INSIGHTS & RECOMMENDATIONS

Insights

- Slight class imbalance (~42%/58%) is manageable but affects minority class recall
- Dataset is more topically skewed to covid limiting generalization and underrepresent domains
- Baseline models capture common patterns but miss subtle misinformation

Recommendations

- Use contextual embeddings (BERT, BioBERT) to capture deeper meaning
- Concider augment False claims using LLMs (to improve class balance, reflect real-world social claims and domain representation)

THANK YOU