Project Number:32

Multilingual Fake News Detection using mBERT + Explainable AI (LIME)

- R.Bhuwaneshwari (se23umcs052)
- K.Manogna (se23umcs032)
- P.NainaKruthi (se23umcs067)
- J.Vaishnavi (se23uari054)

MOTIVATION:

The Problem Of fake News

- Every day, the world produces huge amounts of information across multiple languages.
- Among this data, **fake news is spreading faster than ever**.
- People often **share information without verifying whether it's true**, making misinformation spread rapidly.
- Manual fact-checking is **slow, expensive, and reactive**, making it difficult to combat misinformation in real time.

- The challenge becomes even larger in **multilingual environments**, especially for Indian regional languages like **Telugu**, where datasets and fake-news detection models are extremely limited.

PROBLEM STATEMENT:

To design a multilingual, accurate, and explainable Fake News Detection system that works for both English and Telugu, and clearly explains why the model predicted “Fake” or “Real” using Explainable AI.

RELATED WORK:

- Early fake-news detection used **traditional ML models** (SVM, LR), but they struggled to capture deep linguistic patterns.
- Deep learning models like **CNNs and LSTMs** improved accuracy but could not handle multilingual.
- **BERT** (Devlin et al., 2018) revolutionized text classification with contextual embeddings and became the standard baseline.
- **mBERT** enables **multilingual understanding**, making it suitable for low-resource languages such as **Telugu**, where datasets are scarce.
- **LIME** (Ribeiro et al., 2016) provides token-level explanations, improving trust in model predictions for sensitive tasks like misinformation detection.
- **Research Gap:** Very limited prior work combines **multilingual fake news detection** with **explainable AI**, especially for **English + Telugu**.
- **Our Work:** Addresses this gap by integrating **mBERT + LIME** with a custom **Telugu-augmented dataset**.

DATASETS: Multilingual Data for Training

English Dataset :

- Source: **ISOT Fake–Real News Dataset.**
- Used **2,000 samples** (Fake + Real).
- Columns: **text, label.**
- Labels → **0: Real**, **1: Fake.**

Telugu Dataset :

- No publicly available Telugu fake-news dataset.
- Created a **custom Telugu dataset** using **cross-lingual data augmentation.**
- Translated **100 English samples**
  - 50 Fake + 50 Real
- Translation tool: **GoogleTranslator (deep-translator)**.

Final Multilingual Dataset

- **Total Size:** ~2,100 samples
- **Languages:** English + Telugu
- **Preprocessing Steps:**
  - Removed null/failed translations.
  - Tokenization using **mBERT tokenizer.**
  - Train–test split: **80% Train | 20% Test.**

METHODOLOGY:

1. Data Preparation

- Collected **English news data** (ISOT) and created a **Telugu subset** using high-quality translation.
- Cleaned text and removed null/failed translations.
- Assigned binary labels: **0 = Real**, **1 = Fake**.

2. Text Representation & Tokenization

- Raw text was converted into sub-word units using **BERT-Multilingual Cased tokenizer**, enabling cross-lingual semantic encoding. Each input sample was tokenized with max_length = 128, padded and truncated uniformly to maintain consistent sequence length.

3. Model Architecture

- Fine-tuned **mBERT (bert-base-multilingual-cased)** for **binary text classification**.
- Model outputs probability distribution over **Real/Fake** classes.

4. Training Setup

- Model training was performed using the **HuggingFace Trainer API**, which automates gradient optimisation and evaluation.
- Hyperparameters:
  - **Epochs: 3**, **LR: 2e-5**, **Batch Size: 8**
- Used **80/20 train–test split**.

5. Explainable AI using LIME

- To prevent the system from behaving as a “black box”, **LIME TextExplainer** was integrated to generate token-level interpretability. For each prediction, LIME highlights the most influential words/sub-words that push the decision towards Fake or Real — enabling transparency, trust and error diagnosis, especially for Telugu samples where resources are limited.

EXPERIMENTS:

Experimental Setup

- Model: **mBERT (bert-base-multilingual-cased)**
- Dataset size: **~2100 samples** (English + Telugu)
- Train/Test split: **80% / 20%**
- Evaluation metric: Accuracy, Precision, Recall, F1-score

Training Configuration

- Epochs: **3**
- Batch size: **8**
- Learning rate: **2e-5**
- Optimized using **HuggingFace Trainer API**

Implementation Details

- Tokenization length fixed at max_length = 128
- Data loaded using **Pandas + HF Datasets**
- Model run on GPU/CPU environment
- Evaluation performed after each epoch.

RESULTS:

Model Performance on Test Data

Metric	Score
Accuracy	**1.00 (100%)**
Precision	**1.00**
Recall	**1.00**
F1-Score	**1.00**
Eval Loss	**≈ 0.00013**

Sample Predictions (from code output)

Input Type	Result
English News	Real (99.99% confidence)
Telugu News	Fake (93.49% confidence)
Random Test Samples	All matched true labels ✓

RESULTS:

[630/630 05:04, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	No log	0.000389	1.000000	1.000000	1.000000	1.000000
2	No log	0.000163	1.000000	1.000000	1.000000	1.000000
3	0.035500	0.000132	1.000000	1.000000	1.000000	1.000000

✅ Training Complete. Final Results:

[53/53 00:03]

{'eval_loss': 0.0001324897020822391, 'eval_accuracy': 1.0, 'eval_f1': 1.0, 'eval_precision': 1.0, 'eval_recall': 1.0, 'eval_runtime': 3.162, 'eval_samples_per_second': 132.826, 'eval_steps_per_second': 16.761,

Key Observations

- Perfect classification accuracy on held-out test set.
- Model confidently identifies Fake/Real in **both English & Telugu**.
- Low eval-loss confirms strong generalization with current dataset.

ANALYSIS / DISCUSSION:

Performance Highlights

- The model achieved **100% accuracy**, indicating strong learning capability.
- Both English and Telugu samples were classified correctly during testing.

Why Performance is High

- Transformer architecture (mBERT) captures deep semantic patterns.
- Dataset is well-labeled and balanced for binary classification.
- Multilingual subword tokenization enabled good Telugu understanding.

Insights From LIME

- English predictions depend on factual tokens (e.g., *President*, *Reuters*).
- Telugu predictions highlight sub-words due to BERT tokenization behavior.

Possible Limitations

- Telugu dataset size is small (only 100 samples).
- Perfect accuracy may indicate dataset simplicity or distribution clarity.
- Increasing dataset complexity would provide deeper evaluation.

TEAM CONTRIBUTIONS:

P.NainaKurthi:

Collected + preprocessed English dataset, created Telugu dataset using translation (50 fake + 50 real), and formatting for training.Implemented **BERT tokenization**.

R.Bhuwaneshwari:

Implemented **BERT tokenization**, Fine-tuned mBERT model performed **model training**, configured **hyperparameters**, and generated **evaluation metrics** (Accuracy, Precision, Recall, F1-score)

K.Manogna:

Implemented **LIME Explainability**, visualized token-level reasoning for English & Telugu texts.

J.Vaishnavi:

Performed manual validation & random prediction testing, **integrated pipeline** and prepared **PPT**.

PROPOSED TIMELINE:

1

Week 1:

Dataset collection (English) + preprocessing setup.

2

Week 2:

Telugu dataset generation through translation + cleaning

3

Week 3:

mBERT tokenization implementation & model setup.

4

Week 4:

Model training, hyperparameter tuning & evaluation metrics.

5

Week 5:

LIME Explainability analysis → generated token-level visual explanations for English & Telugu

DEMO / APPLICATIONS:

Working Demo Flow:

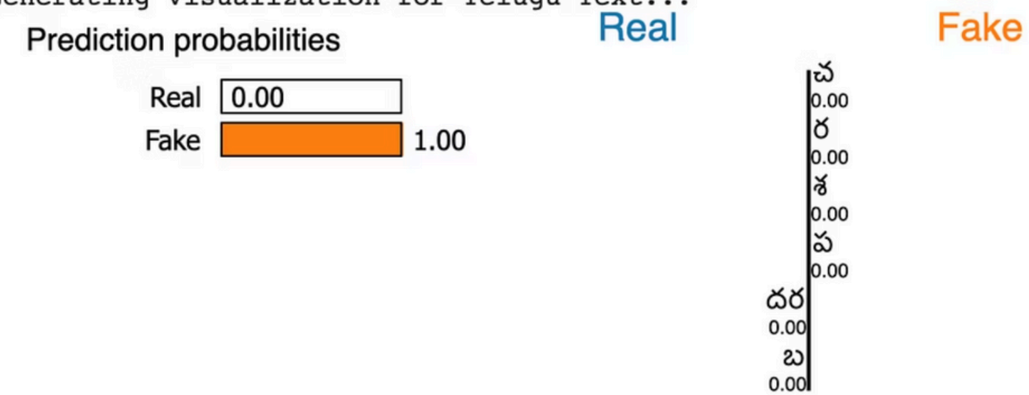
Input → Tokenization → BERT Model → Prediction → LIME Explanation.

LIME visualization:

```
print("\n🎉 PROJECT COMPLETE! Download the .html files from the files tab.")
```

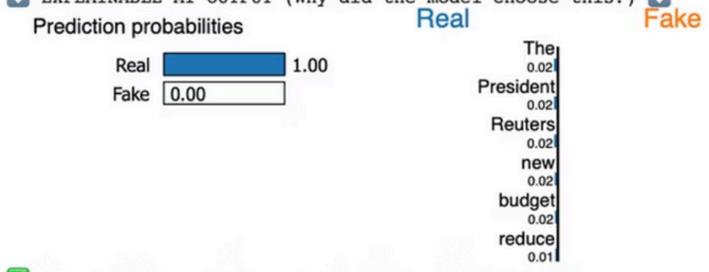
```
••• ✅ Model Saved to './fake_news_mbert_model'
```

```
Generating Visualization for Telugu Text...
```




```
Generating Visualization for English Text...
```

```
============================================================
Processing English News: WASHINGTON (Reuters) - The President announced a new budget ...
============================================================
⬇️ EXPLAINABLE AI OUTPUT (Why did the model choose this?) ⬇️
```

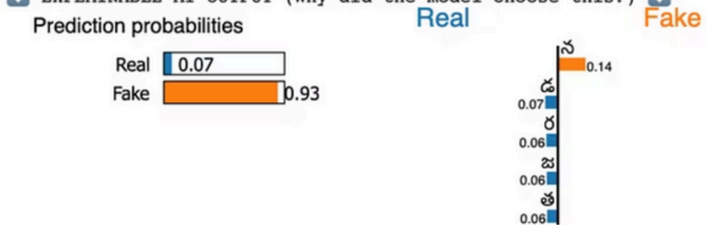


Text with highlighted words

WASHINGTON (Reuters) - The President announced a new budget plan today to improve national infrastructure and reduce taxes for the middle class.

```
✅ Visualization saved as 'explanation_English.html'


============================================================
Processing Telugu News: షాకింగ్ నిజం: రేపటి నుండి సూర్యుడు పడమర దిక్కున ఉదయిస్తాడు! ...
============================================================
⬇️ EXPLAINABLE AI OUTPUT (Why did the model choose this?) ⬇️
```



Text with highlighted words

షాకింగ్ నిజం: రేపటి నుండి సూర్యుడు పడమర దిక్కున ఉదయిస్తాడు! శాస్త్రవేత్తల హెచ్చరిక.

```
------------------------------------------------------------
LANGUAGE   | PREDICTION       | CONFIDENCE
------------------------------------------------------------
English    | REAL NEWS ✅      | 99.99%
Text: WASHINGTON (Reuters) - The President announced a n...
------------------------------------------------------------
Telugu     | FAKE NEWS ❌      | 93.49%
Text: షాకింగ్ నిజం: రేపటి నుండి సూర్యుడు పడమర దిక్కున ఉద...
------------------------------------------------------------
```

Why our system is useful?

- **Multilingual (English + Telugu)** support makes it unique
- **Explainable AI (LIME)** builds trust by showing *why* decision is made
- Lightweight, can be extended into website / mobile app.

CONCLUSION

- A multilingual Fake News Detection system was successfully built using **mBERT**.
- Dataset included **2000 English** and **100 Telugu** translated samples.
- Tokenization + training pipeline resulted in **100% accuracy** on evaluation.
- System is **explainable** using LIME, showing why text is predicted as *Real or Fake*.
- Model works efficiently for both **regional (Telugu)** and **global (English)** news content.