PROJECT ABSTRACT

Fake News Detection using BERT

BERT-based classifier that scores the veracity of news statements, trained on the LIAR benchmark, delivered as a built-to-order student project with code, notebook, report, PPT and viva kit.

01. Title

Fake News Detection using BERT

02. Introduction / Background

Misleading and fabricated news spreads faster than fact-checkers can respond, and manual verification of every claim is impossible at internet scale. Transformer language models such as BERT learn deep contextual representations of text from massive pre-training, and fine-tuning them on labeled claim datasets produces strong text classifiers with modest labeled data. This project fine-tunes bert-base-uncased on the LIAR benchmark dataset of PolitiFact statements and wraps the model in a demo application that scores any pasted claim with a fake/real verdict, confidence and token-level influence highlighting.

03. Problem Statement

Readers, journalists and moderators need a fast first-pass filter that flags suspicious claims for human review. The project addresses this by building a statement-level veracity classifier: given a short news claim, it outputs a fake/real prediction with a confidence score and highlights the words that most influenced the decision, so the verdict is inspectable rather than a black box.

04. Objectives

- Fine-tune bert-base-uncased on the LIAR dataset (12,836 PolitiFact statements) for six-grade truthfulness classification, collapsed to a binary fake/real verdict.
- Reach a design target of F1 approx. 0.70-0.75 on the binary task, with the training notebook logging accuracy, precision, recall, F1 and the confusion matrix on the official test split.
- Build a demo app with statement input, verdict display, confidence bars and token-influence highlighting, plus a dataset-explorer and training-curve views.
- Deliver the complete student kit: source code, training notebook, report, PPT and viva Q&A document.

05. Proposed System / Working

The system has three stages. (1) Data pipeline: LIAR statements are WordPiece-tokenized (max 128 tokens); the six grades (true, mostly-true, half-true, barely-true, false, pants-on-fire) are kept for training and collapsed to binary at inference. Speaker metadata (party, job, credit history) feeds an auxiliary input concatenated with the CLS embedding. (2) Model: bert-base-uncased with a dropout + linear classification head is fine-tuned for about 5 epochs with AdamW and a linear warmup schedule; every epoch logs accuracy, precision, recall, F1 and the confusion matrix. (3) Demo application: a single-page app tokenizes the pasted statement, runs inference, renders the fake/real verdict with a confidence bar, and colors each token by its gradient-based influence on the decision.

06. Technologies / Components Used

- Python 3.10, PyTorch, Hugging Face transformers (bert-base-uncased)
- scikit-learn (metrics, confusion matrix), NumPy, pandas
- Matplotlib (training curves, label distribution)
- Single-file HTML/CSS/JS demo app with statement analyzer and dataset explorer
- LIAR benchmark dataset - Wang, W. Y., ACL 2017 (12,836 statements, 6 truthfulness grades)
- Trained weights exported from the included fine-tuning run

07. Key Features

- Fake/real verdict on any pasted statement with calibrated confidence
- Token-influence highlighting showing which words pushed the decision
- Hybrid text + speaker-metadata input head
- Dataset explorer view with LIAR label distribution
- Training-curve view (F1, train/val loss) reproduced by the notebook
- Batch mode scoring a CSV of statements into a verdict report

08. Applications / Use Cases

- First-pass claim triage for student journalism and fact-check demos
- Teaching material for transformers, fine-tuning and attention-based explanation
- Reference build for other text-classification tasks (spam, hate speech)
- Prototype moderation assistant for campus forums (with human review)

09. Results & Scope

The deliverable is a design-target-driven build, not a pre-measured product. The training notebook computes accuracy, precision, recall, F1 and the confusion matrix on the official LIAR test split; the report presents these as the build's measured results.

The design target is binary F1 approx. 0.70-0.75. Scope is statement-level classification of short political claims in English - it does not verify breaking news, satire or multimedia claims, and a low confidence score must be read as model uncertainty, not ground truth.

10. Conclusion

Fake News Detection using BERT delivers an end-to-end applied NLP project: a real misinformation problem, the standard LIAR benchmark, a transformer fine-tuned with a fully logged experiment, and a demo app that makes the result visible in seconds. The token-influence view and the honest discussion of satire and domain limits give the student strong, defensible material for the report and viva.

11. References

- Wang, W. Y. - 'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection (ACL 2017, arXiv:1705.00648).
- Devlin, J. et al. - BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arXiv:1810.04805).
- Hugging Face - transformers documentation and bert-base-uncased model card.
- PolitiFact (politifact.com) - the fact-checking source behind the LIAR statements.