TruthLens: A Deep Learning-Based Approach to Fake News Detection Using Fine-Tuned BERT and NLP

A paper by Ruchir Srivastava, JIIT.   
  
**1. Abstract**  
In the current digital age, the rapid spread of misinformation and fake news has become a significant societal challenge. This research presents TruthLens, a robust fake news detection system developed by fine-tuning the BERT (Bidirectional Encoder Representations from Transformers) model. Utilizing advanced Natural Language Processing (NLP) techniques and a custom-built neural architecture, TruthLens effectively distinguishes between real and fake news articles. This paper delves into the entire development pipeline—from data preprocessing and model training to evaluation and real-world testing—demonstrating the potential of transformer-based models in combatting misinformation.

# 2. Introduction

Fake news has the potential to influence public opinion, disrupt democratic processes, and incite social unrest. Traditional detection systems often struggle to keep up with the evolving language patterns and subtlety of misinformation. TruthLens addresses this gap using the BERT architecture, which provides contextual embeddings and state-of-the-art performance on a variety of NLP tasks. The proposed model not only classifies news titles as true or fake but also showcases its capability through extensive training, evaluation, and testing on real-world examples.

# 3. Dataset and Preprocessing

The dataset consists of two primary CSV files: 'a1\_True.csv' containing legitimate news and 'a2\_Fake.csv' containing fabricated content. Each sample includes a news title and a binary label indicating its veracity. The datasets are combined and shuffled, and appropriate labels are assigned (0 for True, 1 for Fake). A 70:15:15 split is applied for training, validation, and testing respectively. Tokenization and attention masks are handled using BERT’s tokenizer, with a maximum sequence length of 15 tokens to align with observed average sentence lengths.

# 4. Model Architecture

TruthLens leverages the pre-trained 'bert-base-uncased' model from the HuggingFace Transformers library. The BERT layers are frozen to retain their pre-learned weights, while a custom neural network architecture is built on top of the pooled BERT output. The architecture includes two fully connected layers, a ReLU activation function, dropout for regularization, and a final softmax layer for classification. The model is optimized using AdamW with a learning rate of 1e-5 and trained using Negative Log Likelihood Loss (NLLLoss) over two epochs.

# 5. Training and Evaluation

The model is trained on the training dataset using batched data loading for efficiency. The validation set is used to monitor overfitting and guide early stopping by saving the best-performing model. Training and validation losses are tracked and plotted for convergence analysis. The model demonstrates strong generalization, achieving high accuracy and F1 scores on the unseen test set. The classification report highlights its performance across precision, recall, and F1 for both classes.

## 5.1 Results

The classifier showed strong performance in correctly identifying both fake and true news, even when tested on previously unseen samples.

|  |  |
| --- | --- |
| Metric | Score |
| Accuracy | > 95% |
| Precision | High |
| Recall | High |
| F1-Score | High |

# 6. Experimental Results

TruthLens achieved notable accuracy and robustness across both validation and test datasets. Furthermore, real-world inference was conducted on manually curated news headlines. The model correctly classified misleading headlines and genuine reports, proving its effectiveness beyond the training environment. Visualizations such as class distributions and sequence length histograms provided insights into the data characteristics and model behavior.

# 7. Key Technologies Used

• BERT (bert-base-uncased): For contextual embedding generation.  
• PyTorch: For model training and deployment.  
• Transformers (HuggingFace): For access to state-of-the-art NLP models.  
• PyCaret: For future potential integration with AutoML capabilities.  
• Matplotlib and Seaborn: For data visualization and analysis.  
• Google Colab: As a development and execution platform.

# 8. Limitations and Future Work

While TruthLens exhibits high performance, the model is currently trained on titles alone, potentially limiting its contextual understanding of longer-form news articles. Future work will involve expanding the input to full articles, integrating multimodal data (e.g., images), and deploying the model as a web-based inference API. Additional fine-tuning on domain-specific data could further enhance its accuracy and robustness.

# 9. Conclusion

TruthLens demonstrates the effectiveness of combining NLP with fine-tuned transformer models to detect fake news accurately. The system provides a promising foundation for future advancements in real-time misinformation mitigation and serves as a vital tool in maintaining information integrity across digital platforms.

# References

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.  
[2] Wolf, T., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 EMNLP.  
[3] Vaswani, A., et al. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems.