Don’t Believe Everything You Read

Misinformation Detection in Social Media Posts:

Using Transformer-Based Models for Binary Classification

Christopher Taylor
CSPB 4830: Natural Language Processing

Abstract

This project addresses the critical challenge of automatically detecting misinformation in social media posts and news headlines. Using the FakeNewsNet dataset comprising over 23,000 news headlines from GossipCop and PolitiFact sources, I developed a BERT-based classification model to distinguish between real and fake news. The fine-tuned model achieved 85.13% accuracy and 0.69 F1-score on the test set. To enhance model transparency, LIME (Local Interpretable Model-agnostic Explanations) was implemented to visualize word-level contributions to predictions. Results reveal that certain emotionally charged terms and entities strongly influence classification decisions, with the model performing better on real news (F1: 0.90) than fake news (F1: 0.69). This performance disparity likely stems from dataset imbalance (75% real vs. 25% fake). The interpretability analysis uncovered several key linguistic patterns associated with misinformation, providing insights into future model improvements and potential real-world applications in combating online misinformation.

Introduction

The proliferation of misinformation on social media platforms represents one of today's most pressing challenges to public discourse and democratic processes. False or misleading information can spread rapidly through social networks, often outpacing fact-checking efforts and potentially influencing public opinion on critical issues ranging from health and science to politics and economics. The COVID-19 pandemic and recent election cycles have further highlighted the real-world consequences of unchecked misinformation.

This project aims to develop an automated system for detecting fake news by analyzing news headlines shared on social media platforms. While full-text analysis provides more context, headlines are particularly important as they:

1. Often represent the only content users read before sharing
2. Are crafted to attract attention and engagement
3. Can be misleading even when the associated article is factual

The primary goals of this project were to:

1. Develop a high-performance classifier for distinguishing between real and fake news headlines
2. Implement interpretability techniques to understand the model's decision-making process
3. Identify linguistic patterns and features associated with misinformation

4. Evaluate the practical effectiveness and limitations of transformer-based approaches to this problem

By focusing on both performance and interpretability, this project contributes to the broader effort of creating trustworthy AI systems for misinformation detection that can be understood by human users.

Related Work

The detection of misinformation has been a rapidly evolving field within NLP research. Transformer-based architecture has demonstrated particularly strong performance in this domain. Truică and Apostol (2023) introduced MisRoBÆRTa, an ensemble combining BART and RoBERTa for multi-class misinformation detection, showing significant improvements over standard transformer models. Their approach suggests that ensemble strategies can enhance generalization across varied misinformation tactics.

Earlier work by Singhal et al. (2021) demonstrated that coupling BERT with deep CNN layers (FakeBERT) achieved high accuracy on social media datasets (98.9%) by effectively capturing semantic and contextual features. While these approaches highlight the effectiveness of pretrained transformers for text classification, they also point to limitations in model interpretability and domain adaptation.

Recent research by Su et al. (2023, 2024) has revealed important biases in fake news detectors, particularly their tendency to misclassify LLM-generated text as fake while failing to flag human-written fake news. This raises important questions about model fairness and generalization capabilities. Their 2024 study proposed adversarial training methods using paraphrased genuine news to rebalance detector sensitivity.

The present work differentiates itself by focusing not only on classification performance but also on interpretability through LIME analysis, addressing a gap in understanding how these models make decisions. While most existing literature emphasizes architectural innovations, this project examines both performance and the underlying patterns that drive classification decisions in headline-based fake news detection.

Data

Dataset Source and Composition

This project utilized the FakeNewsNet dataset, a public repository containing news articles and social media metadata. The minimalistic version was used, comprising four CSV files:

- politifact_fake.csv
- politifact_real.csv
- gossipcop_fake.csv
- gossipcop_real.csv

These datasets represent two different domains: political news (PolitiFact) and entertainment/celebrity news (GossipCop). After combining and preprocessing, the dataset contained 23,196 samples, with 17,441 labeled as real (0) and 5,755 as fake (1), revealing a significant class imbalance (75% real vs. 25% fake).

Preprocessing

Data preparation involved:

1. Removing rows with missing title values
2. Filtering out titles shorter than 10 characters to ensure sufficient content for analysis
3. Tokenizing the headlines using the BERT tokenizer (bert-base-uncased)
4. Setting maximum sequence length to 128 tokens
5. Adding special tokens and padding as required by BERT

The dataset was split using stratified sampling to maintain class distribution:

- 80% training (18,556 samples)
- 10% validation (2,320 samples)
- 10% test (2,320 samples)

Limitations

The primary limitation of this dataset is its focus solely on news headlines rather than full articles. While headlines are crucial in social media sharing, they provide limited context. Additionally, the class imbalance potentially biases the model toward predicting the majority class (real news). The temporal range of the data (not explicitly stated in the dataset) may also affect the model's applicability to current news patterns.

Methodology

Model Architecture

This project employed BERT (Bidirectional Encoder Representations from Transformers), specifically the bert-base-uncased variant, fine-tuned for the binary classification task of distinguishing real from fake news headlines. BERT was chosen for its strong performance on text classification tasks and its ability to capture contextual relationships in language. The architecture included:

1. BERT base layer (12 transformer blocks, 12 attention heads, 768-dimensional embeddings)
2. A classification head added on top of the token output
3. Binary cross-entropy loss function for optimization

Training Approach

The model was trained with the following parameters:

- Batch size: 8 (reduced from 16 to prevent out-of-memory errors)
- Optimizer: AdamW with a learning rate of 2e-5 and weight decay of 0.01
- Training epochs: 3
- Device: NVIDIA Tesla T4 GPU on Google Colab

Memory management techniques were implemented, including gradient clearing after each batch and explicit GPU memory cache emptying to maximize available resources.

Evaluation Metrics

To comprehensively assess model performance, particularly given the class imbalance, multiple evaluation metrics were employed:

- Accuracy: Overall proportion of correct predictions
- F1-score: Harmonic mean of precision and recall
- Precision: Proportion of positive identifications that were correct
- Recall: Proportion of actual positives that were identified
- AUC-ROC: Area under the Receiver Operating Characteristic curve
- Confusion matrix: Visual representation of prediction errors

Interpretability with LIME

A key methodological contribution of this project was the implementation of LIME (Local Interpretable Model-agnostic Explanations) to understand the model's decision-making process. LIME works by:

1. Perturbing the input by removing words
2. Generating predictions for these perturbed instances
3. Fitting a local linear model to approximate the complex model's behavior
4. Identifying which words most strongly influence prediction

For each analyzed headline, LIME generated:

- A feature importance ranking showing the top 10 most influential words
- Color-coded visualizations (green for words supporting "real" classification, red for "fake")
- A quantitative weight for each word's contribution to the prediction

Results

Model Performance

The final model achieved the following metrics on the test set:

METRIC	VALUE
Accuracy	0.8513
F1 Score	0.6944
Precision	0.7076
Recall	0.6817
AUC-ROC	0.8879

The classification report revealed an important performance disparity between classes:

	PRECISION	RECALL	F1 SCORE	SUPPORT
REAL	0.90	0.91	0.90	1745
FAKE	0.71	0.68	0.69	575

The confusion matrix showed 1586 true negatives (correctly classified real news), 409 true positives (correctly classified fake news), 158 false positives, and 167 false negatives.

Training exhibited a consistent decrease in loss across epochs:

- Epoch 1: Loss 0.4077, Validation Accuracy 0.8496, F1 0.6330
- Epoch 2: Loss 0.2739, Validation Accuracy 0.8547, F1 0.7020
- Epoch 3: Loss 0.1699, Validation Accuracy 0.8478, F1 0.6901

LIME Analysis Results

LIME analysis of representative examples revealed insightful patterns:

Example 1 (True Class: Fake, Predicted: Real with 0.9891 confidence) "Pregnant Khloe Kardashian enjoys baby shower with sisters and Kris"

Top features included:

- "enjoys" (negative weight: -0.0648, contributes to Fake)
- "Kris" (negative weight: -0.0630, contributes to Fake)
- "shower" (negative weight: -0.0508, contributes to Fake)

Despite these top words contributing to a "Fake" classification, the model incorrectly predicted this headline as "Real," suggesting other features or patterns influenced the final decision.

Example 2 (True Class: Fake, Predicted: Real with 0.9740 confidence) "Camilla's evil plot exposed"

Top features included:

- "exposed" (negative weight: -0.6095, contributes to Fake)
- "Camilla" (positive weight: 0.2588, contributes to Real)
- "plot" (negative weight: -0.1277, contributes to Fake)
- "evil" (positive weight: 0.0713, contributes to Real)

This example reveals that sensationalistic words like "exposed" and "plot" push toward a "Fake" classification, while the royal-related entity "Camilla" contributes to a "Real" prediction.

Example 3 (True Class: Real, Predicted: Real with 0.7933 confidence) "Orlando Bloom Stops Play Twice After Seeing Audience Member With iPad"

Top features included:

- "Member" (negative weight: -0.2265, contributes to Fake)
- "Audience" (positive weight: 0.1180, contributes to Real)

- "Bloom" (positive weight: 0.1180, contributes to Real)
- "Orlando" (positive weight: 0.1014, contributes to Real)

The named entity "Orlando Bloom" strongly influenced the correct classification as "Real," while "Member" contributed to a "Fake" classification.

Discussion

Performance Analysis

The model achieved solid overall performance with 85.13% accuracy and an AUC-ROC of 0.8879, indicating good discriminative ability. However, the F1 score of 0.6944 reflects challenges in handling the class imbalance. The model performs notably better on real news (F1: 0.90) than fake news (F1: 0.69), likely due to two factors:

1. Class imbalance: With 75% of training examples being real news, the model has fewer opportunities to learn fake news patterns.
2. Feature overlap: Certain linguistic features may appear in both real and fake news, making clear discrimination challenging.

The training trajectory showed rapid improvement in the first two epochs followed by a slight decline in validation F1 score in the third epoch (from 0.7020 to 0.6901), suggesting that longer training might lead to overfitting rather than continued improvement.

Interpretability Insights

The LIME analysis revealed several interesting patterns in how the model makes decisions:

1. Named entity influence: Celebrities, politicians, and public figures often influence classification, with some entities (like "Camilla") pushing toward "Real" classification regardless of context.
2. Sensationalistic language: Words like "exposed," "plot," and "evil" typically contribute to "Fake" classifications, reflecting tabloid-style language common in misleading headlines.
3. Misclassification patterns: In both misclassified examples analyzed, the model strongly predicted "Real" (>0.97 confidence) despite multiple words contributing to "Fake" classification, suggesting that certain entity names overwhelm other signals.
4. Context limitations: The model's reliance on individual words without broader context may limit its ability to detect more sophisticated misinformation that uses factual-sounding language.

Practical Implications

These findings have several implications for misinformation detection:

1. Models trained on headline-only data show promise but have clear limitations in accurately detecting fake news, particularly for headlines with factual elements or celebrity mentions.
2. Interpretability tools like LIME can help identify potential biases in model decision-making, such as the overemphasis on certain named entities.

3. The performance gap between real and fake news detection suggests that specialized approaches may be needed for less-represented classes in imbalanced datasets.

Conclusion & Future Work

This project demonstrated that a BERT-based classifier can effectively detect misinformation in news headlines with reasonable accuracy (85.13%). The integration of LIME interpretability provided valuable insights into the model's decision-making process, revealing patterns in how it distinguishes between real and fake news. The model's stronger performance on real news compared to fake news highlights the challenges posed by class imbalance in misinformation datasets.

Several directions for future work could address the limitations identified:

1. Data augmentation: Techniques like SMOTE or synthetic data generation could help balance the class distribution and improve fake news detection.
2. Multi-modal approaches: Incorporating image analysis or user engagement metrics could provide additional signals beyond text content.
3. Ensemble methods: Combining multiple models, as seen in MisRoBÆRTa, might capture different aspects of misinformation and improve overall performance.
4. Domain adaptation: Testing and fine-tuning on different types of misinformation (political, scientific, health-related) could improve generalization.
5. Contextual enhancement: Including article text or linked content when available could provide richer context for classification decisions.

The interpretability findings suggest that misinformation detection models benefit from transparency tools that can expose potential biases or overreliance on specific features. As misinformation becomes increasingly sophisticated, combining high-performance models with interpretability techniques will be essential for developing trustworthy systems that can adapt to evolving challenges in the information ecosystem.

Bibliography

1. Cao, S., Zhang, L., Yao, S., Li, T., Jiang, Y., Yang, Y., Liu, Z., Lin, C., Li, J., Wang, H., & Wu, J. (2024). Exploring Large Language Models for Scientific News Misinformation Detection. arXiv. https://arxiv.org/abs/2401.01979

2. Pavlyshenko, B. (2023). Fine-tuning Llama 2 language models for disinformation detection and sentiment-aware natural language inference. arXiv. https://arxiv.org/abs/2310.04847

3. Singhal, T., Shah, R. R., & Chakraborty, T. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 80(8), 11765–11788. https://doi.org/10.1007/s11042-020-10183-1

4. Su, Y., Sahijwani, H., Li, Y., Chaturvedi, S., Gehrmann, S., Jernite, Y., & Wang, W. Y. (2023). Is your Fake News Detector Fake News? Misclassifying Human-Written News as AI-Generated. arXiv. https://arxiv.org/abs/2311.02260

5. Su, Y., Chaturvedi, S., Gehrmann, S., Jernite, Y., & Wang, W. Y. (2024). Mitigating Bias in Fake News Detectors Using Paraphrased Genuine News. arXiv. https://arxiv.org/abs/2401.12283

6. Truică, C.-O., & Apostol, E. (2023). MisRoBÆRTa: A multi-class classification model for fake and misleading content detection using transformers. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-023-04592-w

7. Zhou, X., Shah, S., & Jin, F. (2024). Correcting Misinformation with Multi-modal Retrieval-Augmented Large Language Models. arXiv. https://arxiv.org/abs/2403.06257

8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778