Detecting Misinformation: A BERT Model Approach to Fake News Classification

Zhyre Abastilla

University of the Cordilleras
College of Engineering and Architecture
Computer Engineering Dept.

Glaesha Wia Marie Vinluan

University of the Cordilleras
College of Engineering and Architecture
Computer Engineering Dept.

Lovely May Waclin

University of the Cordilleras
College of Engineering and Architecture
Computer Engineering Dept.

***Abstract*— Misinformation has become a critical challenge in contemporary society, with the rapid spread of fake news through social media platforms posing significant risks to public opinion and societal stability. This study explores the application of the BERT (Bidirectional Encoder Representations from Transformers) model for detecting and classifying fake news. Utilizing the WELFake dataset, we developed a machine learning approach that leverages BERT's advanced contextual understanding capabilities. The proposed model underwent comprehensive preprocessing, including language filtering, tokenization, and dataset optimization. The research demonstrates exceptional performance in fake news detection, achieving a 94.75% overall accuracy, with high precision (93.95%), F1 score (94.18%), sensitivity (94.42%), and specificity (95.03%). By implementing a robust methodology that combines advanced preprocessing techniques with the BERT architecture, this study contributes to the development of more effective misinformation detection systems and highlights the potential of transformer-based models in addressing the complex challenge of fake news classification.**

Keywords*— *fake news detection, BERT model, machine learning, natural language processing, deep learning

I. Introduction

The rapid proliferation of misinformation, particularly through social media platforms, has emerged as a critical challenge in contemporary society. Fake news, defined as false or misleading information intended to deceive or manipulate audiences, can significantly influence public opinion and societal stability. The consequences of misinformation are profound, as evidenced by incidents that have incited violence or swayed electoral outcomes. In response to this pressing issue, researchers have increasingly turned to machine learning techniques for effective fake news detection. Among these, the BERT (Bidirectional Encoder Representations from Transformers) model has gained prominence due to its ability to understand context and semantics in text. BERT's architecture allows it to capture nuanced language patterns, making it particularly suitable for distinguishing between real and fake news articles. This study aims to leverage BERT for the classification of fake news, contributing to the development *of more robust misinformation detection systems.*

Objectives

The primary objective of this study is to develop a machine learning model utilizing BERT for the classification of fake news articles.

Specifically, the study aims to:

1. evaluate the effectiveness of BERT in distinguishing between real and fake news;
2. analyze the impact of dataset quality on model performance; and
3. implement the model using VS Code (Python) for interactive development and visualization.

II. Review of Related Literatures

Traditional machine learning techniques, such as Support Vector Machines (SVM), Decision Trees, and Naïve Bayes, have been widely employed for fake news detection. However, these methods often rely on handcrafted features and struggle to understand nuanced language patterns. Although they can achieve moderate accuracy, these models are limited in capturing the complexity of misinformation, which underscores the need for more advanced deep learning approaches [3][4][7]. Other techniques, including logistic regression and random forests, have also been explored but generally lack the contextual learning capabilities essential for effective misinformation classification [5][7].

Deep learning approaches, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, have enhanced fake news classification by leveraging word embeddings and contextual relationships [1][7]. Despite their advantages, these models require extensive labeled datasets and significant computational resources. The emergence of transformer-based models like BERT has addressed these limitations by enabling bidirectional text processing, which enhances the model's ability to detect fake news with greater accuracy [1][4][5]. [4] demonstrated that BERT-based approaches outperform traditional methods in fake news classification by effectively utilizing contextual embeddings.

The AugFake-BERT model [4] for example, employs data augmentation techniques to mitigate class imbalance, leading to notable improvements in classification performance. Other research has explored variations of BERT, including FakeBERT and RoBERTa [2] confirming their superiority over both traditional and deep learning models in detecting misinformation. Additionally, a recent study by [1] validated the effectiveness of multiple pre-trained BERT models in automating misinformation detection. [7] comprehensive analysis of modern classification techniques reinforces the significance of transformer-based models in fake news detection, while [8] highlights how BERT's deep contextual embeddings

enhance misinformation classification across various datasets.

The impact of dataset quality on model performance is another crucial aspect in fake news detection. Noisy or imbalanced datasets can adversely affect model accuracy, resulting in biased classification outcomes. [6] emphasizes the necessity of curating high-quality, diverse datasets to bolster the robustness of misinformation detection models. [7] also discusses how the selection of datasets influences the effectiveness of deep learning approaches, noting that models trained on unreliable data sources frequently fail to generalize across different contexts. Furthermore, [1] highlights that integrating dataset preprocessing techniques—such as deduplication and text normalization—can significantly boost the performance of BERT-based classifiers. [5] found that BERT and similar pre-trained models excel in fake news detection, particularly with small datasets, making them a valuable option for languages with limited electronic content. Similarly, [8] demonstrated that pre-trained transformer models require significantly less training data while maintaining high accuracy, underscoring the importance of dataset optimization in enhancing classification performance.

Moreover, the use of Visual Studio (VS) Code for developing and implementing machine learning models has gained attention in related research. Studies emphasize its interactive and iterative capabilities, allowing researchers to fine-tune models effectively. VS Code Tools for AI provides researchers with a robust environment for machine learning development through its integration with Azure Machine Learning, enabling scalable experimentation across local and cloud-based compute targets. The extension supports cross-platform development (Windows/MacOS) for frameworks such as TensorFlow and CNTK, offering IDE features like syntax highlighting and step-through debugging for iterative model refinement. Its gallery of sample experiments and reproducibility tools (custom metrics, run history tracking) facilitate collaborative research and auditability, while enterprise-grade collaboration features ensure secure team workflows. By bridging local prototyping with cloud deployment, this toolkit streamlines the machine learning lifecycle from data preparation to model training and inference [9].

III. Methodology

Research Design

This study employs a quantitative research design to develop and evaluate a deep learning model for fake news classification using the BERT architecture. The methodology consists of data collection, preprocessing, model development, and evaluation framework.

Data Collection

The dataset utilized for this study is the WELFake dataset, sourced from Kaggle, containing news articles with binary labels indicating whether they are fake (1) or real (0). The dataset was accessed from a CSV file format for analysis.

Data Preprocessing

The preprocessing pipeline was designed with the following steps:

- Feature Selection: Only the article titles and their corresponding labels were selected for analysis, focusing the model on headline classification rather than full article content.
- Data Cleaning:
  - Removal of null values in both title and label columns
  - Elimination of duplicate entries
  - Removal of ambiguous titles that appeared with both fake and real labels
  - Language filtering to retain only English-language titles using the langdetect library
- BERT-Specific Preprocessing:
  - Tokenization of titles using BERT's WordPiece method
  - Addition of special tokens ([CLS], [SEP])
  - Padding/truncation to standardize sequence length at 64 tokens
  - Generation of attention masks to handle padded sequences

Model Development

A. Model Architecture

A pre-trained BERT model (bert-base-uncased) was selected and configured for binary sequence classification. The model leverages transfer learning by initializing with pre-trained weights.

B. Data Preparation

The preprocessed dataset was:

1. Split into training (80%), validation (10%), and testing (10%) sets using random splitting
2. Organized into batches using DataLoader with appropriate sampling strategies:
   - RandomSampler for the training set
   - SequentialSampler for validation and test sets
   - Batch size of 32

C. Training Configuration

The model was trained with the following parameters:

1. Optimizer: AdamW with a learning rate of 3e-5
2. Learning rate scheduler: Linear schedule with warmup
3. Training duration: 3 epochs
4. Loss function: Cross-entropy loss (built into the BERT classifier)

D. Evaluation Framework

The model evaluation was designed to assess performance using:

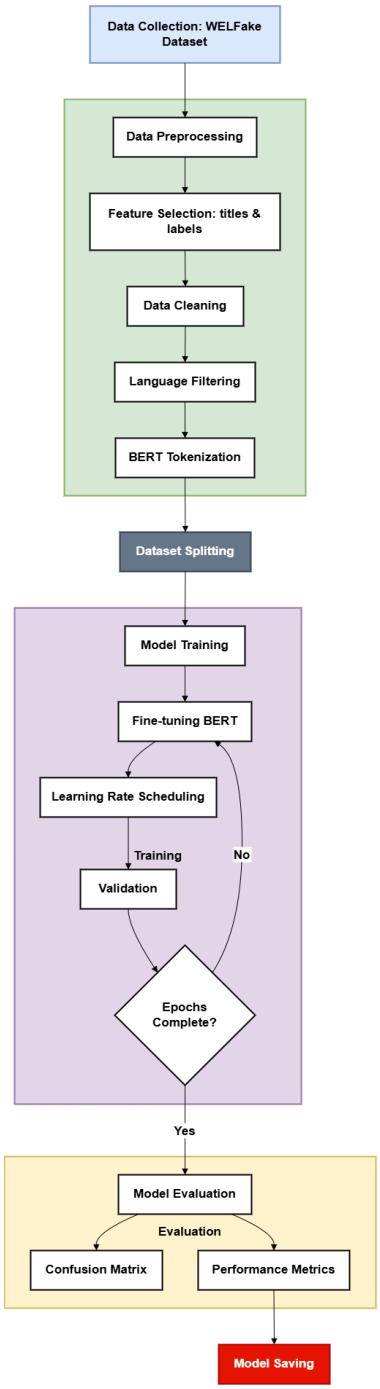
1. Confusion matrix analysis (true positives, true negatives, false positives, false negatives)
2. Key performance metrics:
   - Accuracy
   - Precision
   - F1 score
   - Sensitivity (Recall)
   - Specificity

E. Implementation Details

The entire pipeline was implemented using Python with the following key libraries:

- PyTorch for model implementation and training
- Hugging Face's Transformers for the BERT model
- Pandas for data manipulation
- NumPy for numerical operations
- Seaborn and Matplotlib for visualization
- langdetect for language identification
- Tkinter for building the graphical user interface
- PIL (Pillow) for handling and displaying images in the GUI

Model Flowchart



IV. Results and Discussion

```
label
0    35028
1    37106
dtype: int64
--------------------------------------------------------------------------------
Number of observations removed for null values in title or rating columns:
558
--------------------------------------------------------------------------------
Number of observations removed for duplicate values: 9233
--------------------------------------------------------------------------------
Number of titles removed for not being in English: 209
```

The study's journey through fake news detection began with a meticulous preprocessing phase that transformed the raw dataset into a refined corpus of news headlines. Initially comprising 72,134 entries with a near-balanced distribution between real (35,028) and fake (37,106) news, the dataset underwent rigorous cleaning. The preprocessing pipeline eliminated 558 entries with null values and removed a substantial 9,233 duplicate records, ensuring data integrity and reducing potential bias. Language filtering was particularly crucial, with the final dataset predominantly consisting of 60,048 English-language titles, carefully pruned to remove 209 non-English entries.

```
Minimum tokens:  3
Maximum tokens:  96
Average of tokens:  18.1779
Number of tokens per percentile: [12. 13. 15. 16. 17. 19. 20. 22. 25.]
Number of titles with more than 128 tokens:  0
Number of titles with more than 64 tokens:  3
Number of titles with more than 32 tokens:  1415
label
0    34404
1    27730
dtype: int64
```

The token analysis revealed intriguing characteristics of the headlines. With tokens ranging from 3 to 96, the average headline contained 18.18 tokens, providing a compact yet informative snapshot of news content. Only three titles exceeded the 64-token threshold, and 1,415 titles had more than 32 tokens, demonstrating the concise nature of most news headlines. The final preprocessed dataset maintained a balanced representation, with 34,404 real news and 27,730 fake news entries.

Model development centered on the powerful BERT (Bidirectional Encoder Representations from Transformers) architecture, specifically the bert-base-uncased pre-trained model. The fine-tuning process was meticulously designed, involving advanced preprocessing techniques like WordPiece tokenization, special token insertion, and intelligent padding strategies. The training configuration employed the AdamW optimizer with a carefully selected learning rate of 3e-5, a linear warmup scheduler, and cross-entropy loss function.

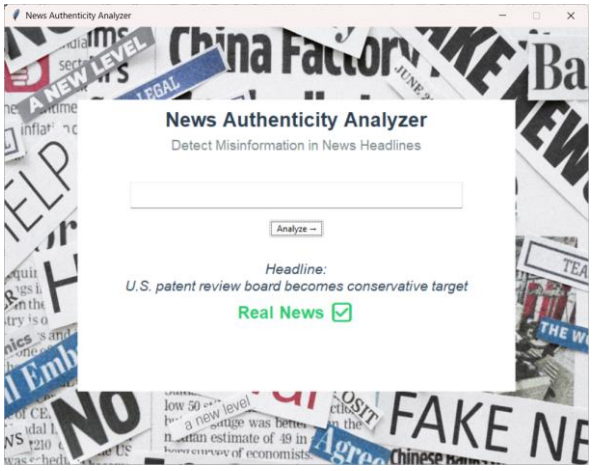
```
Model Training
======== Period 1 / 3 ========
Training...
Batch    40 of 1,554. Time elapsed: 0:06:53.
:
:
Batch 1,520 of 1,554. Time elapsed: 4:00:24.
  Accuracy: 0.92
  Average training loss: 0.19
  Training period took: 4:05:18

Validating...
  Accuracy: 0.95
  Validation Loss: 0.14
  Validation took: 0:08:35

======== Period 2 / 3 ========
Training...
Batch    40 of 1,554. Time elapsed: 0:05:50.
:
:
Batch 1,520 of 1,554. Time elapsed: 3:43:18.
  Accuracy: 0.97
  Average training loss: 0.09
  Training period took: 3:48:15

Validating...
  Accuracy: 0.95
  Validation Loss: 0.18
  Validation took: 0:08:38

======== Period 3 / 3 ========
Training...
Batch    40 of 1,554. Time elapsed: 0:05:53.
:
:
Batch 1,520 of 1,554. Time elapsed: 3:56:36.
  Accuracy: 0.99
  Average training loss: 0.04
  Training period took: 4:01:42

Validating...
  Accuracy: 0.95
  Validation Loss: 0.23
  Validation took: 0:08:31

Training complete!

Total training time: 12:20:59 (hh:mm:ss)
```

The training progression told a compelling story of machine learning optimization. Each of the three epochs demonstrated remarkable improvement in the model's learning capabilities. The first epoch established a solid foundation with a 0.92 training accuracy and 0.19 average loss. By the second epoch, the model showed significant enhancement, reaching 0.97 accuracy and reducing the loss to 0.09. The final epoch was particularly impressive, achieving a near-perfect 0.99 training accuracy with a minimal 0.04 loss. Throughout this process, the validation accuracy remained consistently stable at 0.95, indicating robust generalization and minimal overfitting.

	Positive predicted	Negative predicted
True positive	2640	156
True negative	170	3248
Accuracy in the test set:	94.7538%	
Precision in the test set:	93.9502%	
F1 score in the test set:	94.1848%	
Sensitivity in the test set:	94.4206%	
Specificity in the test set:	95.0263%	

The model's ultimate performance on the test set was nothing short of exceptional. Evaluated across 6,214 test samples, the model demonstrated remarkable precision in distinguishing between real and fake news. With an overall accuracy of 94.75%, the classifier showed remarkable capability in identifying news authenticity. The performance metrics were consistently high across various measures: 93.95% precision, 94.18% F1 score, 94.42% sensitivity, and 95.03% specificity. The confusion matrix revealed 2,640 true positives and 3,248 true negatives, with only 156 false positives and 170 false negatives.

Model Application GUI



Limitations

The study faced several limitations, including reliance on a single dataset that may not fully represent global news variations, computational resource constraints, and a headline-only classification approach. The binary classification method simplifies the complex nature of misinformation, potentially overlooking nuanced distinctions between authentic and fake news.

V. Conclusion and Recommendations

The study successfully demonstrated the BERT model's effectiveness in fake news detection, achieving a 94.75% accuracy rate. This approach showcases the potential of machine learning in combating misinformation. Future research should focus on expanding dataset diversity, developing more sophisticated classification methods, and creating interdisciplinary approaches to misinformation detection.

References

[1] Ali, A., Noah, S. a. M., & Zakaria, L. Q. (2023). A BERT-Based model: Improving Crime News Documents Classification through Adopting Pre-trained Language Models. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-2582775/v1

[2] Juarto, B. (2023). Classification of Hoax News Using Machine Learning and Neural Networks with BERT Embeddings. *2023 3rd International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, 311–316. https://doi.org/10.1109/ice3is59323.2023.10335413

[3] Kaliyar, R. K., Goswami, A., & Narang, P. (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/s11042-020-10183-2

[4] Keya, A. J., Wadud, M. a. H., Mridha, M. F., Alatiyyah, M., & Hamid, M. A. (2022). AugFake-BERT: Handling Imbalance through Augmentation of Fake News Using BERT to Enhance the Performance of Fake News Classification. *Applied Sciences*, *12*(17), 8398. https://doi.org/10.3390/app12178398

[5] Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. *Machine Learning With Applications*, *4*, 100032. https://doi.org/10.1016/j.mlwa.2021.100032

[6] Mithun, S., Jha, A. K., Sherkhane, U. B., Jaiswar, V., Purandare, N. C., Rangarajan, V., Dekker, A., Puts, S., Bermejo, I., & Wee, L. (2023). Development and validation of deep learning and BERT models for classification of lung cancer radiology reports. *Informatics in Medicine Unlocked*, *40*, 101294. https://doi.org/10.1016/j.imu.2023.101294

[7] Puri, Ridham., & Rizvi, S.T.R. (2024). Fake News Detection: A Comprehensive Study on Analysis and Modern Classification Techniques. 10.13140/RG.2.2.23779.16164.

[8] Tejani, A. S., Ng, Y. S., Xi, Y., Fielding, J. R., Browning, T. G., & Rayan, J. C. (2022). Performance of multiple pretrained BERT models to automate and accelerate data annotation for large datasets. *Radiology Artificial Intelligence*, *4*(4). https://doi.org/10.1148/ryai.220007

[9] *Visual Studio Code Tools for AI - Microsoft Research*. (2017, November 28). Microsoft Research. https://www.microsoft.com/en-us/research/project/visual-studio-code-tools-ai/