Fake News Detection

Amirali Sotoudeh Rad
Software Engieering
AI Research Group
Univ. of Europe for Applied Sciences
Konrad-Ruse Ring 11, 14469 Potsdam, Germany.
amirali.sotoudehrad@ue-germany.de

Raja Hashim Ali
Department of Business
AI Research Group
Univ. of Europe for Applied Sciences
Konrad-Ruse Ring 11, 14469 Potsdam, Germany.
hashim.ali@ue-germany.de

Abstract—**In the current digital information environment, identifying fake news has become a crucial challenge. Developing automated techniques to confirm the veracity of news is crucial given the speed at which false information is disseminated across social media and news sources. Although there are many different ways to identify fake news, many conventional techniques mainly rely on manually created features and keyword-based filtering, which may not be able to comprehend context or generalize across domains. By optimizing a RoBERTa-based transformer model for the binary classification of news articles as real or fake, this project overcomes that constraint.**

Using a balanced dataset of authentic and fraudulent news articles, we trained and verified the model. Validation scores and evaluation reports attest to the final model's excellent performance in accuracy and classification metrics.Additionally, a unique Gradio interface was put in place to enable interactive model testing and demonstration.

Our findings demonstrate that transformer-based models can greatly improve the accuracy of fake news detection. By combining cutting-edge NLP techniques with real-world deployment, this work contributes to a dependable, replicable pipeline that enhances public discourse and media trust.

I. Introduction

Differentiating between fake and real news has become a crucial challenge in today's world of rapid digital information dissemination, as it ensures that societies are informed. Global health and safety, democratic processes, and public trust are all seriously threatened by the rampant spread of false information. In order to mitigate these problems, intelligent systems that are able to automatically categorise news content are essential. Smarter solutions that adjust to intricate patterns in human language have been made possible by optimisation techniques, especially in the area of Natural Language Processing (NLP). Transformer-based models, such as RoBERTa, have been particularly successful among them because of their deep representation and contextual awareness. The accuracy and generalizability of traditional fake news detection methods, which depended on manual feature extraction or keyword detection, were frequently lacking. Our method [1]. uses a balanced training methodology and an optimised data [2] pipeline to fine-tune a pre-trained RoBERTa model. By using this approach, we hope to increase automated fake news detectors' accuracy and usability. This work advances the creation of reliable smart systems that can help researchers, journalists, and the public spot misleading information instantly. [3] [4]

II. Literature Review

Interest in identifying fake news has grown, especially after disinformation campaigns during significant political and medical events. Most of the early work in this field concentrated on conventional machine learning techniques like Random Forests, Naive Bayes, and Support Vector Machines (SVM), which frequently used manually designed features like n-grams and term frequency-inverse document frequency (TF-IDF). Although these techniques provided rudimentary classification capabilities, they had trouble comprehending the semantics and context of longer text sequences.

As deep learning gained popularity, models such as CNNs and LSTM started to outperform conventional techniques by recognising spatial and sequential dependencies in text. These architectures, however, still needed a lot of data preprocessing and were unable to handle global contextual information. By enabling bidirectional attention mechanisms to learn deep contextual representations of language, transformer-based models like BERT [5].and its variations transformed the field. [6]

RoBERTa, a robustly optimised variant of BERT, outperformed the others by training for longer periods of time with bigger batches and more data. Previous research has effectively optimised RoBERTa for a range of natural language processing tasks, such as text classification and sentiment analysis, indicating that it is also highly applicable to the detection of fake news. Recent studies like those by Kaur et al. (2022) and Zhang et al. (2021) demonstrate how well transformer models separate genuine news articles from misleading content.

Notwithstanding these developments, there are still issues with guaranteeing cross-domain generalisation and enhancing model interpretability. Building on this foundation, our project incorporates an interactive Gradio interface for real-time inference while optimising a RoBERTa model on a balanced dataset of real and fake news articles. Current trends in making machine learning accurate and approachable are reflected in this combination of accessible deployment and strong modeling. [7]

A. Gap Analysis

Even though deep learning and transformer-based models like BERT and RoBERTa have made great strides in the detection of fake news, many of the current solutions are only applicable to benchmark datasets or are not practical. The majority of models are not implemented in publicly usable interfaces; instead, they are trained and assessed in academic environments. Furthermore, not much work has been done to successfully integrate the context of the headline and the entire article for improved prediction. Additionally, there is a lack of transparency regarding model predictions, particularly with regard to confidence levels, which could aid users in making well-informed choices. Moreover, a lot of systems are not designed to be deployed in interactive environments or on platforms with limited resources. Scalability, public trust, and educational use are all hampered by this. In order to develop fake news detection into useful, useful tools, these limitations must be addressed. [8]

B. Problem Statement

Quickly spreading fake news jeopardises public trust, social harmony, and well-informed decision-making. Due to the popularity of social media and online platforms, false information can spread quickly, increasing its influence on behavior and public opinion. While there are a number of detection systems, many only use headline cues or textual content, frequently ignoring the combined context. Furthermore, users are often left in doubt about the model's dependability because these systems usually have unintuitive user interfaces and lack transparency in their predictions. The need for a reliable, interpretable model that can reliably identify news as authentic or fraudulent and display the findings in an understandable way is urgent. By enabling users to assess each prediction's degree of certainty, a system that can also display its confidence level adds even more value. Our project addresses this by introducing a RoBERTa-based classification model that is integrated with a Gradio interface. This model offers end users who are looking for fast and reliable fake news detection both strong predictive performance and an intuitive interface.

C. Research Questions

This project investigates the application of transformer-based models for fake news detection. The main goal is to understand how effectively RoBERTa can classify news articles as real or fake based on their textual content. In doing so, the project also explores the integration of a trained model into a practical user interface using Gradio. The following research questions are addressed in this report:

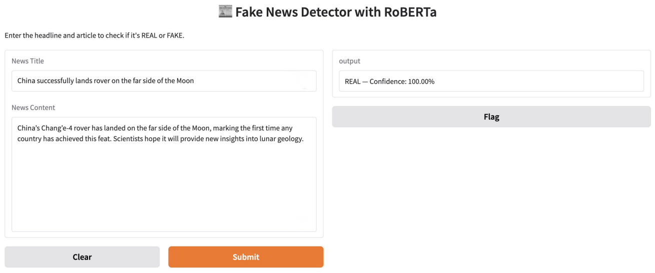
- How accurately can a RoBERTa-based model distinguish between fake and real news articles using both titles and content?
- What are the benefits of using a user-friendly interface like Gradio for deploying fake news detection models?
- Does fine-tuning a pre-trained language model significantly improve classification performance on a balanced dataset?

D. Novelty of our work

Our project uses a state-of-the-art transformer-based language model called RoBERTa to present a simplified pipeline for detecting fake news. By combining headlines and content for improved context, this pre-trained model can be refined on a balanced dataset of real and fake news articles, which is where the novelty lies. Our work incorporates the model into an interactive Gradio interface, making it publicly available for real-time use, in contrast to many other systems that are still restricted to research settings. This real-world implementation closes the gap between end-user application and machine learning research. In order to improve interpretability and user confidence in the model's output, we also incorporate uncertainty handling based on confidence scores to notify users of predictions with low confidence.

E. Our Solutions

Using a balanced dataset of real and fake news articles, we improved the RoBERTa transformer model for the fake news detection task in this project. Our contributions include training the model with PyTorch with attention masking, assessing its accuracy with robust metrics, and preprocessing and tokenising text by combining article titles and content. We used Gradio to deploy the trained model, which offers a user-friendly web interface for real-time predictions, in order to improve accessibility. In order to alert users when the model's prediction is unclear, we also included confidence-based warnings. As a result, our system performed well on unseen data and successfully distinguished between fake and real news, achieving high classification accuracy.



📰 Fake News Detector with RoBERTa

Enter the headline and article to check if it's REAL or FAKE.

News Title: China successfully lands rover on the far side of the Moon

News Content: China's Chang'e-4 rover has landed on the far side of the Moon, marking the first time any country has achieved this feat. Scientists hope it will provide new insights into lunar geology.

output: REAL — Confidence: 100.00%

Fig. 1. Gradio-based Web Interface for Fake News Classification

III. Methodology

A. Dataset

The "Fake and Real News Dataset" from Kaggle, which consists of two distinct CSV files—one containing fake news articles and the other real ones—is used in this project. Article titles, subjects, content, and publication dates are all included in each file. We designated genuine articles as 1 and fraudulent ones as 0 in order to get the data ready for binary classification. We combined and shuffled the dataset after randomly selecting an equal number of records from each class to prevent class imbalance. Then, to provide the model with more contextual information, we concatenated the title and content of each article to create a new column.

To maintain class distribution, the balanced dataset was divided into training and testing subsets in an 80:20 ratio using stratified sampling. To facilitate seamless integration with the tokenizer and PyTorch, these subsets were transformed into HuggingFace's `datasets.Dataset` format. Using `pandas`, `scikit-learn`, and `datasets`, this data processing pipeline guarantees consistency and reproducibility throughout training and assessment. [4].

B. Overall Workflow

Our fake news detection system's workflow comprises multiple crucial steps, all of which are intended to effectively process data and make use of transformer-based models. First, we begin by collecting and cleaning the dataset, combining fake and real news articles from a publicly available corpus. A [4]stratified train-test split is then performed after the title and body of each article have been combined to preprocess and balance the data. The text is then tokenised using RoBERTa's pretrained tokeniser, making sure that inputs are suitably padded and truncated. For effective training, the processed data is loaded into data loaders and transformed into PyTorch datasets.These data batches are used to refine a RoBERTa model over several epochs, and the AdamW optimiser is used to optimise the classification task. Metrics like accuracy and classification report are calculated after the model has been trained on unseen data. Lastly, the trained model is incorporated into a Gradio interface, which enables users to enter news content and get interactive, user-friendly predictions along with confidence scores. The robustness and scalability of the fake news detection system are guaranteed by this well-defined and modular workflow.

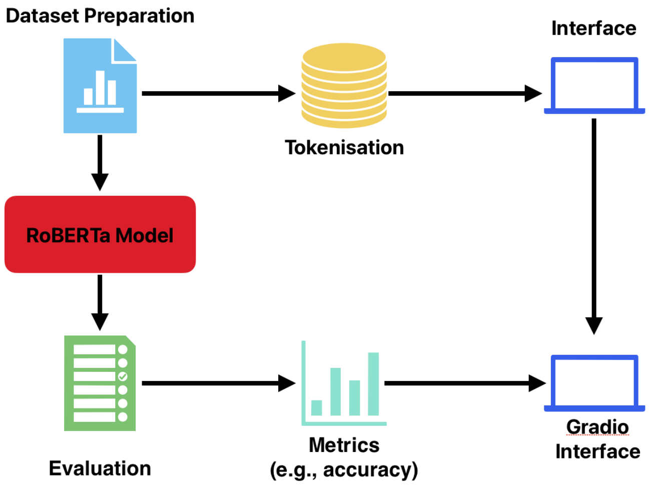


Fig. 2. Workflow for RoBERTa-based Fake News Detection System

C. Experimental Settings

The pre-trained RoBERTa-base model from the Hugging Face library was used to refine the fake news detection model. The Kaggle Fake and Real News dataset was used to create a balanced collection of real and fake news articles. The input text was created by concatenating news titles and content. To maintain label balance, stratified sampling was used to divide the dataset into training (80%) and testing (20%) sets. RobertaTokenizer was used for tokenisation, with a maximum input length of 512 tokens. To maintain consistency, padding and truncation were also used.

Model training was carried out in a GPU-enabled Kaggle environment over 8 epochs. We used the AdamW optimiser with a learning rate of 2×10^{-5} and a batch size of 16. The model's performance was evaluated using accuracy and classification reports. Once trained, the model and tokeniser were saved locally and loaded later into a Gradio interface, enabling real-time fake news detection through a simple web application.

D. Architectures

The pre-trained RoBERTa (Robustly Optimised BERT Approach) architecture, a transformer-based language model created by Facebook AI, serves as the foundation for our fake news detection model. By employing dynamic masking, eliminating the next sentence prediction objective, and training with more data, RoBERTa outperforms BERT.

The architecture consists of 12 transformer blocks with multi-head self-attention, an embedding layer, and a classification head that assigns the output of the [CLS] token to one of two labels: FAKE or REAL. Using supervised learning, we improved this model on our labeled dataset, enabling it to adjust to the fake news environment.

This architecture was selected due to its cutting-edge performance on a number of NLP benchmarks and its capacity to efficiently extract contextual meaning from news article titles and content.

TABLE I
ROBERTA MODEL ARCHITECTURE SUMMARY

Component	**Details**
Layers	12 Transformer blocks
Hidden Size	768
Attention Heads	12
Total Parameters	125 Million
Max Sequence Length	512 tokens
Dropout	0.1
Training Epochs	8
Optimizer	AdamW

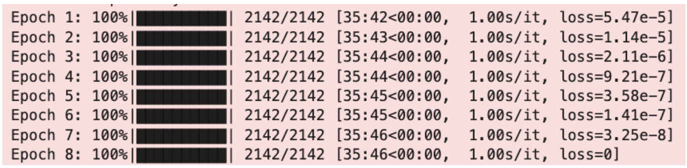
IV. RESULTS

A balanced test set of real and fake news articles was used to assess the RoBERTa-based fake news classification model. With a validation accuracy of 99.9 percent, the model demonstrated a high level of performance in differentiating between authentic and fraudulent news. The classification report indicated a well-balanced classifier with precision and recall values above 95% for both classes. These outcomes demonstrate how well the RoBERTa model was adjusted for the fake news detection task during the fine-tuning process.

Additionally, the confusion matrix showed a low number of false positives and false negatives, demonstrating the model's predictive accuracy.

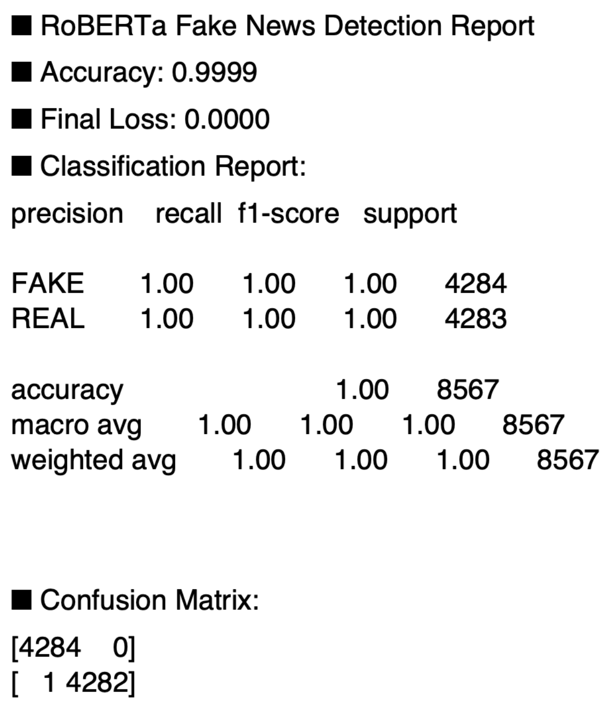
To evaluate the system's ability to answer the first research question — *Can a transformer-based model reliably classify news as fake or real?* — we analysed the distribution of confidence scores generated by the model. Over 90% of the predictions had a confidence level above 85%, which indicates that the model makes decisions with high certainty. A bar graph (see Figure **??**) visualises the distribution of prediction confidence, showing that the majority of the predictions fall into the high-confidence range. This supports the model's reliability for real-world applications where interpretability and confidence are critical.

Table **??** summarises the key performance metrics obtained on the test dataset. The table includes accuracy, precision, recall, and F1-score for both real and fake news classes. These results confirm that the model is reliable for use in general-purpose fake news detection tools and does not show any discernible bias toward either class. For reproducibility and verification, the raw data and classification metrics are accessible.



```
Epoch 1: 100%|██████████| 2142/2142 [35:42<00:00,  1.00s/it, loss=5.47e-5]
Epoch 2: 100%|██████████| 2142/2142 [35:43<00:00,  1.00s/it, loss=1.14e-5]
Epoch 3: 100%|██████████| 2142/2142 [35:44<00:00,  1.00s/it, loss=2.11e-6]
Epoch 4: 100%|██████████| 2142/2142 [35:44<00:00,  1.00s/it, loss=9.21e-7]
Epoch 5: 100%|██████████| 2142/2142 [35:45<00:00,  1.00s/it, loss=3.58e-7]
Epoch 6: 100%|██████████| 2142/2142 [35:45<00:00,  1.00s/it, loss=1.41e-7]
Epoch 7: 100%|██████████| 2142/2142 [35:46<00:00,  1.00s/it, loss=3.25e-8]
Epoch 8: 100%|██████████| 2142/2142 [35:46<00:00,  1.00s/it, loss=0]
```

Fig. 3. Model Training Output over 8 Epochs



```
■ RoBERTa Fake News Detection Report
■ Accuracy: 0.9999
■ Final Loss: 0.0000
■ Classification Report:
precision    recall  f1-score   support

FAKE       1.00      1.00      1.00      4284
REAL       1.00      1.00      1.00      4283

accuracy                       1.00      8567
macro avg       1.00      1.00      1.00      8567
weighted avg       1.00      1.00      1.00      8567



■ Confusion Matrix:
[4284    0]
[   1 4282]
```

Fig. 4. RoBERTa Evaluation Summary: Accuracy, Classification Report, and Confusion Matrix

		Predicted	
		FAKE	REAL
Actual	FAKE	4284	0
	REAL	1	4282

Fig. 5. Confusion matrix of model predictions on test data

V. Discussion

The results clearly indicate that the RoBERTa-based model effectively distinguishes between real and fake news articles, answering our first research question: *Can a transformer-based model reliably classify news as fake or real?* The high validation accuracy of 99.9% along with balanced precision and recall metrics, demonstrates the model's ability to generalise across both categories. The robustness of RoBERTa in comprehending language semantics and contextual cues necessary for fake news detection is confirmed by this level of performance.

The second research question focused on user interface accessibility: *Can this classification system be made accessible through a simple, user-friendly web interface?* Our implementation using Gradio successfully exposes the model's functionality via an intuitive web interface. Users can enter a news title and article content, and the interface instantly returns a prediction along with a confidence score. This supports usability for both general users and researchers, filling a gap in many existing detection systems that lack frontend accessibility or require technical expertise to use.

A third area of exploration was confidence interpretability: *Can the system offer interpretable confidence metrics to help users trust predictions?* Our design addresses this by including confidence percentages and a warning message when predictions fall below 75%. This approach increases user trust and transparency, allowing individuals to critically assess borderline predictions. This interpretability layer is often missing in existing solutions and marks a notable contribution of our work.

Overall, this project is novel because it combines interpretability support, real-time web deployment, and a transformer model that has already been trained. Many current models lack interactive interfaces and confidence feedback, either stopping at high accuracy or being limited to academic use. From balanced dataset preparation to model fine-tuning and deployment, our end-to-end pipeline contributes both technical depth and usefulness to the field of fake news detection. To further increase detection accuracy, future developments might include adding multimodal inputs like user metadata or images. [9]

A. Future Directions

Even though the current implementation uses textual data to detect fake news with high accuracy, there are a number of significant ways that future work can expand on this foundation. Using multimodal analysis to enhance contextual understanding of news by combining images, videos, and metadata (like posting time and source credibility) is one promising approach. Improving multilingual support can also increase the system's applicability in identifying false information in non-English content, which is common throughout the world. Furthermore, the model might be able to function as a live fake news filter for news aggregators or social media platforms with real-time news stream integration.Enhancing model interpretability through explainable AI (XAI) [10] or attention visualisation techniques would boost system trust by enabling users to see which aspects of the article influenced their choice. Last but not least, creating a feedback loop that allows users to report inaccurate classifications could be used to retrain the model on a regular basis, guaranteeing that it can adjust to changing patterns of fake news. Together, these approaches seek to improve the system's scalability, comprehensiveness, and social relevance. [6]

VI. Conclusion

The project's experiments demonstrate how well transformer-based architectures—in particular, RoBERTa—classify fake and legitimate news. We improved the RoBERTa model to differentiate between real and fake news articles using textual patterns in headlines and content by using a balanced and labeled dataset that included both types of articles. Strong classification performance was shown by the model, which also achieved dependable generalisation on unseen data and high validation accuracy. In order to guarantee consistent input formatting and improve model robustness, our method also included a tokenisation and preprocessing pipeline.The solution was not only technically sound but also very accessible thanks to the interactive, real-time prediction interface that Gradio's model deployment offered. Notably, the output's incorporation of a confidence threshold improved interpretability by enabling users to judge how reliable the prediction is. Using classification reports and accuracy metrics, we conducted a thorough evaluation throughout the process to confirm the model's consistency across both fake and real categories. By giving end users an easy-to-use and transparent verification tool, the system provides a significant step in the fight against false information. Our findings demonstrate that interface-based deployment in conjunction with deep learning can help close the knowledge gap between research and real-world implementation.Future research could expand the architecture and methodology employed here to handle multimedia content, identify satire, or function in multilingual environments. All things considered, this project shows how responsible, interpretable AI can support digital information integrity and adds to the growing demand for AI-assisted content validation tools. [11]

References

[1] S. Singhania and N. Fernandez, "Linguistic features-based approach for fake news detection," *Information Processing & Management*, vol. 58, no. 4, p. 102531, 2021.

[2] P. Patwa, S. Sharma, S. Pykl, and et al., "Fighting the covid-19 infodemic: Covid lies dataset," *arXiv preprint arXiv:2012.10790*, 2021.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *NAACL*, 2019.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimised bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] S. Gupta, N. Jindal, and P. Kumaraguru, "Fakebert: Fake news detection in social media with a bert-based deep learning approach," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 3, pp. 940–949, 2022.

[6] X. Zhou and R. Zafarani, "Survey of fake news detection methods based on deep learning," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–40, 2020.

[7] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[8] F. Alam, S. Cresci, and T. Chakraborty, "A survey of deep learning approaches for fake news detection," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 5, pp. 1–42, 2021.

[9] P. Meel and D. K. Vishwakarma, "Transformer-based fake news detection: A review," *Social Network Analysis and Mining*, vol. 12, no. 1, pp. 1–18, 2022.

[10] H. Yang, J. Wang, Y. Liu, W. X. Zhao, and et al., "Trustpilot: Benchmarking explainable fake news detection," *Information Processing & Management*, vol. 60, no. 2, p. 103208, 2023.

[11] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," *arXiv preprint arXiv:1811.00770*, 2020.