

Fake News Detection

CS3008: Deep Learning

Thribhuvan S (EC21B1093)
Chetan Vasista (EC21B1099)
Amar Rohith (EC21B1106)

Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram

April 20, 2025

Overview

1. Data Preparation
2. Vocabulary and Dataset
3. Model Architecture
4. Positional Encoding
5. Training Process
6. Model Saving and Deployment
7. Evaluation Metrics
8. Results Visualization
9. Custom Input Prediction
10. Attention Visualization

Data Preparation

- True.csv, Fake.csv
- Combined 'title' and 'text' fields into a single 'content' column.
- Text cleaning: lowercase conversion, trimming, removal of non-alphanumeric characters.

The image shows two side-by-side Jupyter Notebook cells. The left cell is titled 'True.csv' and displays the first 10 rows of a dataset. The right cell is titled 'Fake.csv' and displays the first 10 rows of a different dataset. Both datasets have columns: title, text, subject, date.

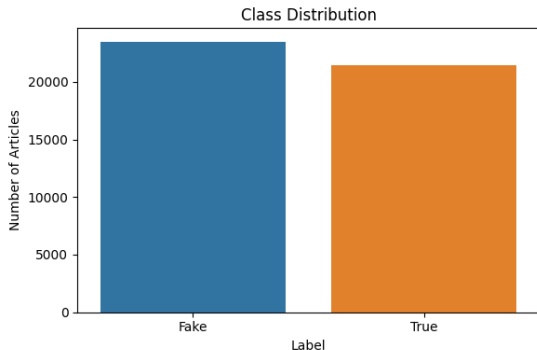
Index	title	text	subject	date
1	"As U.S. budget fight looms, Republicans flip their fiscal script",	"WASHINGTON		
2	U.S. military to accept transgender recruits on Monday: Pentagon,	"WASHINGTON		
3	Senior U.S. Republican senator: 'let Mr. Mueller do his job',	"WASHINGTON (R		
4	FBI Russia probe helped by Australian diplomat tip-off: NYT,	"WASHINGTON (Re		
5	Trump wants Postal Service to charge 'much more' for Amazon shipments,"	"SEAT		
6	'White House, Congress prepare for talks on spending, immigration',	"WEST PA		
7	"Trump says Russia probe will be fair, but timeline unclear: NYT",	"WEST PAL		
8	"Factbox: Trump on Twitter (Dec 29) - Approval rating, Amazon",	"The followi		
9	Trump on Twitter (Dec 28) - Global Warming,"	The following statements were p		
10				

Index	title	text	subject	date
1	Donald Trump Sends Out Embarrassing New Year's Eve Message; This is Distur			
2	Drunk Bragging Trump Staffer Started Russian Collusion Investigation,"	"Hous		
3	Sheriff David Clarke Becomes An Internet Joke For Threatening To Poke Peop			
4	Trump Is So Obsessed He Even Has Obama's Name Coded Into His Website (IMAGI			
5	Pope Francis Just Called Out Donald Trump During His Christmas Speech,"	"Pop		
6	Racist Alabama Cops Brutalize Black Boy While He Is In Handcuffs (GRAPHIC	"		
7	" Fresh Off The Golf Course, Trump Lashes Out At FBI Deputy Director And Jai			
8	" Trump Said Some INSANELY Racist Stuff Inside The Oval Office, And Witness			
9	" Former CIA Director Slams Trump Over UN Bullying, Openly Suggests He's Act			
10				

Sample of the dataset

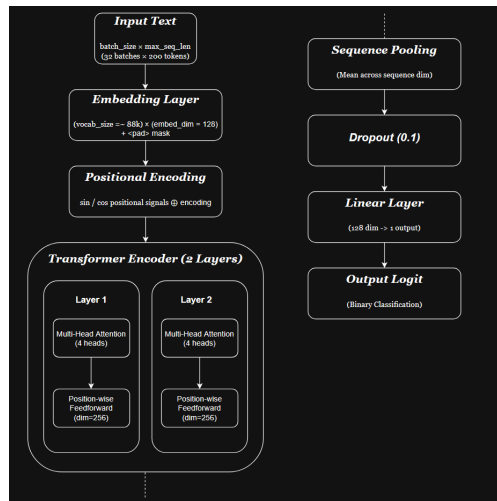
Vocabulary and Dataset

- Fake.csv slightly larger than True.csv
- Built vocabulary from training texts with minimum token frequency threshold of 2.
- Defined custom 'NewsDataset' class to tokenize, numericalize, and pad sequences up to a maximum sequence length of 200.
- Created PyTorch 'DataLoader's for train, validation, and test splits (60-20-20 split).



Model Architecture

- Embedding layer: Maps token indices to 128-dimensional vectors.
- Positional encoding added to embeddings for sequential information.
- Transformer encoder with 2 layers, 4 attention heads, and 256-dimensional feedforward network.
- Classification head: global average pooling + dropout + linear layer to output logit.

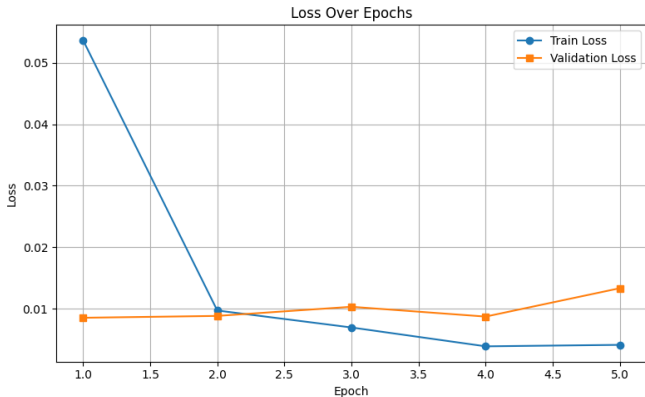


Positional Encoding

- Implemented sinusoidal positional encoding.
- Adds unique position-based vectors to embeddings to retain token order.
- No additional parameters (fixed during training).

Training Process

- Loss: Binary Cross-Entropy with Logits (*BCEWithLogitsLoss*).
- Optimizer: Adam with learning rate 1×10^{-3} .
- Trained for 5 epochs, tracking training and validation loss.



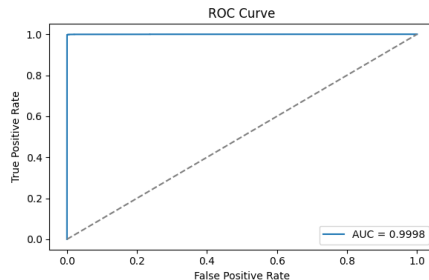
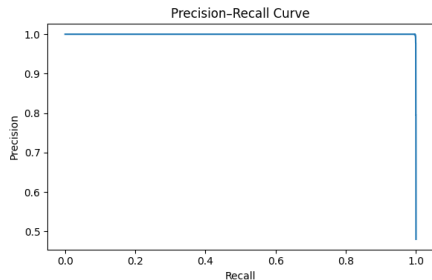
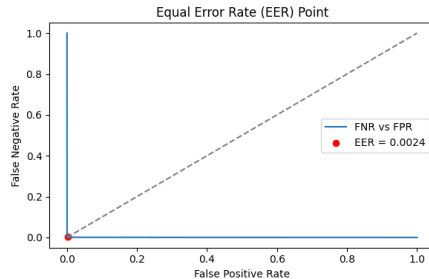
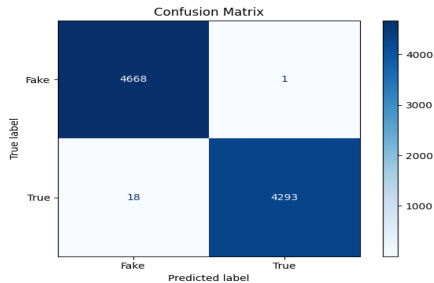
Model Saving and Deployment

- Saved trained model state to 'transformer_fake_news_model.pth'.
- Vocabulary serialized to 'vocab.pkl' using Python 'pickle'.
- Ready for inference: Load model and vocab, run prediction on custom text input.

Evaluation Metrics

- Accuracy: Correct classifications over total samples = 0.9979
- AUC: Area under ROC curve for model discrimination = 0.9998
- Precision: True positives divided by predicted positives = 0.9998
- Equal Error Rate (EER): Point where False Positive Rate equals False Negative Rate = 0.0024 at threshold 0.0131

Results Visualization



Custom Input Prediction

- Cleans and tokenizes input text.
- Pads/truncates to max sequence length.
- Outputs predicted class ('True' or 'Fake') and confidence score.

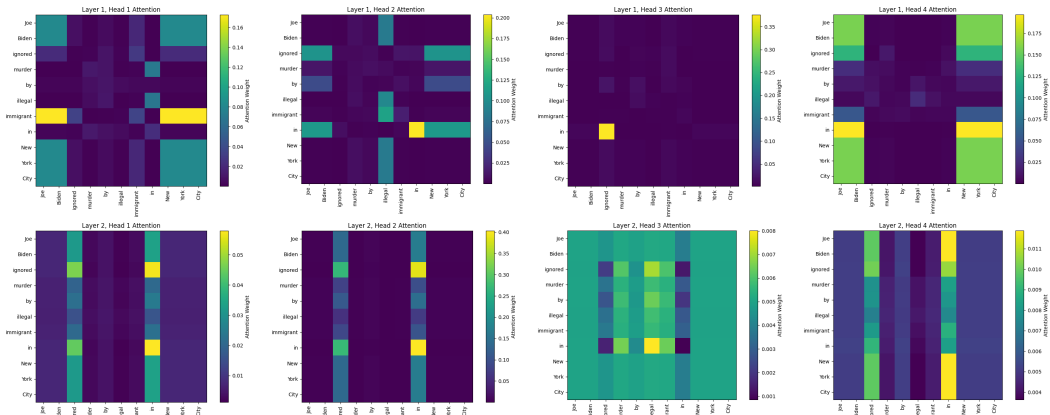
```
89 model = TransformerClassifier(  
90     vocab_size=vocab_size,  
91     embed_dim=128,  
92     nhead=4,  
93     num_encoder_layers=2,  
94     dim_feedforward=256,  
95     dropout=0.1,  
96     max_seq_len=max_seq_len,  
97     num_classes=1  
98 )  
99  
100 device = torch.device("cuda" if torch.cuda.is_available() else "cpu")  
101 model.load_state_dict(torch.load(model_save_path, map_location=device))  
102 model = model.to(device)  
103 model.eval()  
104  
105 if __name__ == "__main__":  
106     custom_text = "Joe Biden ignored murder by illegal immigrant in New York City" # Change this prompt  
107     predicted_label, confidence = predict_custom_input(custom_text, model, vocab, max_seq_len, device)  
108     print("Custom Input Prediction: ")  
109     print("(Only works on US geopolitics and internal politics between 2016 and 2017 due to the nature of the dataset)")  
110     print("-----")  
111     print(f"Input Text      : {custom_text}")  
112     print(f"Predicted Label  : {predicted_label}")  
113     print(f"Confidence Score (0 for Fake, 1 for True): {confidence:.4f}")  
✓ 41s
```

Custom Input Prediction:
(Only works on US geopolitics and internal politics between 2016 and 2017 due to the nature of the dataset)

Input Text : Joe Biden ignored murder by illegal immigrant in New York City
Predicted Label : Fake
Confidence Score (0 for Fake, 1 for True): 0.0001

Attention Visualization

- Patched multi-head attention to extract weight matrices for selected layer and head to interpret model focus on token relationships.



Thank You