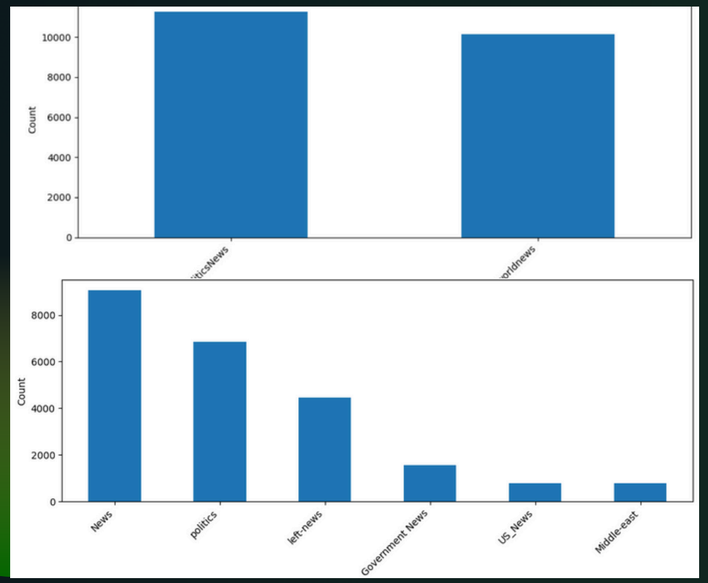
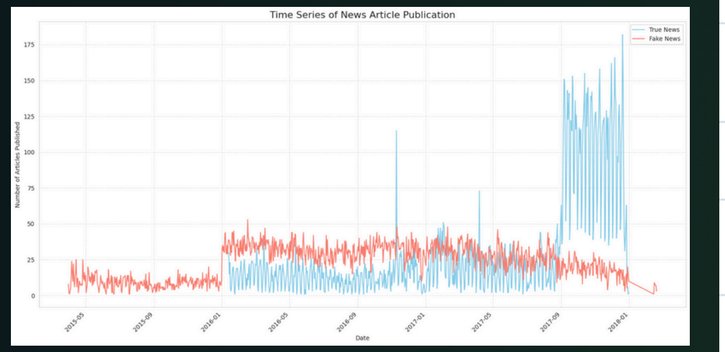
Fake News Detection

CS22B1019 KEERTHANA SARAVANAN

Preprocessing





	title	text	date
count	21417	21417	21417
unique	20826	21192	716
top	os for his administ...	.S. President Dona...	December 20, 2017
freq	14	8	182

	title	text	subject	date
count	23481	23481	23481	23481
unique	17903	17455	6	1681
top	MEDIA IGNORES Time That Bill Clinton FIRED His...		News	May 10, 2017
freq	6	626	9050	46

- Number of True Data : 21,417
- Number of Fake Data : 23,481

- Converted Data column to Day , Month , Year for easy uniform access

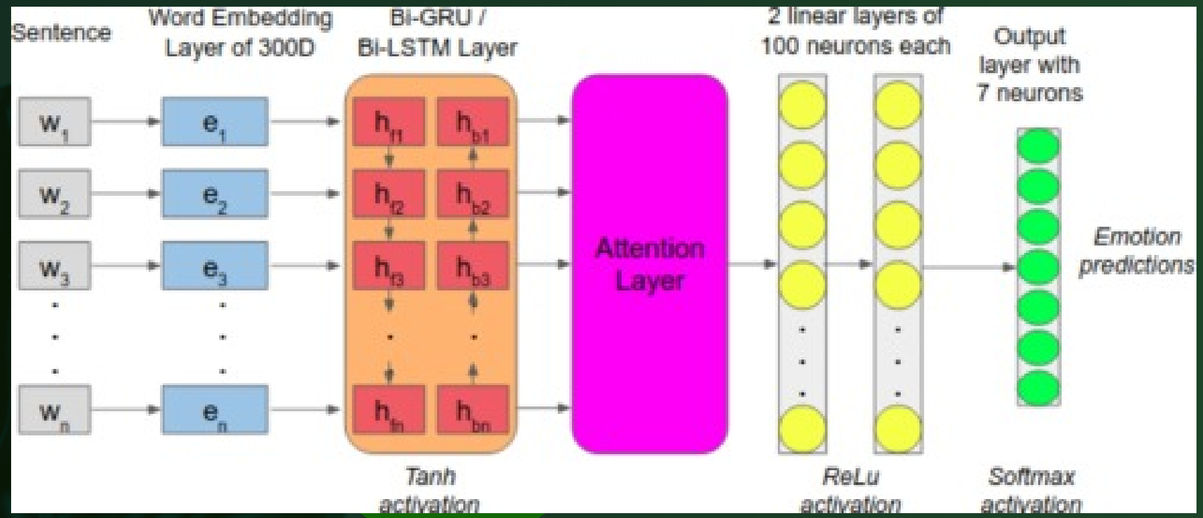
Models

Component	BiLSTM	CNN + BiLSTM	Custom Transformer	DistilBERT + Metadata
Model Flow	Embedding → BiLSTM → Dense	Conv1D → MaxPool → BiLSTM → Dense	Embedding → Transformer ×2 → MeanPool → Dense	Tokenizer → DistilBERT → MetadataConcat → Dense
Embedding Dim	128 (trainable)	128 (trainable)	128 (trainable)	768 (pretrained)
Main Layers	BiLSTM (64 units)	Conv1D(128, kernel=5) → MaxPool → BiLSTM(64)	Transformer Encoder ×2, 4 heads, FFN=256	6-layer Transformer (768-dim, FFN=3072, 12 heads)
Sequence Length	200	200	128	128
Activation Functions	ReLU (hidden), Sigmoid (output)	ReLU (Conv), Sigmoid (output)	ReLU (FFN), Softmax (output)	GELU (internal), Softmax (output)
Loss Function	Binary Crossentropy	Binary Crossentropy	CrossEntropyLoss (for 2-class output)	CrossEntropyLoss
Optimizer	Adam (lr=0.001)	Adam (lr=0.001)	Adam (lr=0.0005)	AdamW (lr=2e-5)
Dropout	0.5	0.5	None	None
EarlyStopping	Yes (patience=3, stopped at epoch 4)	Yes (patience=3, stopped at epoch 4)	None	None
Total Parameters	~220K	~270K	~4.11M	~66M (fully finetuned)

Bi LSTM

WORKFLOW

- Raw text is **cleaned** and **tokenized** into integer sequences.
- Sequences are **padded** to **200 tokens**.
- Words are **embedded** and passed through **BiLSTM** layers for both **forward & backward** directions.
- **Dropout** is used as a **regularization** method to prevent **overfitting**.
- Output goes through a **Dense layer**, then sigmoid to predict **fake/real.**
- **Early stopping (patience = 3)** was used; though set for 10 epochs, training **stopped at 4.**



Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 200, 128)	1,280,000
bidirectional_4 (Bidirectional)	(None, 200, 128)	98,816
dropout_6 (Dropout)	(None, 200, 128)	0
bidirectional_5 (Bidirectional)	(None, 64)	41,216
dropout_7 (Dropout)	(None, 64)	0
dense_4 (Dense)	(None, 16)	1,040
dropout_8 (Dropout)	(None, 16)	0
dense_5 (Dense)	(None, 1)	17

Total params: 4,263,269 (16.26 MB)
Trainable params: 1,421,089 (5.42 MB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 2,842,180 (10.84 MB)

Results

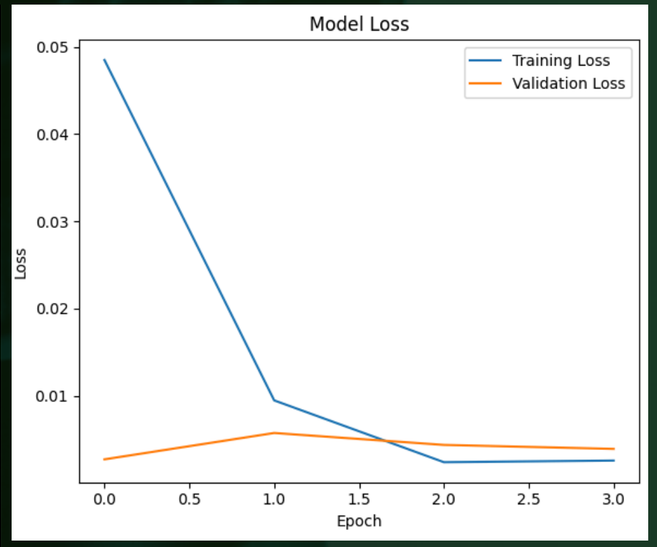
Advantages:

- LSTM reads left to right, but BiLSTM reads both directions, capturing full context — helpful when fake cues are at sentence ends.
- Better at capturing long-term dependencies than standard LSTM, improving accuracy on complex inputs.

Disadvantages:

- Needs custom setup to use metadata (like date/category), unlike Transformers that handle it via special tokens.
- Deep BiLSTMs may face vanishing gradients or slow training — less efficient than Transformers for long texts.

```
===== Performance Metrics =====
           Training  Validation  Testing
Accuracy     0.9992      0.9993   0.9989
Precision    0.9996      0.9991   0.9986
Recall       0.9988      0.9995   0.9991
F1 Score     0.9992      0.9993   0.9988
AUC-ROC      0.9999      1.0000   0.9998
```




```
Training set size: 35918
Validation set size: 4490
Test set size: 4490
```

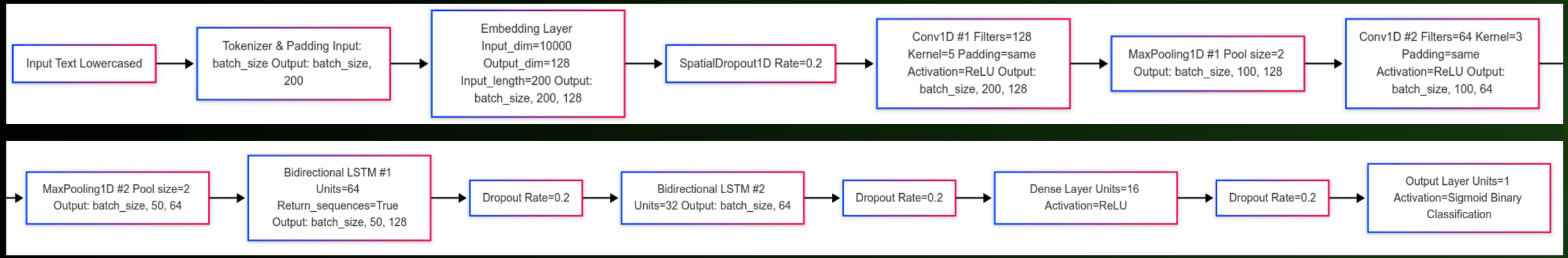
```
=== Cohen's Kappa Scores ==
    Dataset  Cohen's Kappa
   Training         0.9985
Validation         0.9987
    Testing         0.9978
```

CNN BiLSTM

- Input text is **cleaned**, **tokenized**, and **padded** to fixed length (**200 tokens**).
- Tokens are passed through an **embedding layer**.
- A **1D Convolutional layer** (CNN) extracts local n-gram features.
- Outputs are fed into **Bidirectional LSTM** to capture context in both directions.
- Followed by **Dropout** to reduce overfitting.
- **Dense layer** then outputs a sigmoid score for fake/real classification.
- **EarlyStopping** (patience = 3) was used; although trained for 10 epochs, model stopped at epoch 4.

Layer (type)	Output Shape	Param #
embedding_3 (Embedding)	(None, 200, 128)	1,280,000
spatial_dropout1d_1 (SpatialDropout1D)	(None, 200, 128)	0
conv1d_2 (Conv1D)	(None, 200, 128)	82,048
max_pooling1d_2 (MaxPooling1D)	(None, 100, 128)	0
conv1d_3 (Conv1D)	(None, 100, 64)	24,640
max_pooling1d_3 (MaxPooling1D)	(None, 50, 64)	0
bidirectional_6 (Bidirectional)	(None, 50, 128)	66,048
dropout_9 (Dropout)	(None, 50, 128)	0
bidirectional_7 (Bidirectional)	(None, 64)	41,216
dropout_10 (Dropout)	(None, 64)	0
dense_6 (Dense)	(None, 16)	1,040
dropout_11 (Dropout)	(None, 16)	0
dense_7 (Dense)	(None, 1)	17

Total params: 4,485,029 (17.11 MB)
Trainable params: 1,495,009 (5.70 MB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 2,990,020 (11.41 MB)



Results

Advantages

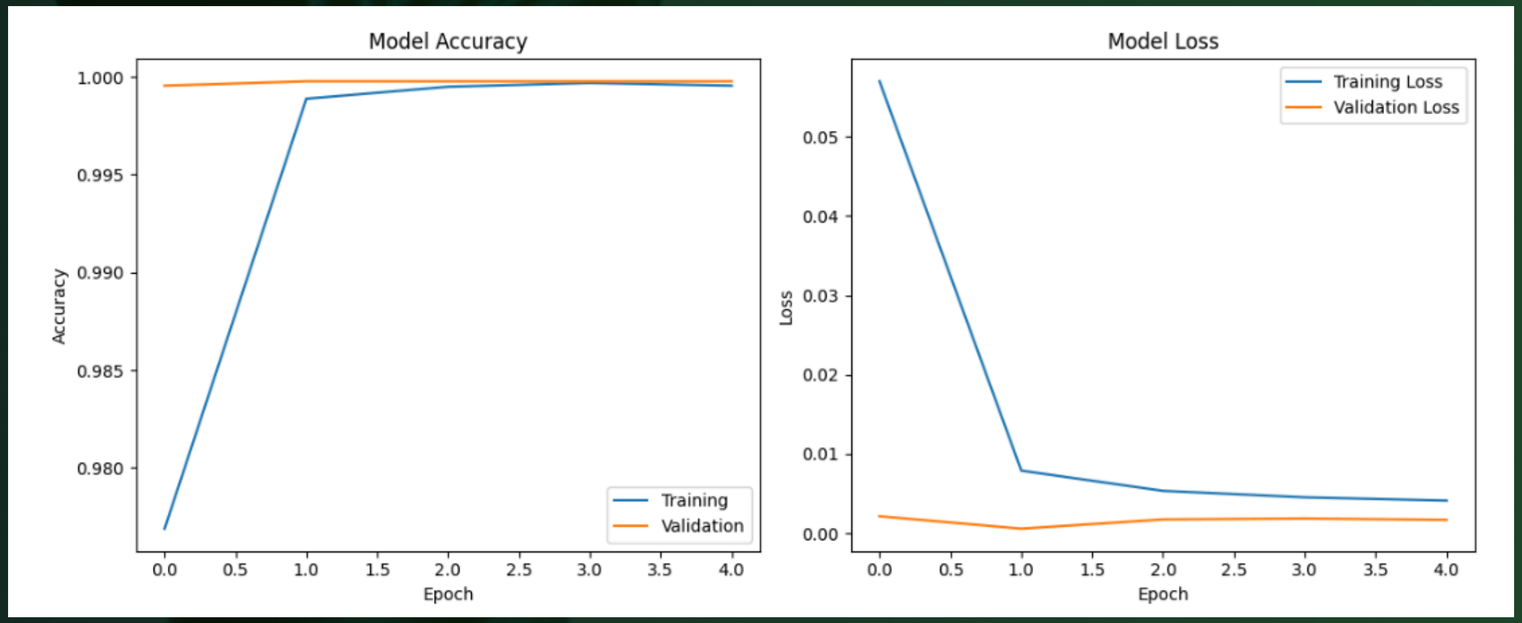
- Captures both local (CNN) and sequential (BiLSTM) patterns, improving accuracy.
- More efficient than pure BiLSTM, as CNN reduces input size before LSTM.

Disadvantages

- Slightly slower training because it stacks CNN + BiLSTM, increasing the number of operations and layers to compute .
- More hyperparameters make tuning complex causes Gradient flow issues which slows or destabilizes training.

```
Training set size: 35918
Validation set size: 4490
Test set size: 4490
```

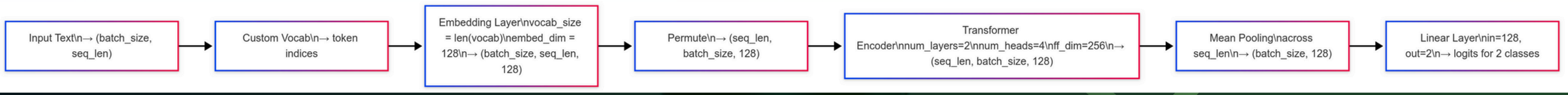
```
===== Performance Metrics =====
                Training  Validation  Testing
Accuracy          0.9997      0.9998   0.9991
Precision         0.9999      1.0000   0.9995
Recall            0.9996      0.9995   0.9986
F1 Score          0.9997      0.9998   0.9991
AUC-ROC           1.0000      1.0000   0.9998
Cohen's Kappa     0.9995      0.9996   0.9982
```



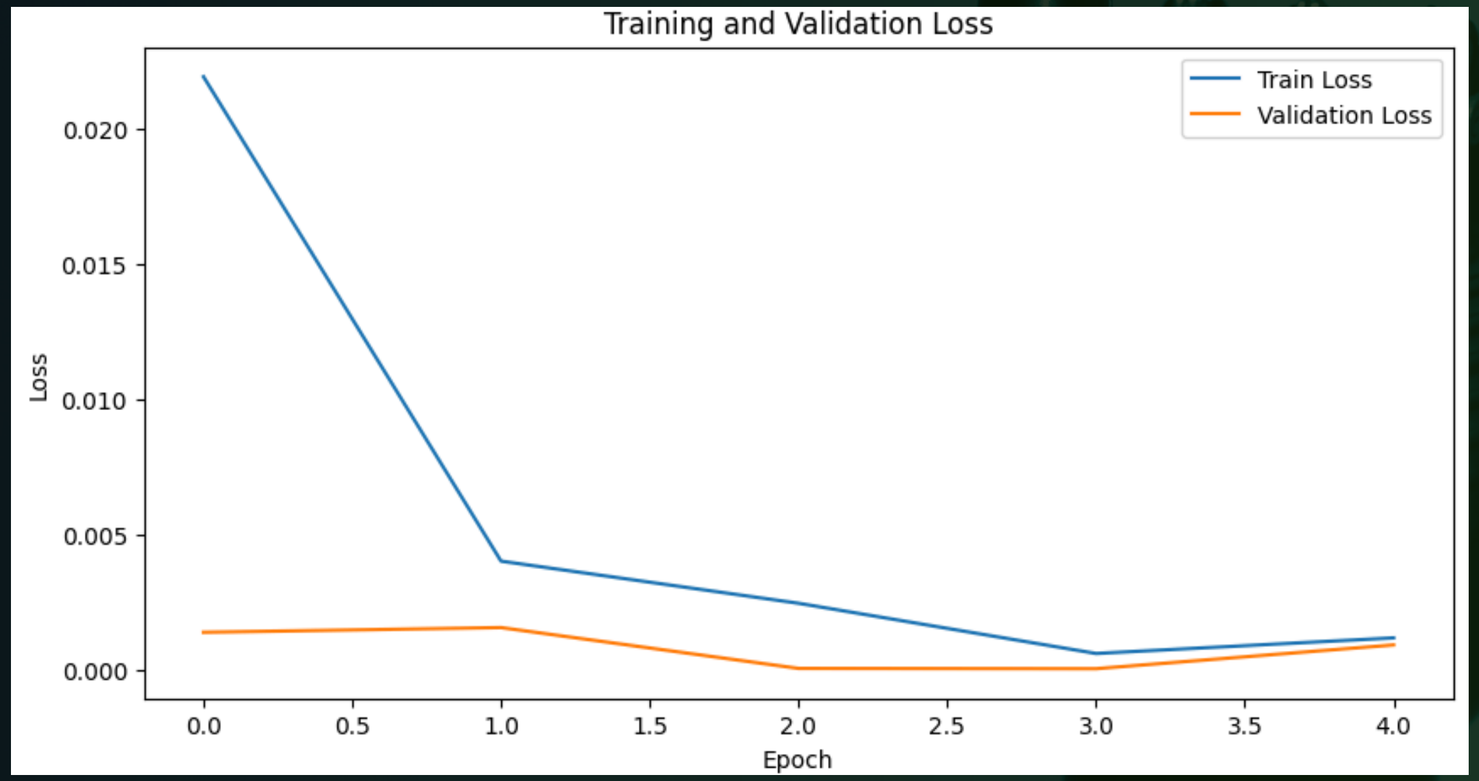
Transformer Based Fake News Classifier

- Title and text are initially combined.
- A custom vocabulary with tokenization and indexing includes <PAD> and <UNK>.
- Input sequences are padded to a fixed length of 128.
- Tokens are embedded into 128-dimensional vectors.
- A 2-layer Transformer Encoder with self-attention captures contextual relationships and word order (implicitly via positional encoding).
- Mean pooling and a final fully connected layer (2 outputs) are used for binary classification with CrossEntropyLoss and Adam optimization.

Layer	Input Size	Output Size	Details
Input (token indices)	(batch, 128)	(batch, 128)	Tokenized & padded news text
Embedding	(batch, 128)	(batch, 128, 128)	Learnable 128-dim embeddings, **~3.84M params**
Positional Encoding	(batch, 128, 128)	(batch, 128, 128)	Learnable, added to embeddings, **~16K params**
Transformer Encoder ×2	(batch, 128, 128)	(batch, 128, 128)	Each with 4-head attention, FFN with 256 units, **~262K params total**
Mean Pooling	(batch, 128, 128)	(batch, 128)	Sequence averaged along time axis
Fully Connected Layer	(batch, 128)	(batch, 2)	Linear layer with **128×2 + 2 = 258 params**
Total Parameters	-	-	**~4.11 million**



Results



Advantages :

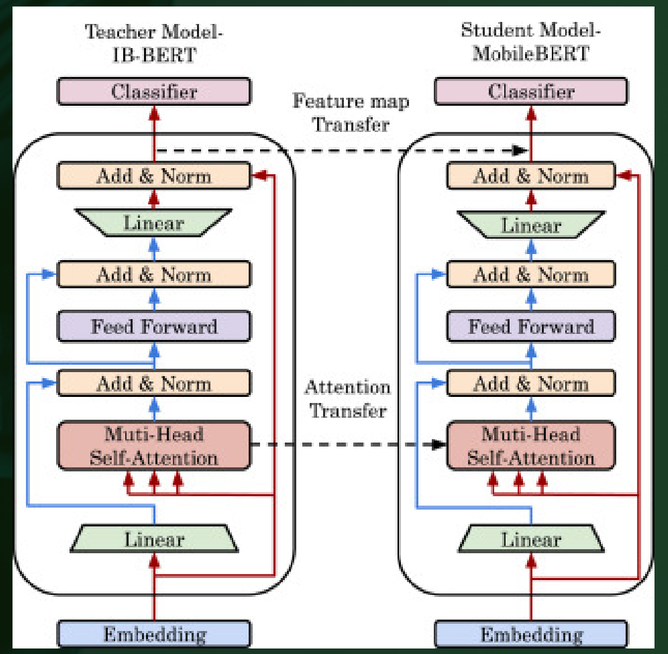
- **Better Parallelism**: Processes all tokens parallely , unlike sequential BiLSTMs, leading to faster training.
- **Handles Long-Range Dependencies effectively**: Directly attends to distant words, overcoming LSTMs' limitations in long-term context.
- Disadvantages :
- Requires **More Data**: Needs larger datasets to **generalize** as well as pretraining helps BiLSTMs with smaller sets.
- **Less effective** for capturing **Local Features**: CNN layers in CNN-BiLSTMs excel at local pattern recognition (e.g., phrases), which Transformers may miss without specific tuning. It instead captures long dependencies very well.

```
Final Metrics:
TRAIN  acc:0.9997 prec:0.9997 rec:0.9997 f1:0.9997 auc:0.9997 coe:0.9989
VAL    acc:0.9997 prec:0.9994 rec:1.0000 f1:0.9997 auc:0.9997 coe:0.9989
TEST   acc:0.9997 prec:1.0000 rec:0.9993 f1:0.9997 auc:0.9997 coe:0.9987
```

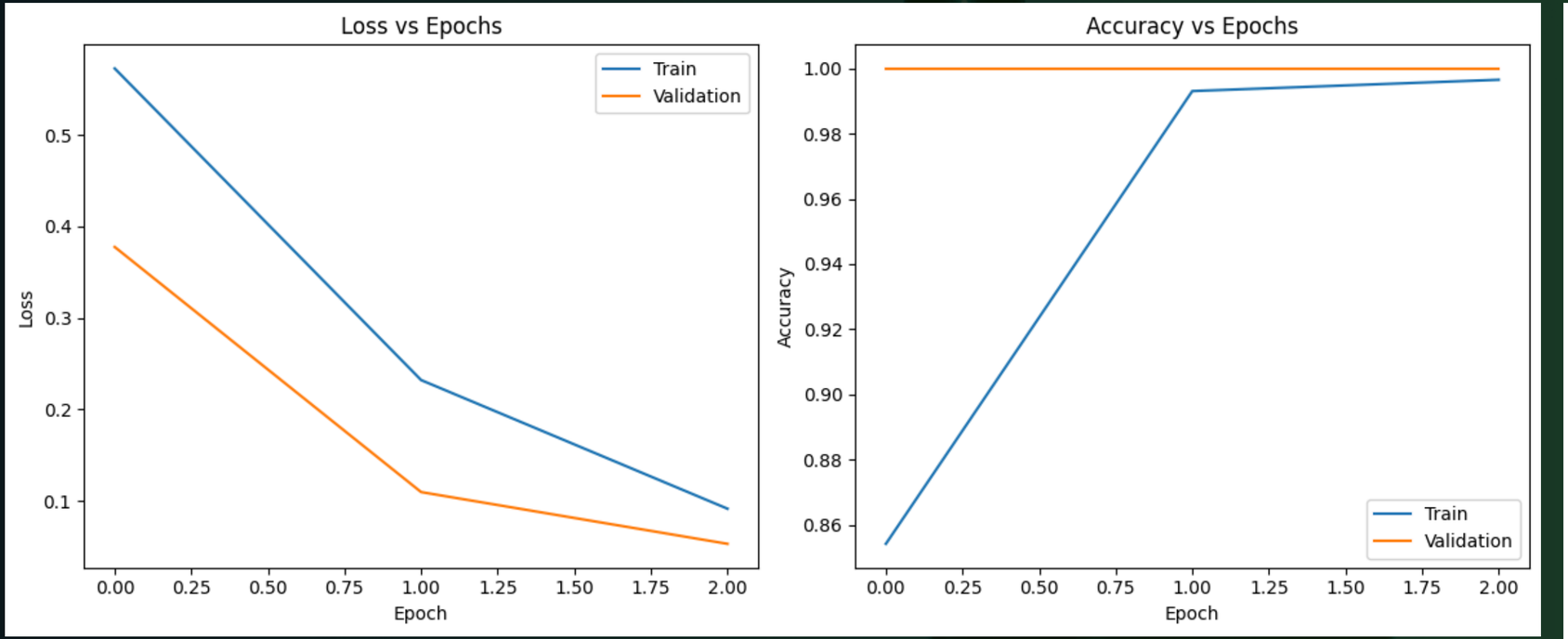
Finetuned Multimodal DistilBERT

- **Text**, **year**, and **category** are cleaned/preprocessed (**year: normalized; category: numerical**).
- **DistilBERT** tokenizer creates **token IDs/masks**; inputs are **padded**.
- A **PyTorch Dataset** (used by **DataLoader**) provides **batches** of **tokenized text, structured data,** and **binary labels**.
- **DistilBERT's 6 transformer layers** produce **contextual embeddings** for each token, including a **[CLS] token** representing the entire sequence; these are **fine-tuned**.
- The **[CLS] token embedding** is combined with **year/category vectors**, then passed through a **ReLU**-activated **dense layer** with **dropout** for classification.
- **End-to-end training**, using **binary cross-entropy loss** and the **AdamW optimizer**, updates all network parameters.

Layer	Input Size	Output Size	Details	# Parameters
Token + Position Embeddings	(batch, 128)	(batch, 128, 768)	Learned embeddings from HuggingFace; includes segment embeddings	**~23.5M**
Transformer Encoder ×6	(batch, 128, 768)	(batch, 128, 768)	Each layer: MHA with 12 heads, FFN(768→3072→768), GELU activation	**~42M**
[CLS] Pooling	(batch, 128, 768)	(batch, 768)	Use [CLS] token representation	0
Metadata Concat	(batch, 768) + m	(batch, 768 + m)	Additional inputs like date, category; small extra parameters if projected	e.g., **~1K** (if proj.)
Classification Layer	(batch, 768 + m)	(batch, 2)	Final dense layer: Linear(768+m → 2)	768×2 + 2 = **~1.5K**



Results



Test Metrics:

Loss: 0.0513

Accuracy: 1.0000

Precision: 1.0000

Recall: 1.0000

F1: 1.0000

AUC-ROC: 1.0000

COE: 1.0000

Advantages

- Multi-modal Feature Fusion: Combines semantic power of DistilBERT with structured metadata like year and category to improve classification.
- Metric-rich Evaluation: Goes beyond accuracy with full metric tracking, enabling deeper model performance understanding.

Disadvantages

- Training Complexity: Integrating structured and unstructured data requires careful tensor alignment and tuning.
- Compute Intensive: Using BERT models and tracking multiple metrics demands more memory and

CONCLUSION

- Traditional models like Bi-LSTM and CNN-BiLSTM:
  - Perform well with limited data.
  - Excel at capturing local and sequential patterns.

- Transformer-based models, especially fine-tuned DistilBERT:
  - Offer better contextual understanding of the text.
  - Can integrate structured metadata (e.g., year, category).
  - Achieve higher accuracy and generalization on real-world data.

- Trade-offs:
  - Require more computational resources and memory.
  - More complex to train and fine-tune.

- Future directions:
  - Explore lightweight transformer variants for faster training.
  - Use domain-specific pretraining to boost performance on news data.
  - Add Explainable AI (XAI) tools to improve model transparency and trust.

REFERENCES

- https://pmc.ncbi.nlm.nih.gov/articles/PMC10800750/
- https://www.researchgate.net/publication/382939584_A_novel_iteration_scheme_with_conjugate_gradient_for_faster_pruning_on_transformer_models

IMAGES

- https://www.sciencedirect.com/science/article/abs/pii/S1574013721000733
- https://paperswithcode.com/method/bilstm
- https://www.sciencedirect.com/topics/computer-science/bidirectional-long-short-term-memory-network
- https://www.sciencedirect.com/topics/computer-science/bidirectional-long-short-term-memory-network

LINK TO CODE :
https://github.com/Keerthana-1024/Fake-News-Detection.git OR
https://www.kaggle.com/code/keerthana1024/fake-news

Thank You!