NLP PROJECT

News Headline Classification: Comparison of Traditional ML and Transformer Models

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Prresented By:

TEAM 9

Sanjith Ganesh(sg2151)

Pranav Senthilkumaran(ps1471)

DATASET - AG News

- AG News dataset from HuggingFace
- 4 categories: World, Sports, Business, Sci/Tech
- Training samples, Test samples: 120,000, 7,600

3. Why This Dataset?

- Headlines are short, information-dense, and real-world
- Ideal for comparing classical ML vs transformer models
- Widely used benchmark to makes results meaningful & comparable

Preprocessing

- Converted text to lowercase for consistent matching
- Applied light normalization (fixing hyphens, removing extra spaces)
- Removed no punctuation since headlines sometimes rely on symbols

No stopword removal

- Even common words (“in”, “on”, “at”) carry positional or contextual meaning
- **Removing stopwords would remove meaningful cues needed for classification**

No stemming or lemmatization

- Lemmatization adds unnecessary computation for very small gain
- Headlines rely on exact word forms.
- Preserving the original tokens helps SVM capture critical noun phrases

KeyVectorization

- **TF-IDF with unigrams + bigrams:**
  Allows the model to capture both individual keywords and meaningful short phrases

- **50,000-word vocabulary:**
  Provides wide coverage of important terms while keeping the feature space computationally manageable.

Baseline Models — Why Two, What They Are, Results

WHY? We trained two baselines to compare probabilistic and margin based linear classifiers, and to set a solid TF-IDF benchmark before evaluating transformers.

Baseline Model 1: Logistic Regression

- **Simple linear classifier** – Provides a clear probabilistic baseline and is easy to understand, making it a good starting point.
- **Tuned multiple C values**- Hyperparameter sweep helps find the best balance between regularization and model complexity.
- **Macro-F1 ≈ 85–86%**- Performs reasonably well but leaves room for improvement, especially compared to stronger linear models.
- Struggles with **overlapping categories**

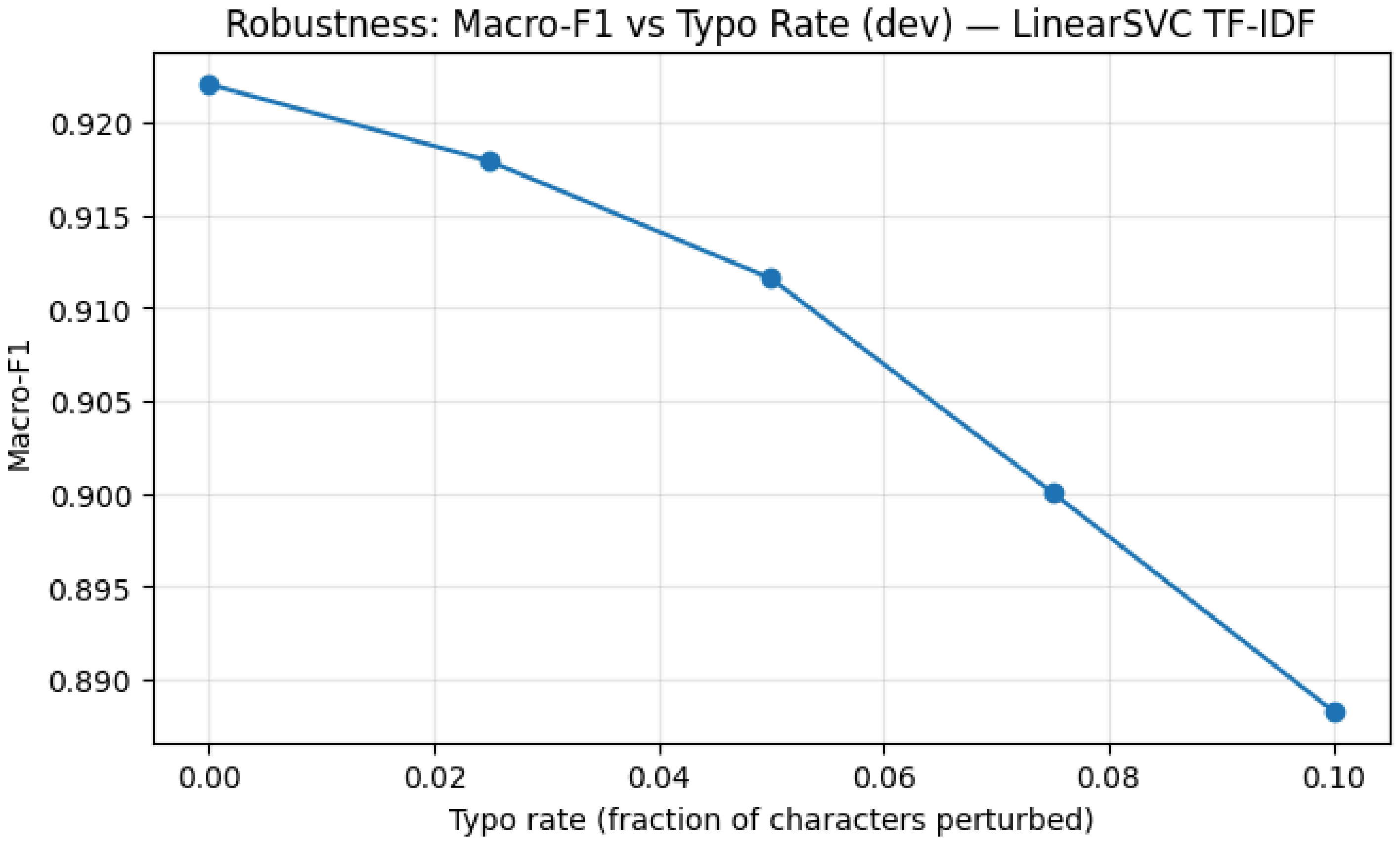
Baseline Model 2: Linear SVM

- **Strong margin-based classifier for text data** – Maximizes class separation, making it highly effective for high-dimensional TF-IDF vectors.
- **Same TF-IDF features + hyperparameter sweep** – Used the same feature space as Logistic Regression, with multiple C values tuned for optimal performance.
- **Macro-F1 ≈ 92%, Accuracy 93–94%** – Significantly stronger performance across all classes, becoming the final chosen classical baseline.

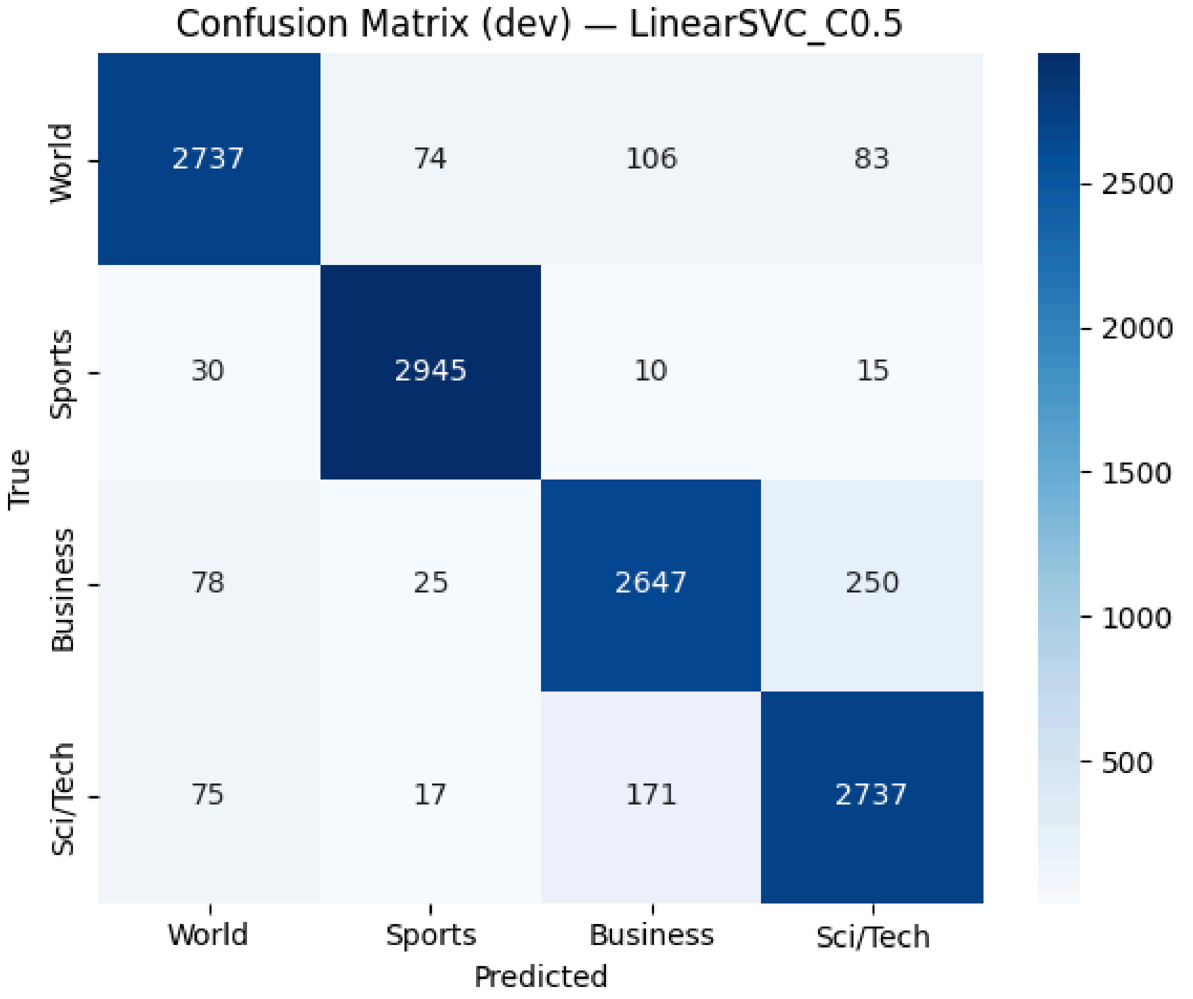
Analysis & Insights

Classification Report (Precision/Recall/F1 Table)

	precision	recall	f1-score	support
World	0.937	0.912	0.925	3000
Sports	0.962	0.982	0.972	3000
Business	0.902	0.882	0.892	3000
Sci/Tech	0.887	0.912	0.900	3000
accuracy			0.922	12000
macro avg	0.922	0.922	0.922	12000
weighted avg	0.922	0.922	0.922	12000



Confusion Matrix (Dev Split)

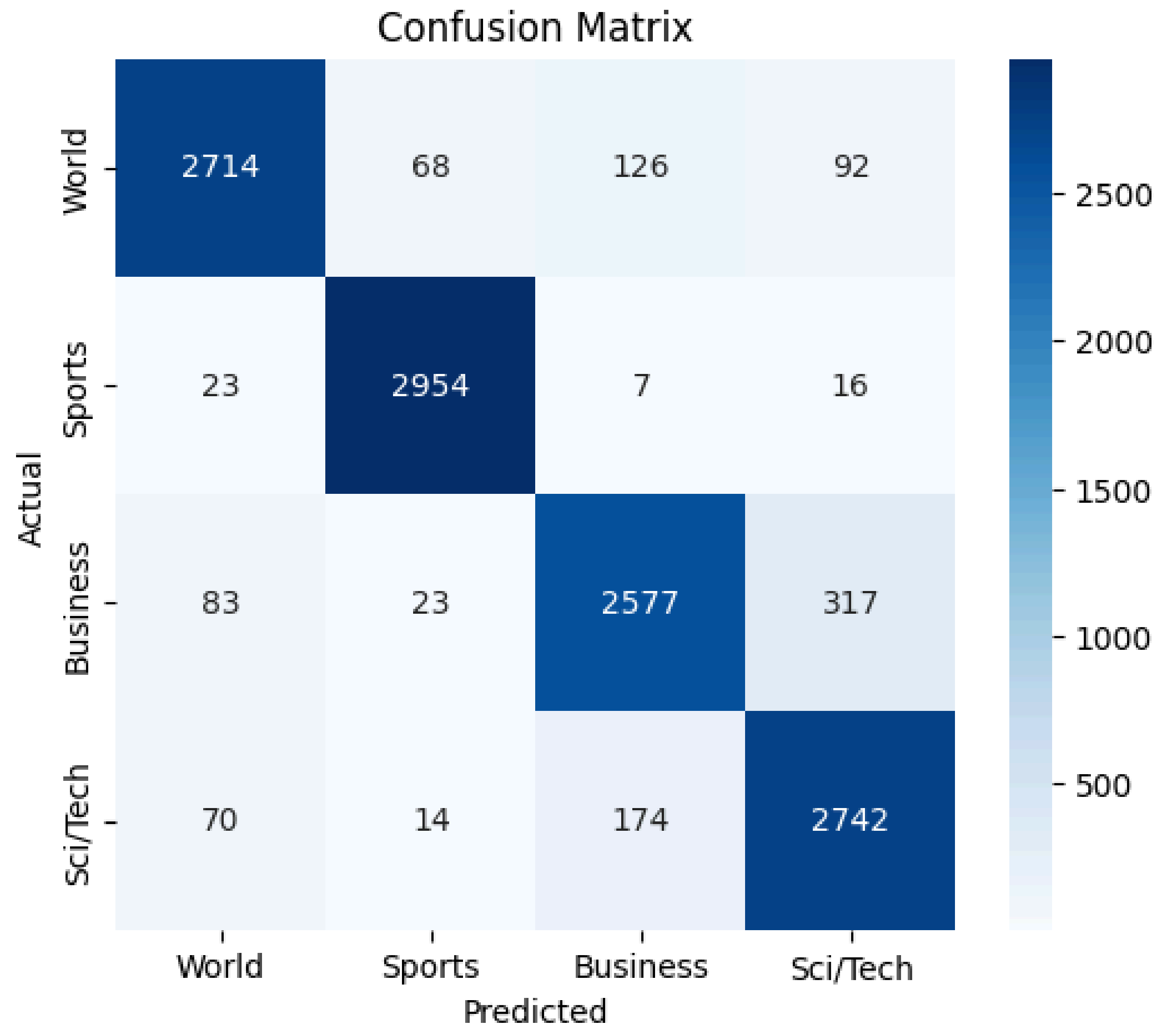


- Other visuals include the class distribution plot, per-class F1 scores, top n-grams per class, a qualitative error table with 10 representative mistakes, top-confidence predictions per class, and correct-prediction samples showing both positive and negative outputs.

DistilBERT - Training Setup & Evaluation Results

- **DistilBERT:** A 40% smaller, 60% faster version of BERT **(Transformer-based Masked language model)** with **bidirectional understanding.**
- **Handles short text well** → Headlines are typically 10–30 words, so full BERT's 512 token capacity is unnecessary. **(max_length=64 enough to capture full content)**

- **Standard Supervised Full Fine-Tuning:** core fine-tuning used in NLP classification tasks
- **Model**: DistilBERT fine-tuned for 4-class news headline classification
- **Hardware**: Google TPU v5e-1 (fast & memory-efficient)
- **Hyperparameters**:
  - Batch size: 32 (train), 64 (eval)
  - Learning rate: 3e-5
  - Epochs: 2
  - Optimizer: **AdamW** (TPU-compatible), bf16 mixed precision for faster TPU training
- **Monitoring**: W&B logs + evaluation every 1,000 steps



Evaluation Results:

- Overall Accuracy: **92%**
- Macro F1: **92%**
- Best class: **Sports (F1 = 0.98)**
- **Bottleneck**:
  Slight confusion between World vs Business, and Business vs Sci/Tech as often share similar language.

	precision	recall	f1-score	support
World	0.94	0.90	0.92	3000
Sports	0.97	0.98	0.98	3000
Business	0.89	0.86	0.88	3000
Sci/Tech	0.87	0.91	0.89	3000
accuracy			0.92	12000
macro avg	0.92	0.92	0.92	12000
weighted avg	0.92	0.92	0.92	12000

Why These Confusions Happen: (Misclassification Patterns)

- Headlines often include economic, political, and tech terms all mixed together.
- **Example**: World & Business share words like **trade, market, economy, global policy**.
- Business & Sci/Tech share **tech-company names, launch, platform, innovation**.

PEFT (Parameter-Efficient Fine-Tuning): LoRA technique

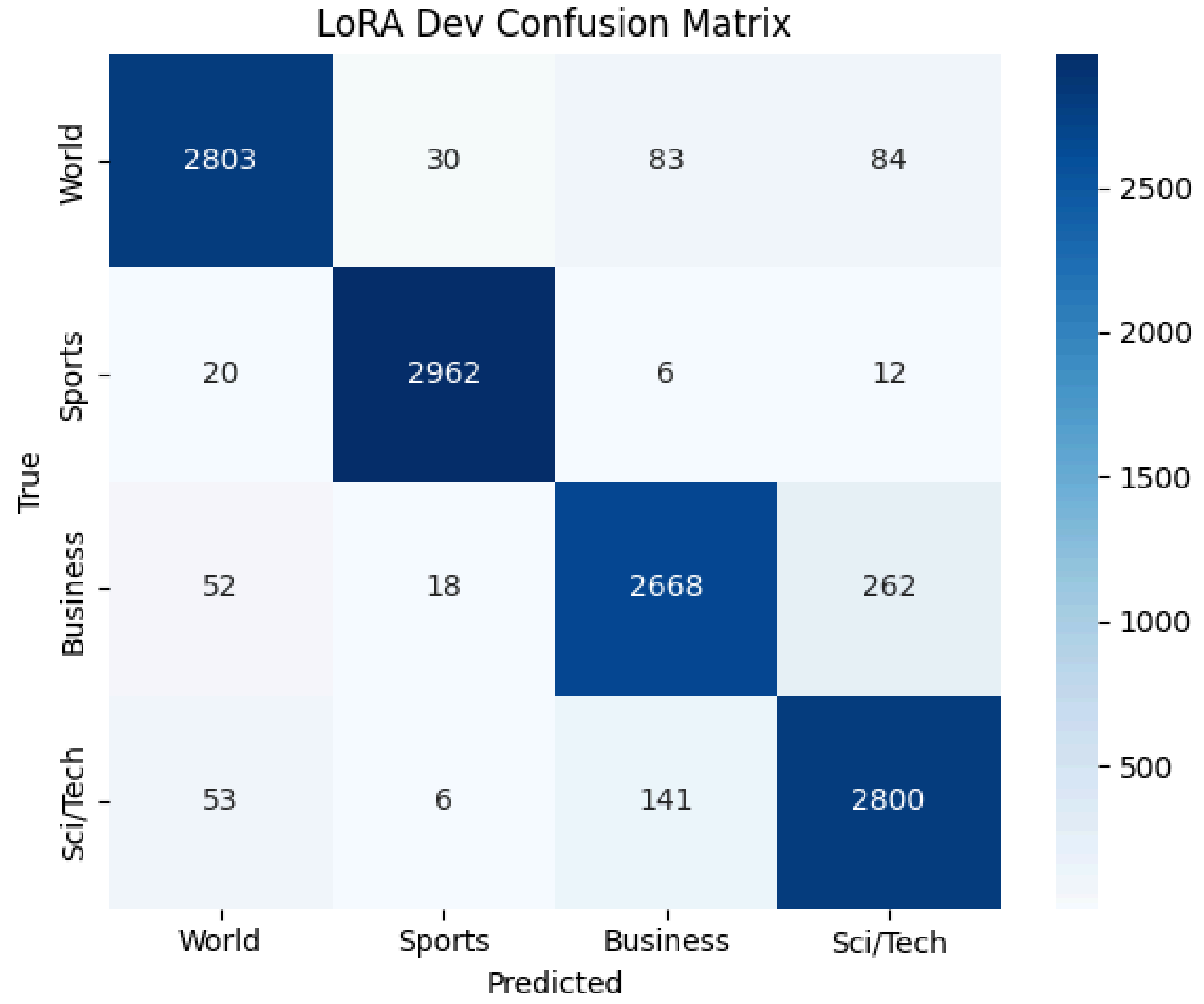
- Instead of updating all model weights (~66M for DistilBERT), LoRA adds **small trainable low-rank matrices** to specific layers (attention layers like query, value).
- Base model weights remain frozen, reducing memory usage and training time.
- Only a **fraction of the parameters are trained** (thousands vs millions), making it fast and scalable, especially on TPUs or GPUs.
- **trainable params: 1,183,492 || all params: 68,140,040 || trainable%: 1.7369**

Why not Key?

- **increases parameter count** (30–40% more LoRA weights) and minimal or no accuracy gain.
- can sometimes reduce stability on smaller datasets like AG News and increases memory usage.
- Generally, add k only for very large models (GPT-J, LLaMA) or generation tasks. For **classification, Q+V is ideal**.

Evaluation Results:

- Overall Accuracy: 94% (**2% increase** than full fine-tuning)
- Macro F1: **94%**
- Performance is stronger than full fine-tuning, showing clear gains.
- Best class: Sports (F1 = 0.98), **Business and Sci/ Tech did much better**.
- Overall, LoRA delivers improved stability, higher consistency across categories, and better generalization.



LoRA Dev Classification Report:

	precision	recall	f1-score	support
World	0.96	0.93	0.95	3000
Sports	0.98	0.99	0.98	3000
Business	0.92	0.89	0.90	3000
Sci/Tech	0.89	0.93	0.91	3000
accuracy			0.94	12000
macro avg	0.94	0.94	0.94	12000
weighted avg	0.94	0.94	0.94	12000

RoBERTa - Model Overview & Evaluation Results

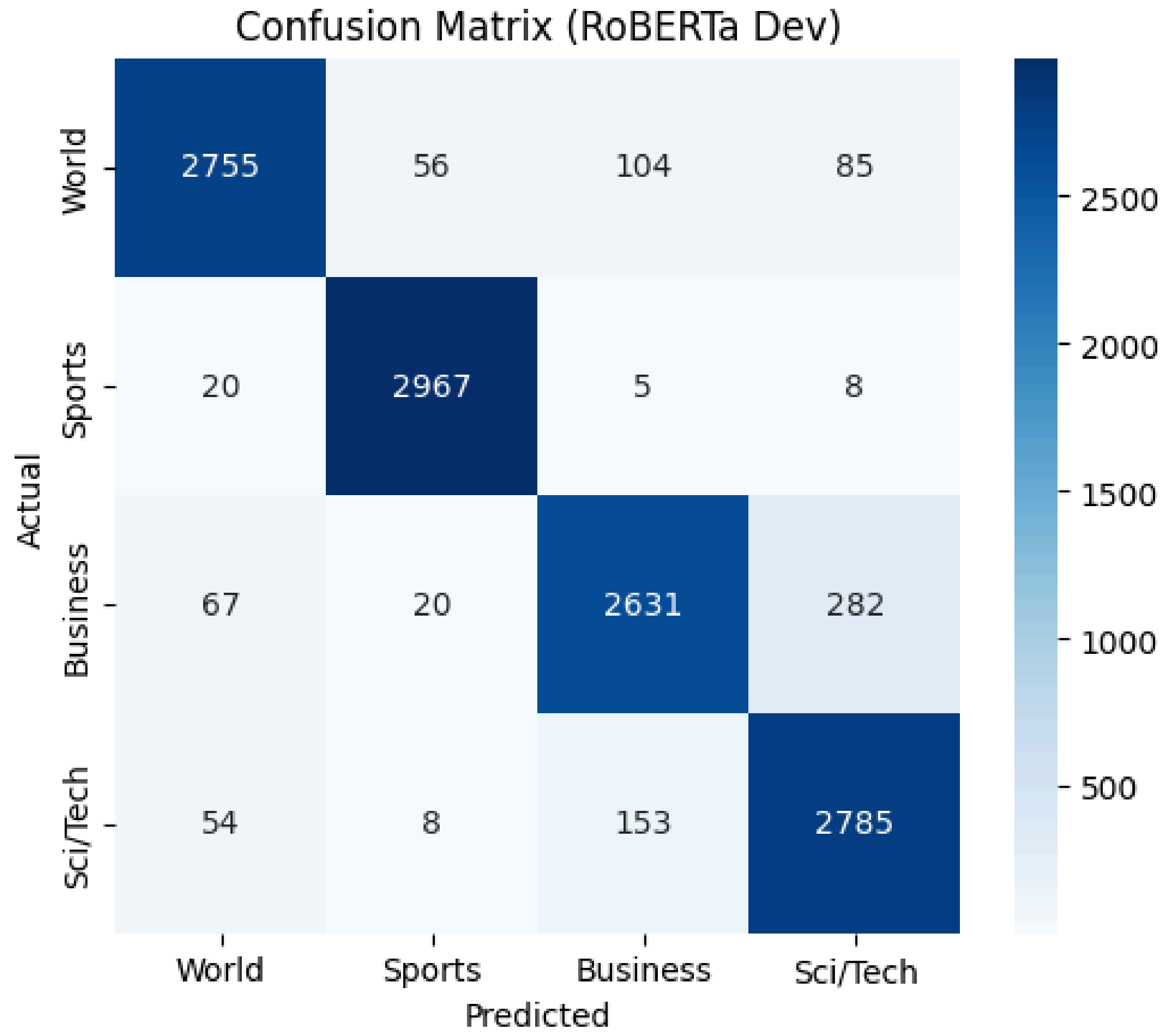
- **Standard Supervised Full Fine-Tuning:** core fine-tuning used in NLP classification tasks.

Why RoBERTa?

- **More powerful than BERT →** Trained longer, on more data, with dynamic masking for stronger language understanding.
- **No Next Sentence Prediction** → Removes an unnecessary objective, leading to more stable and efficient training.
- **Strong generalization** → **Large pretraining corpus** makes it highly effective even with limited labeled data.

Evaluation Results:

- Overall Accuracy: 93% (**1%** better than DistilBERT)
- Macro F1: 93% (**1%** increase)
- Strong and balanced performance across all four categories, **Business and Sci/ Tech did much better than DistilBERT**.
- Best class: Sports (F1 = 0.98)



	precision	recall	f1-score	support
World	0.95	0.92	0.93	3000
Sports	0.97	0.99	0.98	3000
Business	0.91	0.88	0.89	3000
Sci/Tech	0.88	0.93	0.90	3000
accuracy			0.93	12000
macro avg	0.93	0.93	0.93	12000
weighted avg	0.93	0.93	0.93	12000

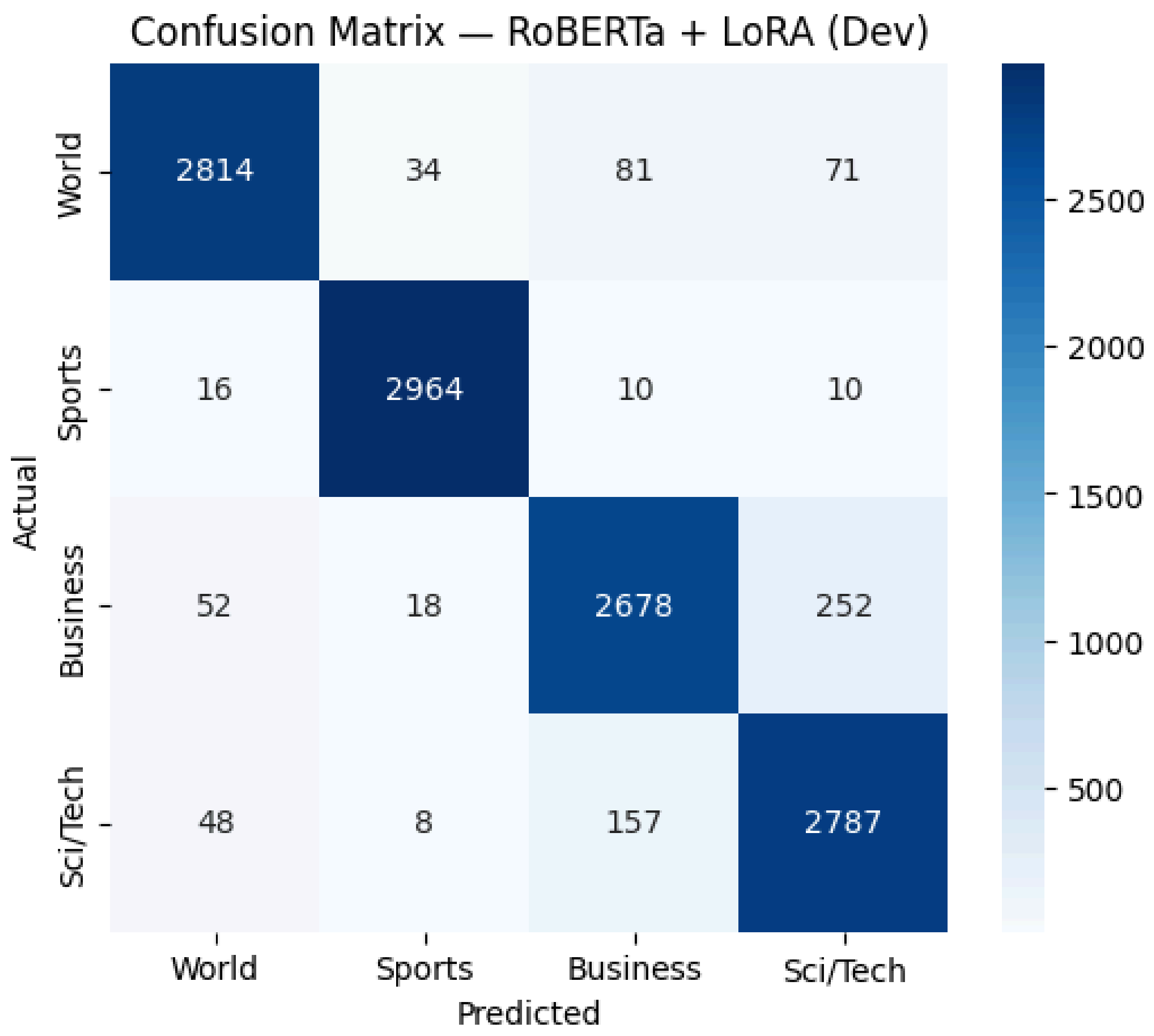
RoBERTa : LoRA technique (PEFT)

- Only **~0.7% of parameters** are trained, making fine-tuning faster.
- **Trainable params: 888580 | Total params: 125537288 | Trainable%: 0.7078%**
- Used a **higher Learning Rate (2e-4)**, **larger effective batch size (64)** via gradient accumulation. **(Hyper-parameter Optimization)**
- Trained for **4 epochs** with LoRA-friendly LR (2e-4) for faster convergence.

Evaluation Results:

- Overall Accuracy: 94% **(93% → 94%)**
- Macro F1: 94% **(93% → 94%)**
- Strongest class: Sports (F1 = 0.98)
- **World and Sci/Tech improved in recall**
- **Business and Sci/ Tech improved in F1**

	precision	recall	f1-score	support
World	0.96	0.94	0.95	3000
Sports	0.98	0.99	0.98	3000
Business	0.92	0.89	0.90	3000
Sci/Tech	0.89	0.93	0.91	3000
accuracy			0.94	12000
macro avg	0.94	0.94	0.94	12000
weighted avg	0.94	0.94	0.94	12000

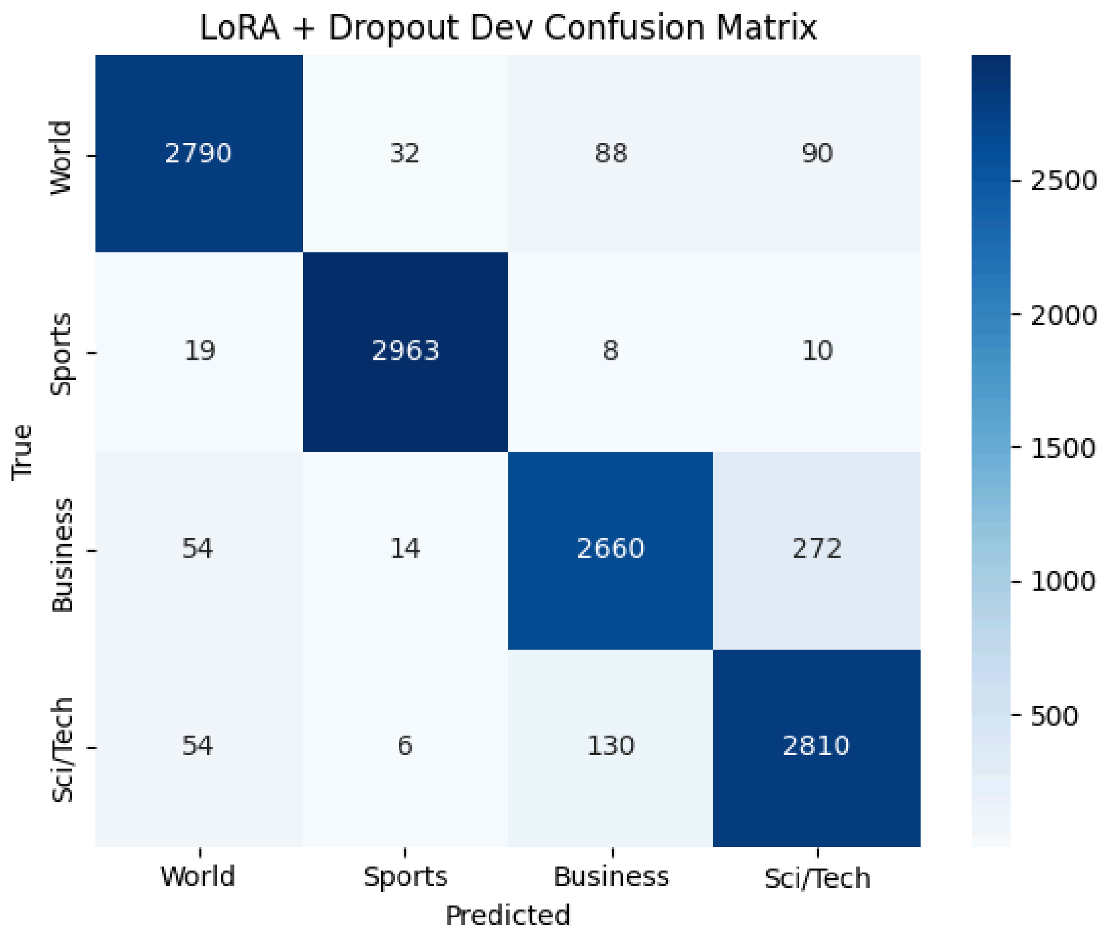


LoRA-Optimized Dropout Regularization

- Added **0.3 dropout to the classifier head** and used **LoRA's internal 0.05 dropout**, giving stronger regularization and helping reduce overfitting. **(main point of dropout regularization)**
- Kept the pretrained encoder layers frozen, allowing the model to rely on RoBERTa's learned representations while only adapting the LoRA layers.
- Maintained the same LoRA setup, and the added dropout improved stability, generalization, and overall accuracy.

DistilBERT

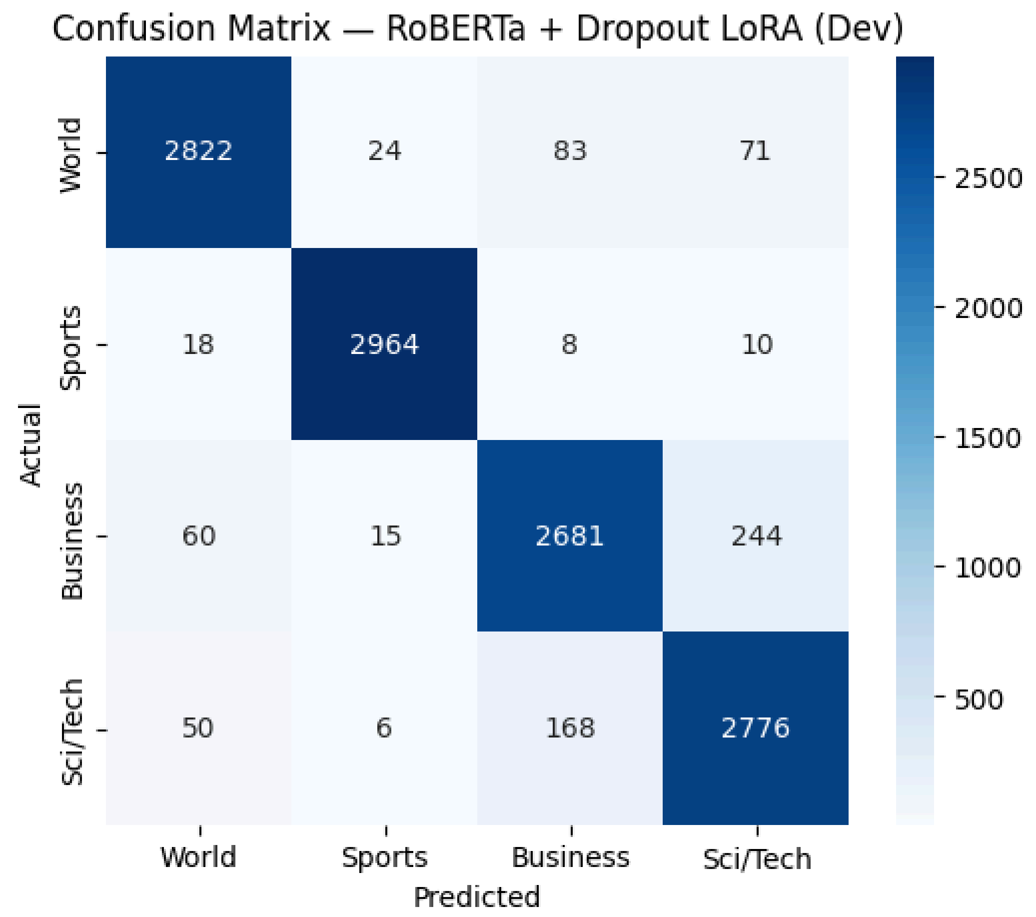
- DistilBERT shows slightly stronger recall pattern in Sci/ Tech
- DistilBERT shows slightly stronger precision patterns in Business.
- DistilBERT offered similiar strong performance with a smaller model.



	precision	recall	f1-score	support
World	0.96	0.93	0.94	3000
Sports	0.98	0.99	0.99	3000
Business	0.92	0.89	0.90	3000
Sci/Tech	0.88	0.94	0.91	3000
accuracy			0.94	12000
macro avg	0.94	0.94	0.94	12000
weighted avg	0.94	0.94	0.94	12000

RoBERTa

- RoBERTa performs slightly better on the World category **(0.95 vs 0.94)**, showing stronger recall.
- **RoBERTa delivered the highest accuracy and F1 score compared to all other models.**



	precision	recall	f1-score	support
World	0.96	0.94	0.95	3000
Sports	0.99	0.99	0.99	3000
Business	0.91	0.89	0.90	3000
Sci/Tech	0.90	0.93	0.91	3000
accuracy			0.94	12000
macro avg	0.94	0.94	0.94	12000
weighted avg	0.94	0.94	0.94	12000

Model Performance Comparison

Model	Accuracy	Macro F1	Trainable Parameters	Training Time
Logistic Regression (TF-IDF)	92%	85–86%	~50k–100k	< 10 sec
Linear SVM (TF-IDF)	93–94%	92%	~50k–100k	< 10 sec
DistilBERT	92%	92%	68M	20–25 min
DistilBERT + LoRA	94%	94%	1.18M (~1.7–1.8% of full params)	≈15 min
DistilBERT + LoRA + Dropout Regularization	94%	94%	1.18M (~1.7–1.8% of full params)	≈15 min
RoBERTa	93%	93%	68M	20–25 min
RoBERTa + LoRA	94%	94%	8.88M (~0.7–0.8% of full params)	20–25 min
RoBERTa + LoRA + Dropout Regularization	94%	94%	1.14M (~1.4–1.5% of full params)	25–30 min

- Even though TF-IDF is strong, it still showed clear mismatches between the four classes, especially Business vs Sci/Tech.
- Full transformer fine-tuning didn't change overall metrics much, but **LoRA + Dropout sharply reduced those pairwise confusions, especially Business ↔ Sci/Tech**.
- So, the headline numbers (94% / 94%) stay the same, but the per-category behavior is much cleaner after LoRA fine-tuning.
- This means **parameter-efficient finetuning doesn't just save compute → it actually fixes the hardest boundary in our dataset**.