

News Headline Classification: Comparison of Traditional ML and Transformer Models

Project Team 9

Pranav Senthilkumaran (ps1471@scarletmail.rutgers.edu)

Master's in Data Science, Department of Statistics

Sanjith Ganesh (sg2151@scarletmail.rutgers.edu)

Master's in Data Science, Department of Statistics

1) Abstract

News headline classification is an important task in Natural Language Processing (NLP) that involves assigning short text sequences to predefined categories. This project addresses a multiclass text classification problem using the AG News dataset, which contains news headlines labeled into four main categories: World, Sports, Business, and Science/Technology.

The study begins by establishing strong baseline models using traditional machine learning techniques, including **Logistic Regression** and **Linear Support Vector Machines (SVM)** with **TF-IDF** feature representations. These baseline approaches provide a meaningful reference point and demonstrate that classical models can achieve competitive performance on short text data.

Building on these baselines, the project explores **encoder-based transformer** architectures, specifically **DistilBERT** and **RoBERTa**, which use bidirectional contextual representations to capture semantic and syntactic information in news headlines. The models are trained using standard supervised fine-tuning, where all model parameters (model weights) are updated for the classification task.

To improve training efficiency and generalization, parameter-efficient fine-tuning (**PEFT**) is implemented using **Low-Rank Adaptation (LoRA)**. The **Q+V (Query and Value)** projection approach is used. The **key projection** is excluded, as it increases the parameter count by approximately 30–40% while providing minimal or no accuracy improvement for classification tasks.

In addition, **dropout regularization** is applied to both the classifier head and LoRA layers to reduce overfitting and improve model stability. **Hyperparameter optimization** is performed by tuning learning rates, batch sizes, number of epochs, and regularization settings to achieve optimal performance.

Experimental results show that **RoBERTa with LoRA and dropout regularization** delivers the highest accuracy and macro F1 score compared to all other models, while training only ~1.4% of the full model parameters. The results demonstrate that parameter-efficient fine-tuning not only reduces computational cost but also improves classification performance, making it a practical and effective approach for real-world NLP problems.

Keywords: TF-IDF, Logistic Regression, Support Vector Machine, Encoder Transformers, DistilBERT, RoBERTa, PEFT, LoRA, Q+V Attention, Key, Dropout Regularization, Hyperparameter Optimization.

2) Introduction

Natural Language Processing (NLP) focuses on enabling computers/ devices to understand and analyze human language. One common NLP task is text classification, where a model assigns a piece of text to one of several predefined categories. This task is widely used in applications such as news categorization, spam detection, sentiment analysis, and question answering.

2.1) Problem Statement

The NLP problem addressed in this project is news headline topic classification. The main objective is to determine the topic of a news headline using only the textual content of the headline. Since headlines are generally short summaries of longer news articles, they contain limited text but still convey the core idea of the story, making the classification task both important and challenging.

The goal of this project is to **find the best model** that can accurately understand the meaning of short text and distinguish between different news topics, even when categories share similar vocabulary. Successful headline classification supports practical applications such as news recommendation systems, content filtering, search indexing, and automated news aggregation platforms.

2.2) Type of Classification

This task is a **multinomial (multiclass) classification** problem and each news headline belongs to exactly one of four topic categories: **World, Sports, Business, Science/Technology**.

Since there are more than two possible output labels, the problem is not binary classification. The model must learn to distinguish between multiple classes that may share similar vocabulary and themes.

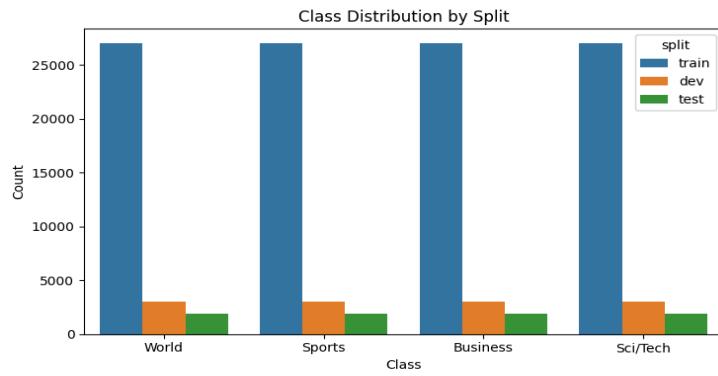
Mathematically, the task is to learn a function: **$f:X \rightarrow Y$** , where **X → set of input news headlines**, and

Y={World,Sports,Business,Sci/Tech} represents the topic labels.

3) Data Collection

The AG News dataset from the **Hugging Face Datasets library** has been used for this project. This is a standard benchmark corpus of news headlines labeled into four classes: **World, Sports, Business, and Sci/Tech**. The original training split contains about 120,000 examples and the test split contains 7,600 examples. The official train/test split of the Hugging Face Dataset was kept unchanged. From this original training split, 10% (12,000 examples) is set aside as a validation/dev set, leaving 108,000 examples for final training.

The Dev split was done using `train_test_split` with `test size = 0.10`, `seed = 42` for reproducibility. And also a stratified by the label column was created so that the four classes remain balanced in both train and dev. Overall, the dataset is very well balanced and each has roughly similar numbers of headlines across train, dev, and test, which was confirmed by plotting class distribution bar charts. This balance makes it straightforward to interpret accuracy and macro-F1.



Preprocessing is intentionally minimal. The AG News dataset is already clean, so the TF-IDF vectorizer and the DistilBERT tokenizer are used to handle tokenization, punctuation, and lowercasing, along with basic whitespace trimming. **Stopwords are not removed and stemming is not applied**, because headlines are short and even common function words can be part of informative phrases.

	A	B	C	D	E
	headline	TRUE	pred	correct	margin
1	NASA To Test Laser Communications With Mars Spacecraft (Sci/Tech	TRUE	Sci/Tech	TRUE	5.419924713
2	UN suspends aid operations in South Darfur after killing of world	World	World	TRUE	5.371419959
3	India's prime minister offers unconditional talks over Kashmir (World	World	World	TRUE	5.217089706
4	Berlusconi talks terrorism with Thai PM (AFP) AFP - Italian Pri	World	World	TRUE	5.182507188
5	Britain's press split over Blair speech (AFP) AFP - Britain's pres	World	World	TRUE	5.145445332
6	Death toll in Darfur region of Sudan could reach 300,000, U.S. World	World	World	TRUE	5.119144046
7	Heinz Profit Meets Forecasts; Shares Rise CHICAGO (Reuter Business	Business	Business	TRUE	4.935498831
8	CommScope Says Costs Rise, Sales to Fall NEW YORK (Re Business	Business	Business	TRUE	4.914406067
9	Cerberus to Buy LNR for \$1.9 Billion NEW YORK (Reuters) - Business	Business	Business	TRUE	4.780574306
10	Is Google the Next Netscape? Microsoft is trailing badly in the :Sci/Tech	Sci/Tech	Sci/Tech	TRUE	4.743127836
11	Use Shuttle To Fix Hubble, NASA Is Told The space shuttle sh:Sci/Tech	Sci/Tech	Sci/Tech	TRUE	4.721498214
12	NBA suspends nine players. Arrest for rest of season NBA on :Sports	Sports	Sports	TRUE	4.711309789
13	Wal-Mart Give Retailers Upbeat '05 Start NEW YORK (Reuter Business	Business	Business	TRUE	4.695069971
14	Hewlett-Packard debuts 'Apple iPod from HP' (MacCentral) MacSci/Tech	Sci/Tech	Sci/Tech	TRUE	4.671930393
15	RIM Intros Souped-Up BlackBerry for Mobile Enterprise (NewsSci/Tech	Sci/Tech	Sci/Tech	TRUE	4.648153694
16	UPDATE 4-Marsh to scrap fees Spitzer faulted, sets reforms Business	Business	Business	TRUE	4.614031117
17	USC Strengthens BCS No. 1 Hold; Oklahoma No. 2 PHILADE Sports	Sports	Sports	TRUE	4.60137425
18	NFL Wrap: Steelers End Bills' Playoff Hopes ORCHARD PAR Sports	Sports	Sports	TRUE	4.573811733
19	Grizzlies Name Fratello Head Coach MEMPHIS, Tenn. (Sport Sports	Sports	Sports	TRUE	4.516451828
20	UPDATE 1-FIFA orders Germany to play World Cup opener FI Sports	Sports	Sports	TRUE	4.474970055
21					

Sample rows from the dataset. Each row corresponds to a single news headline with its ground-truth topic label (TRUE). For analysis, the model’s predicted label (pred) is shown, along with an indicator of whether the prediction is correct and the confidence margin computed from the classifier’s scores.

4) Experimental Setup

A comparison between traditional machine learning models and modern transformer-based neural network models on a real-world text classification task.

4.1) Baseline Models

Two traditional machine learning models were used as baselines. These models help establish a strong reference point before moving to neural network architectures. Both baseline models used the same text representation to ensure a fair comparison.

4.1.1) Baseline 1: Logistic Regression

Logistic Regression was used as the first baseline model to provide a simple and interpretable reference for news headline classification. Each headline was converted into a TF-IDF feature vector, and the model learned a weight for each feature to estimate the probability of a headline belonging to each of the four news categories.

For multi-class classification, the model used a **softmax function** to convert raw scores into class probabilities:

$$P(y = k|x) = \frac{e^{\frac{w_k^T x}{4}}}{\sum_{j=1}^4 e^{\frac{w_j^T x}{4}}}$$

where x represents the TF-IDF feature vector and w_k represents the learned weights for class k.

To control overfitting, L2 regularization was applied. The regularization strength was adjusted using the parameter C, where smaller values of C impose stronger regularization. A systematic hyperparameter sweep over multiple values of C was conducted using the validation set to select the configuration that produced the best macro-F1 score.

The best-performing configuration achieved a **macro-F1 score of approximately 85–86%**. While the model learned clear keyword-based patterns, it struggled when different categories shared similar vocabulary, particularly between **Business** and **Science/Technology**.

4.1.2) Baseline 2: Linear Support Vector Machine

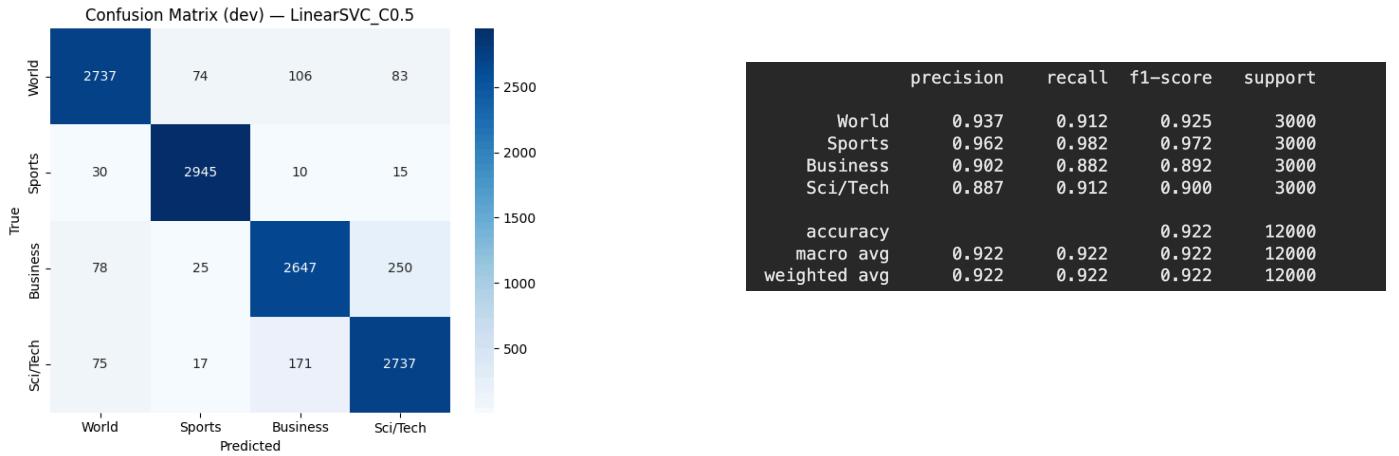
The Linear Support Vector Machine (SVM) was used as the second baseline model due to its strong performance on high-dimensional text data. Like Logistic Regression, the model was trained using TF-IDF representations of news headlines.

The Linear SVM learns a decision boundary that maximizes the margin between classes. Its objective function can be written as:

$$\min_w \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \max(0, 1 - y_i(w^T x_i))$$

where x_i is the TF-IDF vector, y_i is the class label, and C controls the trade-off between margin size and classification error.

The Linear SVM, maximizes the margin between classes while penalizing misclassifications based on the regularization parameter C . This approach allows the model to generalize well, even when classes are not perfectly separable. Unlike Logistic Regression, which models probabilities, SVM focuses on the boundaries between classes, making it particularly effective for text classification tasks with sparse and high-dimensional features like TF-IDF vectors.



Similar to Logistic Regression, a hyperparameter sweep was conducted over different values of C . The Linear SVM achieved a **macro-F1 score of around 92%** and an **accuracy of 92.2%**, showing significantly stronger performance across all classes and becoming the final chosen classical baseline.

4.2) Misclassifications - Bottlenecks

Error analysis shows that misclassifications are not spread evenly across all class pairs. The **most common confusion occurs between World and Business, and between Business and Sci/Tech**. In other words, headlines that are truly World are often predicted as Business and vice versa, and similarly, Business headlines are frequently mistaken for Sci/Tech and the other way around. This pattern appears consistently in both the confusion matrix and the qualitative error tables.

These mistakes usually arise because the categories share very similar vocabulary, especially in short, context-light headlines. World and Business headlines both refer to trade, currencies, markets, and global economic policy, and stories involving technology companies, products, or innovations appear in both Business

and Sci/Tech. As a result, the model sometimes cannot distinguish whether the primary angle of a headline is political/geopolitical, economic, or technological.

Example 1: “EC promises Microsoft anti-trust fight will continue.” is labeled Business, but the model predicts Sci/Tech, likely because the presence of “Microsoft” and “anti-trust” strongly suggests a technology-related story.

Example 2: ‘Roh: Korea needs to be vigilant on won’s slide.’ The actual label is World, but the model predicted Business. Words like ‘won’ and ‘slide’ look very similar to financial or market-related content, so the model pushes it toward Business even though the story is geopolitical. Additional misclassification examples are shown below.

== World → Business (misclassified) ==				== Business → Sci/Tech (misclassified) ==			
	text	actual	predicted		text	actual	predicted
100	Arthritis drug warnings 'ignored' \An arthriti...	World	Business	34	EC promises Microsoft anti-trust fight will co...	Business	Sci/Tech
124	Deal on offshore revenue for Nfid. and N.S. co...	World	Business	52	CA Taps IBM Vet John Swainson As CEO Swainson,...	Business	Sci/Tech
396	Roh: Korea needs to be vigilant on won #39;s s...	World	Business	91	Update 1: Texas Instruments May See Wireless B...	Business	Sci/Tech
432	Business deals a highlight of Chirac #39;s Bei...	World	Business	206	Dolby Labs files for IPO worth up to \\$460 mil...	Business	Sci/Tech
461	Stock Futures Indecisive in Early Going NEW YO...	World	Business	216	Obesity Solution: Nuke It Sharp unveils a new ...	Business	Sci/Tech
== Business → World (misclassified) ==				== Sci/Tech → Business (misclassified) ==			
	text	actual	predicted		text	actual	predicted
26	Japan's troubled Daiei considering seeking aid...	Business	World	126	New Start-Up Breed: Born in the USA, Made in I...	Sci/Tech	Business
131	Losses rise at British Energy after nuclear sh...	Business	World	268	Cybertrust to open for business in 30 days The...	Sci/Tech	Business
149	Colo. Wal-Mart Makes Effort to Unionize (AP) A...	Business	World	311	Brief: eBay snaps up Rent.com for \\$415M EBay ...	Sci/Tech	Business
305	Iranian MPs back investment veto Iran's conser...	Business	World	346	Computer Associates Posts Wider Loss (Reuters)...	Sci/Tech	Business
332	Japanese Central Bank Holds Policy Steady (AP)...	Business	World	414	Symantec Nearly Doubles Quarterly Revenue Syma...	Sci/Tech	Business

4.3) Neural Network Models: Transformer-Based Approaches

Several transformer-based neural network models were used to evaluate whether deep contextual representations improve headline classification. Unlike TF-IDF methods, transformers do not rely only on word frequency. Instead, they learn how words relate to each other based on their **position and surrounding context** and process the entire sentence at once and capture relationships between words.

Hardware Setup: Google TPU v5e-1 using BF16 mixed precision ((fast & memory-efficient)

4.3.1) DistilBERT Model

DistilBERT is **40% smaller** and **60% faster** version of the original BERT model with **bidirectional understanding and masked language modeling**. It was chosen because news headlines are short and do not require very long input sequences. DistilBERT retains most of BERT’s language understanding ability while using fewer parameters, which makes training more efficient.

Each headline was converted into tokens and limited to a maximum length of 64 tokens as headlines are typically 10–30 words, so full BERT’s 512 token capacity was unnecessary.

Model used: Pre-trained **DistilBERT-base-uncased** model obtained from the **HuggingFace Model Hub**

Model link: <https://huggingface.co/distilbert/distilbert-base-uncased>

Inputs: Tokenized news headlines (**input IDs and attention masks**)

Outputs: Probability distribution over four news categories

Activation Function: Softmax in the classification head

Hidden Layers: Pre-trained transformer encoder layers with a single classification head

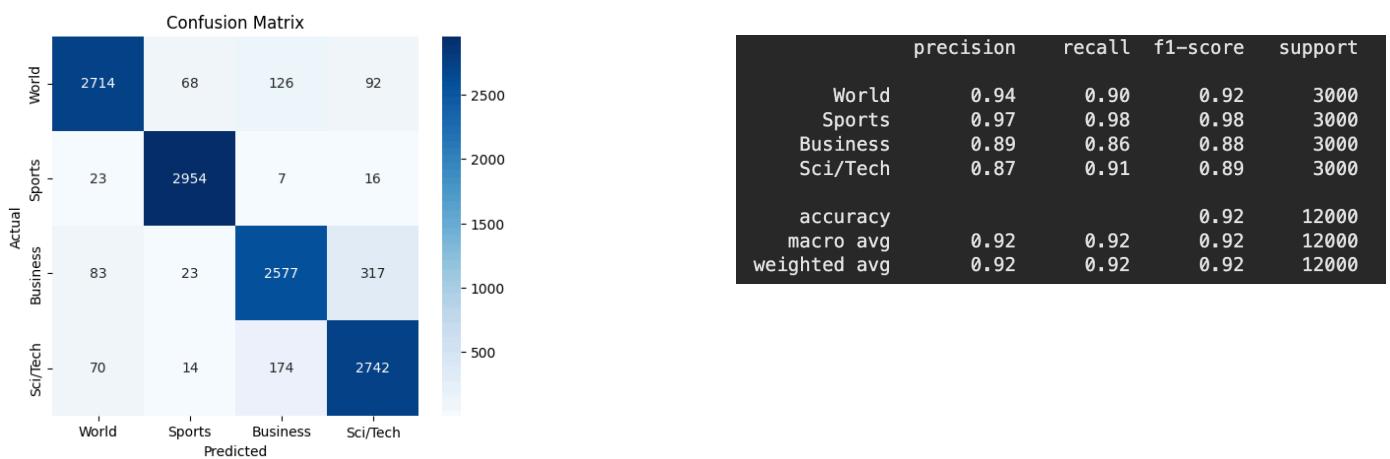
Word Embeddings: Pre-trained contextual embeddings from DistilBERT

4.3.1.1) Standard Fine-Tuning: Full Fine-Tuning

In the standard fine-tuning approach, a pre-trained DistilBERT model was adapted for updating all model parameters during training. Each news headline was first tokenized and converted into input IDs and attention masks. These inputs were passed through the pre-trained transformer encoder layers, which produce contextual word embeddings that capture the meaning of each word based on its surrounding context. A single classification head was added on top of the encoder to map the learned representations to the four output classes.

The classification head produced a set of logits corresponding to the four news categories: World, Sports, Business, and Science/Technology. A softmax activation function was applied to convert these logits into a probability distribution, where each value represents the model’s confidence for a given category.

During training, all DistilBERT parameters, including embedding, attention, and feed-forward layers were fully updated. The model was trained for two epochs using the AdamW optimizer with a learning rate of 3e-5 and a weight decay of 0.01, with a training batch size of 32 and an evaluation batch size of 64. The total training time was approximately 20–25 minutes, and the model contained around 68 million trainable parameters.



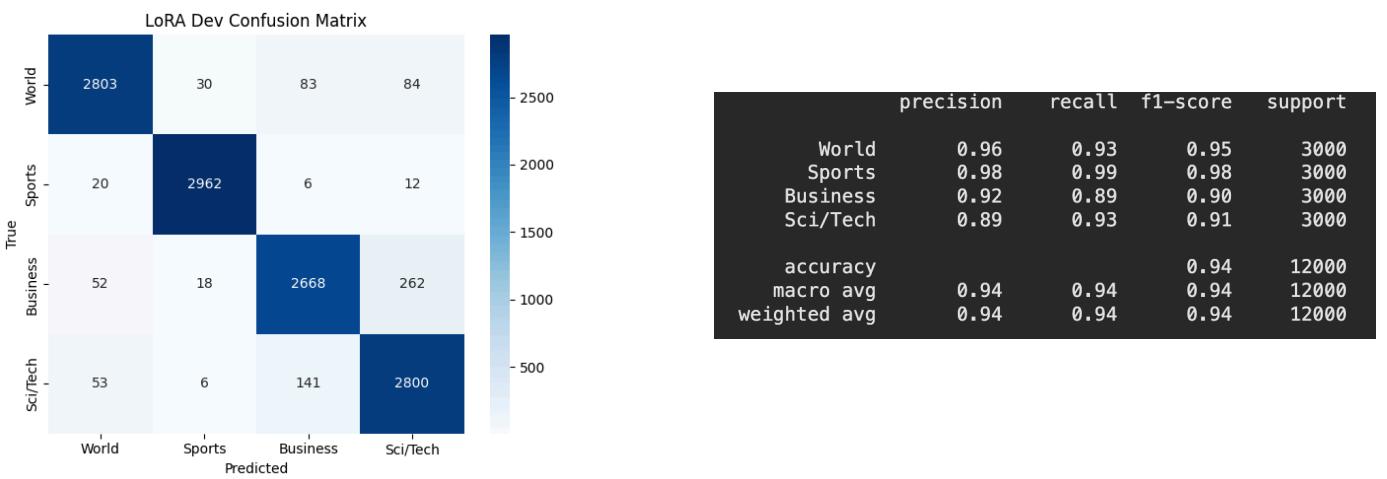
After training, the fully fine-tuned DistilBERT model achieved an **overall accuracy of 92%** and a **macro-averaged F1 score of 92%**. The Sports category showed the strongest performance, as sports-related headlines contain clear and distinctive keywords. However, some confusion remained between World and Business, and between Business and Science/Technology.

4.3.1.2) Fine-tuning 2: Parameter Efficient Fine-tuning (LoRA)

To reduce computational cost and improve training, a parameter-efficient fine-tuning approach using **LoRA (Low-Rank Adaptation)** was applied to DistilBERT. Instead of updating all model parameters, LoRA introduces small trainable low-rank matrices into selected attention components while keeping the original pre-trained weights frozen. In this setup, LoRA was applied to the **query and value (Q+V) attention layers**, as these layers are most effective for classification tasks.

Each news headline was processed using the same tokenization strategy as full fine-tuning, producing input IDs and attention masks. The frozen DistilBERT encoder generated contextual embeddings, while only the LoRA parameters and the classification head were updated during training. This drastically reduced the number of trainable parameters to approximately **1.18 million**, representing about **1.7% of the full model size**.

Training was performed for four epochs using the AdamW optimizer with a higher learning rate of 2e-4, which is well suited for LoRA-based fine-tuning. A training batch size of 16 was used, with an evaluation batch size of 32, and gradient accumulation was applied to achieve a larger effective batch size. The total training time was approximately 15 minutes, significantly shorter than full fine-tuning.



The LoRA-fine-tuned DistilBERT achieved **94% accuracy** and **94% macro-F1**, outperforming full fine-tuning. Performance improved notably for Business and Science/Technology, where class overlap was previously high. Overall, LoRA delivered **better generalization, higher consistency across categories and stability** while training only a small fraction of model parameters.

4.3.2) RoBERTa Model

RoBERTa is a more powerful transformer model built on top of BERT with improved pretraining strategies. It was trained on a large **corpus**, allowing it to learn more robust language representations. RoBERTa **removes the next sentence prediction** objective used in BERT, which leads to more stable training. Due to its stronger contextual understanding, RoBERTa is well suited for handling subtle differences between news categories.

Similar to DistilBERT, headlines were short in length, so each input was tokenized and truncated to a maximum sequence length of **64 tokens**, which was sufficient to capture the full content of most news headlines.

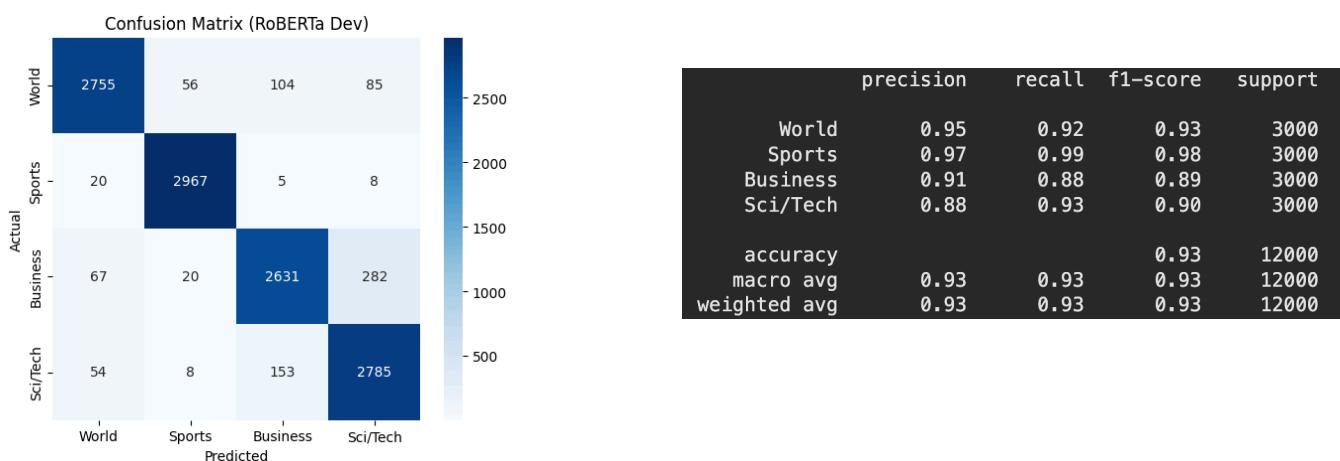
Model used: Pre-trained **RoBERTa-base** model obtained from the **HuggingFace Model Hub**

Model link: <https://huggingface.co/roberta-base>

Same **inputs, outputs, activation functions, hidden layers, and word embeddings** of DistilBERT were used.

4.3.2.1) Standard Fine-Tuning: Full Fine-Tuning

For full fine-tuning, a pre-trained RoBERTa model was used with the same input format, output structure, hidden layer architecture, and tokenization approach as DistilBERT. During training, all RoBERTa parameters, including embeddings, attention, and feed-forward layers, were fully updated. The model was trained for two epochs using the AdamW optimizer with a learning rate of 2e-5 and a weight decay of 0.01. The training batch size was 16, evaluation batch size 64, and training time took around 20-25 minutes. The fully fine-tuned model contained about **125 million trainable parameters**.



The fully fine-tuned RoBERTa model achieved an **overall accuracy of 93%** and a **macro-F1 score of 93%**. Performance improvements were observed across all categories, with more balanced results compared to DistilBERT, especially for World and Science/Technology.

4.3.2.2) Fine-tuning 2: Parameter Efficient Fine-tuning (Lora)

For RoBERTa with LoRA, only a small portion of the model was trainable, **with 0.88 million trainable parameters out of 125 million (0.71%)**. The same input format, output structure, and hidden layer architecture as full fine-tuning were used. Training was done for four epochs with the AdamW optimizer at a higher learning rate of 2e-4, a batch size of 16, evaluation batch size of 32, and gradient accumulation to simulate a larger batch. Total training time was about 20 minutes with better performance.



The RoBERTa model with LoRA achieved an **overall accuracy of 94%** and a **macro-F1 score of 94%**. It showed strong performance across all categories, with the highest precision in Sports (0.98) and balanced results for World, Business, and Sci/Tech, outperforming full fine-tuning.

5) Dropout Regularization

Dropout regularization is a common technique used for both RoBERTa and DistilBERT to reduce overfitting and improve generalization. In this work, a **dropout rate of 0.3** was applied to the classification head, along with **LoRA's internal dropout of 0.05**, to prevent the model from relying too heavily on a small set of features. By randomly turning off neurons during training, dropout encourages the model to learn more robust and well-distributed representations, which is especially important for short, information-dense news headlines.

5.1) Lora-Optimized DistilBERT — Dropout Regularization

The results of Lora optimized DistilBERT - Dropout Regularization are attached below:



5.2) Lora-Optimized RoBERTa — Dropout Regularization

The results of Lora optimized RoBERTa - Dropout Regularization are attached below:



6) Overall 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)	92%	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%	125M	20-25 min
RoBERTa + LoRA	94%	94%	0.88M (~0.7% of full params)	20-25 min
RoBERTa + LoRA + Dropout Regularization	94%	94%	1.77M (~1.4% of full params)	25-30 min

This table summarizes the performance and efficiency of all models we evaluated. The TF-IDF baselines already reach **around 92% accuracy**, but fully fine-tuned DistilBERT and RoBERTa match or slightly improve on this at the cost of training all 68M parameters.

Adding **LoRA and dropout regularization achieves the best accuracy and macro-F1 of about 94%** while updating only a small fraction of the parameters and reducing training time, showing that parameter efficient fine tuning can be both more accurate and more computationally efficient than full fine-tuning.

7) Conclusion

In summary, this project compared traditional machine learning baselines with modern transformer-based models for news headline classification on the AG News dataset. A carefully tuned **TF-IDF + Linear SVM** achieved **strong performance** and remains a competitive, interpretable baseline.

However, encoder transformers such as **DistilBERT** and **RoBERTa**, especially when combined with parameter-efficient fine-tuning via **LoRA and dropout regularization**, achieved the **best overall accuracy and macro-F1** while training only a small fraction of the model parameters.

These findings suggest that, for short-text classification problems like headline categorization, classical linear models are a solid starting point, but transformer models with PEFT provide a practical and scalable path to the best performance. The project highlights how combining strong baselines, careful evaluation, and efficient fine-tuning techniques can lead to models that are both accurate and resource-efficient, making them suitable for real-world NLP applications.

8) Relevant Literature: Paper Summary

Paper: [Research Paper Link](#)

Title: Comparing BERT against traditional Machine Learning text classifications

Authors: González-Carvajal and Garrido-Merchán, 2021

Summary:

The paper systematically evaluates how BERT compares to classic TF-IDF-based machine learning pipelines for text classification. Traditional approaches are outlined in which documents are represented as a bag of words or n-grams and then fed into models such as Logistic Regression, Support Vector Machines, Naive Bayes, or Gradient Boosting. In contrast, BERT is presented as a bidirectional Transformer model that is pre-trained on

large unlabeled text corpora and subsequently fine tuned on specific labeled tasks, using the transfer learning to achieve strong performance with relatively little task specific feature engineering. Experiments are conducted on four different tasks and languages: English IMDB movie review sentiment analysis, English disaster-tweet classification, Portuguese news categorization, and Chinese hotel review sentiment analysis. Across all of these settings, BERT consistently outperforms the best TF-IDF plus classical machine learning baselines, often by several points in accuracy.

For example, on the IMDB sentiment task, BERT reaches an accuracy of about 0.94, compared to roughly 0.90 for the best traditional models such as linear SVM built on TF-IDF features. In the multilingual experiments on Portuguese news and Chinese hotel reviews, BERT again clearly surpasses the best AutoML/gradient boosting baselines, showing that its advantage is robust across different languages and domains. Although BERT is more computationally expensive, the study emphasizes that it can reduce the amount of manual feature engineering and tuning required compared to traditional pipelines.

Model	Accuracy
BERT	0.9387
Voting Classifier	0.9007
Logistic Regression	0.8949
Linear SVC	0.8989
Multinomial NB	0.8771
Ridge Classifier	0.8990
Passive Aggresive Classifier	0.8931

Overall, the paper supports the idea that Transformer-based models such as BERT can serve as a strong default choice for text classification, while traditional models remain useful as lighter, less resource-intensive baselines.

9) Appendix: Referenced Research Paper

- [1] S. González-Carvajal and E. C. Garrido-Merchán, “Comparing BERT against traditional machine learning text classification,” *arXiv preprint arXiv:2005.13012*, 2021.
- [2] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
- [3] V. Sanh, L. Debut, J. Chaumond and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 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 optimized BERT pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.