Technical Report

Fine-Tuning DistilBERT for Airline Sentiment Analysis on Twitter

---

1. Introduction

This project explores fine-tuning a pretrained transformer model (DistilBERT) on the Twitter US Airline Sentiment dataset. The goal is to accurately classify tweets as positive, neutral, or negative, leveraging transfer learning to handle the nuances of short-form user-generated content. The model is trained to interpret tweets related to airline services, which often contain informal language, sarcasm, or emotionally charged expressions.

Using Hugging Face's `transformers`, `datasets`, and `Trainer` API, the pipeline includes data cleaning, tokenization, fine-tuning, hyperparameter tuning, and evaluation. The final deliverables include a reproducible notebook, performance visualizations, and a real-time prediction function.

---

2. Dataset Overview

- Source: Kaggle – Twitter US Airline Sentiment Dataset
- Labels: `negative`, `neutral`, `positive`
- Size: ~14,640 labeled tweets
- Preprocessing Steps:
    - Removed `@mentions`, `#hashtags`, and URLs
    - Added a `clean_text` field
    - Stratified splitting: 80% train, 10% validation, 10% test
    - Converted to Hugging Face `DatasetDict` format

---

3. Model and Methodology

- Base Model: `distilbert-base-uncased`

- Tokenizer: WordPiece tokenizer with truncation enabled
- Loss Function: Cross-Entropy
- Optimizer: AdamW
- Training Configuration:
    - Batch Size: 16
    - Epochs: 3
    - Learning Rates Tested: 1e-5, 2e-5, 5e-5
    - Evaluation: Per epoch using accuracy and F1 score

The model was fine-tuned using Hugging Face's `Trainer` class, which wrapped the training logic, evaluation steps, and logging.

---

4. Hyperparameter Experiments

Three training configurations were tested to identify optimal learning rates:

Experiment	Learning Rate	Training Size	Validation Accuracy	F1 Score
Baseline	2e-5	Full dataset	82.8%	77.2%
Exp 2	5e-5	1,000 samples	Approx. 80.1%	Approx. 74.6%
Exp 3	1e-5	1,000 samples	Approx. 81.0%	Approx. 76.0%

The best performance came from the full training run with a learning rate of 2e-5. Smaller and larger learning rates either underfit or converged prematurely.

---

5. Evaluation and Results

- Test Accuracy: Approximately 82.8%
- F1 Score: Approximately 77.2% (macro average)

- Confusion Matrix: Most errors involved confusion between neutral and positive tweets
- Error Trends:
    - Tweets with ambiguous tone were harder to classify
    - Sarcasm and indirect complaints were often misclassified as neutral

---

6. Error Analysis

Key areas where the model struggled:

1. Ambiguity in tone (e.g., "Great job, United...")
2. Tweets that mixed sentiment (e.g., complaint followed by a thank-you)
3. Very short tweets lacking context

The confusion matrix and `classification_report` provided insight into which classes were frequently confused.

---

7. Inference Pipeline

The `predict_sentiment(text)` function enables real-time sentiment predictions using the fine-tuned model. It:

- Tokenizes the input
- Passes it through the model
- Returns the predicted sentiment and confidence scores

This makes the model practical for interactive use cases such as dashboards or customer service filtering tools.

---

8. Limitations and Future Improvements

Limitations:

- Class imbalance (majority negative class)
- Domain bias (airline complaints dominate)
- DistilBERT has lower representational capacity than larger models

Future Improvements:

- Try models like `bert-base-uncased` or `roberta-base` for better accuracy
- Apply class weighting or oversampling techniques to handle imbalance
- Build a Streamlit or FastAPI UI for public deployment

---

9. Ethical Considerations

- Mentions, usernames, and URLs were removed to preserve user privacy
- The dataset inherently contains bias toward negative experiences; this may influence model behavior
- No demographic information was used or modeled
- The inference model is intended for internal analysis or tooling, not for public-facing sentiment labeling

---