Model Card

ATIS Intent Classification using BERT-Tiny

Kuntian Tang   Yurui Feng
University of Southern California

December 2025

Model Card: ATIS Intent Classification Models

Model Details

We compare three architectures for intent classification on the ATIS dataset:

- **TextCNN**: 3 convolutional branches (kernel sizes 3, 4, 5), each with 100 filters; max-pooling; dropout 0.5; FC classifier. Total parameters: ∼500k.
- **BiLSTM**: 1-layer bidirectional LSTM with hidden size 128 per direction; dropout 0.5; FC classifier. Total parameters: ∼700k.
- **BERT-tiny**: 2-layer transformer encoder (`prajjwal1/bert-tiny`); hidden size 128; fine-tuned with a linear classification head. Total parameters: 4.4M.

All models were implemented in PyTorch 2.1 and trained on CPU.

Training Data

Training, validation, and test splits follow standard ATIS:

- 4,481 training examples
- 497 validation examples
- 893 test examples (888 usable after removing unseen labels)

The dataset contains 22 intents with a highly imbalanced distribution. No data augmentation was applied.

Performance Metrics

The following table summarizes the final performance of the three models on the ATIS test set. Across all metrics, the fine-tuned `bert-tiny` model achieves the strongest overall performance.BERT-tiny achieves the highest overall performance among the three models, with the strongest test accuracy and weighted F1 score. While TextCNN and BiLSTM perform well on high-frequency classes, their macro F1 scores are lower due to difficulties in handling rare intents. BERT-tiny mitigates many of these issues through contextual representations, resulting in improved robustness across most categories.

BERT-tiny achieves the highest test accuracy and strongest macro-level performance, particularly on classes with semantic overlap. While TextCNN and BiLSTM remain competitive baselines—especially on high-frequency intents—BERT provides superior contextual understanding and overall robustness.

Table 1: Performance comparison of all models on the ATIS test set.

Model	**Validation Acc**	**Test Acc**	**Macro F1**	**Weighted F1**
TextCNN	0.984	0.9392	0.5186	0.9332
BiLSTM	0.982	0.9437	0.5093	0.9341
BERT-tiny	**0.986**	**0.9600**	**0.6200**	**0.9600**

Intended Use and Limitations

These models are designed for educational research in short-text classification. They perform well on ATIS-style utterances involving airline-related queries. While BERT-tiny provides the strongest performance overall, its computational cost is higher than that of the CNN and BiLSTM baselines, making it less suitable for extremely resource-constrained deployments.

Limitations include:

- Limited vocabulary outside airline/travel domain
- Difficulty handling rare composite intents
- Sensitivity to class imbalance

Failure Modes

All models struggle with:

- Rare intents (*city*, *meal*, *flight+airfare*)
- Composite labels that mix multiple semantic frames
- Extremely short or ambiguous queries

All models show reduced performance on extremely low-frequency intents, which have very limited or zero representation in the training set. TextCNN often confuses semantically similar labels due to its reliance on local patterns, and BiLSTM struggles with long-tail classes for similar reasons. BERT-tiny substantially reduces several of these confusion patterns thanks to contextualized embeddings, although it still exhibits occasional errors on the rarest categories.

Fairness and Bias

Because ATIS concerns airline booking—not social topics—traditional fairness issues (gender, race, etc.) do not apply. However, dataset bias exists in the form of:

- Strong skew toward the *flight* intent (70% of training data)
- Underrepresented classes with ¡10 examples

These biases affect model calibration and macro F1.

Ethical Considerations

This model is not intended for deployment. It is a course project for understanding model architectures, training dynamics, and evaluation practices. No sensitive or personal data is used.