## Natural Language Processing: Challenge

Handout: 07.04.2025 08:00
Due: 15.04.2025 23:59

Spring 2025
**Hands-On Deep Learning**

**</> Challenge** Open in IDE

# Sentiment Analysis on IMDB

**Prototype Notebook**: Open in Google Colab.

**Important**: you ahve to submit your solution on CodeExpert.

In this week's notebook, you have explored the entire NLP pipeline, from processing textual data to evaluating models for sentiment analysis.

Your final task is to apply your new knowledge by building a working model to classify movie reviews from the IMDB dataset as positive or negative.

Once you have a working model, experiment with the learning rate, batch size, sequence length, and different model architectures (e.g., varying hidden sizes, layers, or using pretrained embeddings) to achieve the highest possible classification accuracy.

See your position on the leaderboard.

## Scoring System

Points are awarded based on the model's final score, according to the thresholds below:

| Model accuracy (running) | CodeExpert score (submitting) | Points Earned |
|---|---|---|
| < 0.70 | < 8.3% | 0 |
| ≥ 0.70 | ≥ 8.3% | 1 |
| ≥ 0.85 | ≥ 25.0% | 2 |
| ≥ 0.89 | ≥ 41.7% | 3 |
| ≥ 0.90 | ≥ 58.3% | 4 |
| ≥ 0.91 | ≥ 75.0% | 5 |
| ≥ 0.917 | ≥ 91.7% | 6 |

The model accuracy is your models performance on the test set, see logs after running a job. The CodeExpert score is the percentage displayed in CodeExpert when submitting the job.

## Task

Your goal is to complete the `preprocess_function`, `init_model` and `train_model` functions. Your functions are imported and called in for evaluation. **Function signature must remain unchanged.** Your code will be called by the following function:

```
def run():
  set_seed(42)

  # Get datasets for training and testing
  train_dataset, test_dataset = get_data()

  # Initialize the model using student's init_model function
  model = init_model()

  # Train the model using student's train_model function
  model = train_model(model, train_dataset)

  # Evaluate the model on the test set
  score = evaluate_model(model, test_dataset)

  return score
```

The `preprocess_function` is applied to both the train and test datasets:

```
dataset = dataset.map(preprocess_function, batched=True)
```

The training dataset is provided to you. Please use the `Trainer` method from the `transformers` library to train and evaluate your model (as it will automatically select the device).

## Rules

You are not allowed to use samples in the test dataset for training, or to use a model pre-trained on the test dataset.

You cannot set a new seed.

Your code must run in 15 minutes.

## Tutorial

Tutorial on how to use CodeExpert.

## Toblerone

The top three solutions will receive a Toblerone. Good luck! 🚀