NAME: Antonio Cordero

INTRODUCTION

In this project, I explored the task of automated topic classification of news headlines using simple neural networks. The goal was to assign a news title to a relevant thematic category (e.g., sports, finance, health) based solely on its wording. This task is interesting because it allows for rapid content filtering and organization, particularly in large-scale media environments.

I tried to answer if a simple feedforward neural network trained on TF-IDF representations of news titles achieve meaningful topic classification despite class imbalance. I expected good performance for dominant categories (like sports) and poor generalization for underrepresented ones.

MATERIAL AND METHODS

I used the MIND (Microsoft News Dataset), a large-scale English-language dataset for news recommendation. From it, I extracted only the Title and Category columns from the news.tsv file. After removing null entries, I encoded categories using LabelEncoder. Text was transformed into numerical features using TF-IDF vectorization (limited to 5000 terms), and data was split into 80% training and 20% testing sets. Features and labels were converted into PyTorch tensors.

The model architecture was a two-layer feedforward neural network:

- Input: 5000-dimensional TF-IDF vectors.
- Hidden layer: 100 ReLU-activated units.
- Output: 18-category softmax layer (via CrossEntropyLoss).

The model was trained for 10 epochs using Adam optimizer (learning rate = 0.001) and mini batches of size 32.

RESULTS

The model achieved an overall accuracy of 67.16% on the test set. As expected, performance varied significantly by class:

- High performance in dominant categories:

    - Sports: F1-score = 0.89
    - News: F1-score = 0.68
- Low or no performance in underrepresented classes (e.g., kids, games, northamerica: F1 = 0.00)

These results align with expectations. The model generalizes reasonably well for high-frequency categories but fails for underrepresented ones due to class imbalance and the simplicity of the architecture. An interesting result is the example “Doctors go strike”, a text that could be News, but the model classifies it as Health, probably because of the word *Doctor*.

CODE

https://github.com/AntonioCorderoBrummer/Exercise-3.git