AG News Headline Classification

Yunmin CHA (2022123028)

1 Introduction

This report summarizes a lightweight AI pipeline for classifying AG News headlines into four topics (World, Sports, Business, Sci/Tech). The goal is to compare a naive keyword baseline against a small, pre-trained text encoder paired with a simple classifier. All experiments were run locally on CPU with a small subset of the public dataset to keep runtime short while still revealing clear performance gaps between the approaches.

2 Task Definition

- **Task description:** Assign one of four AG News topics to a headline.
- **Motivation:** News topic tagging is a common routing/filtering step for downstream products such as personalized feeds.
- **Input / Output:** Input is a single news headline string; output is one of {World, Sports, Business, Sci/Tech}.
- **Success criteria:** High accuracy and balanced macro-F1 on a held-out test split; qualitative improvements over a keyword rule baseline.

3 Methods

3.1 Naïve Baseline

A hand-written keyword matcher scores each class by counting keyword hits in the headline (e.g., *war, minister* for World or *coach, season* for Sports). Ties and zero-hit cases fall back to the majority class in the training subset. This baseline ignores context, negation, and ambiguous terms, so it often fails when headlines lack obvious cue words.

3.2 AI Pipeline

Headlines are lowercased, encoded with the SentenceTransformer model `all-MiniLM-L6-v2`, and fed into a multinomial logistic regression classifier (`C=4.0`, `max_iter=1000`, `n_jobs=-1`). The encoder runs on GPU if available, otherwise CPU. This keeps the pipeline fast and compact while leveraging a modern embedding model for semantic signals that keyword rules miss.

4 Experiments

4.1 Datasets

The Hugging Face `ag_news` dataset was used. A reproducible seed (42) sampled 2,000 training examples and 500 test examples. Headlines were lowercased and stripped of surrounding whitespace;

no additional preprocessing or data augmentation was applied.

4.2 Metrics

Accuracy and macro-F1 were computed on the held-out test set to balance class-specific performance.

4.3 Results

Method	Accuracy	Macro-F1
Naive keyword baseline	0.460	0.447
MiniLM + logistic regression	0.850	0.852

The embedding-based pipeline improves accuracy by +0.39 and macro-F1 by +0.41 over the keyword rules, showing the value of contextual representations even with a small linear head.

Qualitative differences. Examples where the pipeline corrected the baseline:

- *"paris tourists search for key to 'da vinci code' mystery"* (true: World; baseline: Sports; pipeline: World)
- *"profit plunges at international game tech"* (true: Business; baseline: Sports; pipeline: Business)
- *"general mills goes whole grains"* (true: Business; baseline: Sports; pipeline: Business)

5 Reflection and Limitations

The keyword baseline was fast but brittle, collapsing many unrelated headlines into the majority class. MiniLM embeddings paired with logistic regression delivered a large jump in both accuracy and macro-F1, with notably better handling of headlines lacking explicit cue words. Remaining errors often involved Business vs. World ambiguity and Sci/Tech items that looked like market news. The small 2k/500 split kept runtime under a minute but limits robustness; more data or light class rebalancing could help. Hyperparameters for the classifier were only lightly tuned, so a quick sweep or calibration step might yield incremental gains. Future work could try prompt-based zero-shot classifiers or fine-tuned small transformers to see if marginal improvements justify extra compute.