CAS2105 Homework 6: Mini AI Pipeline Project 🤗

AG News Topic Classification with a Keyword Baseline and DistilBERT Pipeline

Chaewon Kim (2024149035)

1 Introduction

In this project, I designed and evaluated a small AI system using the principles of a mini AI pipeline. Rather than training a large model, I selected a lightweight text classification problem and compared a naive baseline with a pretrained transformer model used purely for inference.

I chose the **AG News topic classification task**, where the system predicts one of four categories: World, Sports, Business, or Sci/Tech. This task is simple and well-studied, making it ideal for demonstrating the performance gap between heuristic approaches and modern pretrained models. The project highlights core AI development steps such as baseline construction, pipeline design, evaluation, and reflection.

2 Task Definition

- **Task description:** The task is to perform news headline classification using the AG News dataset, where each sample consists of a headline and a short accompanying description. The goal is to assign each news text to one of four categories: {World, Sports, Business, Sci/Tech}.
- **Motivation:** Topic classification is widely used in search, recommendation, and information filtering. It provides a clear comparison point between simple heuristics and transformer-based pipelines.
- **Input / Output:**
  - **Input:** A news headline together with its short accompanying description.
  - **Output:** One of the four topic labels: {World, Sports, Business, Sci/Tech}.
- **Success criteria:** We judge how good the system is by its classification accuracy on a held-out test set; higher accuracy means better performance. Specifically, the AI pipeline should outperform the naive keyword baseline by a significant margin.

3 Methods

3.1 Naïve Baseline

Your Baseline

- **Method description:** The baseline classifier is a keyword-matching rule system. For each category, I manually defined several indicative keywords (e.g., “war”, “minister” for World; “team”, “league” for Sports). The classifier counts the number of keywords appearing in the

text for each class and selects the class with the highest count. If no keywords match, it defaults to the World category.

- **Why naïve:**
  - It relies on literal substring matching only.
  - It cannot understand synonyms, paraphrases, grammar, or context.
  - It fails when keywords are missing, ambiguous, or used metaphorically.
- **Likely failure modes:**
  - Headlines with subtle topical cues (e.g., sports stories without explicit sports terms)
  - Overlapping words across categories.
  - Headlines requiring contextual or semantic reasoning.

3.2 AI Pipeline

Your Pipeline

- **Models used:** The improved system uses the pretrained transformer model `textattack/distilbert-base-uncased-ag-news`, accessed through the HuggingFace Transformers library. No fine-tuning was performed; the model is used strictly in inference mode.
- **Pipeline stages:**
  1. **Preprocessing:** Convert each news text (headline + short description) into token IDs using the DistilBERT tokenizer with padding and truncation.
  2. **Embedding:** Feed the tokenized batch through the pretrained DistilBERT model to obtain contextual representations and class logits.
  3. **Decision Module:** Select the predicted label by taking the `argmax` over the logits for each example.
  4. **Post-processing:** Compare predictions with ground-truth labels and compute accuracy.
- **Design choices and justification:** DistilBERT provides a good balance between computational efficiency and performance. Since it is pretrained on a text corpus and specifically adapted to the AG News domain, it can perform well even without fine-tuning.

4 Experiments

4.1 Datasets

Your Dataset Description

- **Source:** AG News dataset from the HuggingFace `datasets` library.
- **Total examples:** 2,500 examples in total
- **Train/Test split:** I randomly sampled 2,000 training examples and 500 test examples from the AG News dataset.
- **Preprocessing steps:** I used short news text (headline + short description) as input, and rely on the DistilBERT tokenizer to handle tokenization, padding, and truncation to a maximum length of 128 tokens.

4.2 Metrics

Use at least one quantitative metric appropriate for your task:

- **Classification:** accuracy, precision, recall, F1,
- **Retrieval:** precision@k, recall@k,
- **Simple generation:** exact match, ROUGE-1.

It's worth considering how the metrics you select align with your tasks.

I used classification accuracy as the primary evaluation metric.

4.3 Results

Results table(metric values for baseline vs. pipeline

Method	Accuracy
Baseline	0.530
AI Pipeline	0.928

Qualitative Examples

- **Example 1**
  - **Text:** Indian board plans own telecast of Australia series. The Indian cricket board said on Wednesday it was making arrangements on its own to broadcast next month's test series against Australia, which is under threat because of a raging TV rights dispute.
  - **True label:** Sports
  - **Baseline pred:** Sci/Tech
  - **Pipeline pred:** Sports
- **Example 2**
  - **Text:** REVIEW: 'Half-Life 2' a Tech Masterpiece (AP). It's been six years since Valve Corp. perfected the first-person shooter with "Half-Life." Video games have come a long way since, with better graphics and more options than ever. Still, relatively few games have mustered this one's memorable characters and original science fiction story.
  - **True label:** Sci/Tech
  - **Baseline pred:** Sports
  - **Pipeline pred:** Sci/Tech
- **Example 3**
  - **Text:** ADV: Try Currency Trading Risk-Free 30 Days. 24-hour commission-free trading, 100-to-1 leverage of your capital, and Dealbook Fx 2—our free advanced trading software. Sign up for our free 30-day trial and receive one-on-one training.
  - **True label:** Business
  - **Baseline pred:** World
  - **Pipeline pred:** Business

5 Reflection and Limitations

Your Reflection

Overall, the project went more smoothly than I expected once I switched from a naive rule-based approach to a pretrained model. The keyword baseline was easy to code, but it was surprisingly difficult to choose keywords that covered many cases without introducing obvious failure modes, and its final accuracy was lower than I had hoped. In contrast, the DistilBERT pipeline required more setup with libraries and the model, but once everything was installed correctly it produced strong results with relatively little custom code. One practical difficulty was dealing with library and version issues, such as getting PyTorch and the Transformers library to work together without errors.

Using accuracy as the main metric worked reasonably well because the dataset is balanced and the task is straightforward multi-class classification. However, accuracy alone does not explain why certain examples are misclassified, so the qualitative examples were important to understand the model's behavior. If I had more time or compute, I would like to explore additional metrics such as per-class accuracy or confusion matrices, and possibly fine-tune DistilBERT on the selected subset of AG News.

References

[1] AG News Dataset. Hugging Face Datasets. Available at https://huggingface.co/datasets/ag_news. Accessed 2025.

[2] `textattack/distilbert-base-uncased-ag-news`. Pretrained DistilBERT model for AG News classification. Available at https://huggingface.co/textattack/distilbert-base-uncased-ag-news. Accessed 2025.

[3] Thomas Wolf, Lysandre Debut, Victor Sanh *et al.*
HuggingFace's Transformers: State-of-the-art Natural Language Processing.
In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, 2020.