AG News Headline Classification: From Keywords to DistilBERT

ZHANG ZEHAO (2021147540)

1 Introduction

In this project, we build a small AI system that can read a news headline and guess which topic it belongs to. We use the AG News[1] dataset, where each English headline is labeled as one of four categories: World, Sports, Business, or Sci/Tech. For example, a headline about a soccer match should be Sports, a story about a stock market crash should be Business, and an article about a new programming language release is usually Sci/Tech. News classification of this kind is common in real applications (news apps, portals, and recommendation systems all need it), but the problem is still small enough that we can understand and debug the whole pipeline step by step.

The interesting part of this project is not just getting a high accuracy, but seeing how different design choices change the behavior of the system. We first build a very simple keyword-based baseline: a set of hand-written rules that look for words like "goal", "shares", or "software" and map them to one of the four classes. Then we build a stronger AI pipeline using a pretrained Transformer model (DistilBERT[2]) and fine-tune it on the same dataset. This lets us directly compare "old school" rules with a modern language model under the same conditions. Along the way, we have to decide how to split the data, how to tokenize the text, which metrics to use, and how to interpret the errors the models make. Our goal is to end up with a pipeline that is easy for a classmate to understand, clearly improves over the naïve baseline, and gives us some intuition about what pretrained models actually buy us on a simple text classification task.

2 Task Definition

- **Task description:** We want to classify each AG News headline into one of four topic categories: World, Sports, Business, or Sci/Tech. Given only the short headline text, the system should decide which label best describes the main topic of the news.
- **Motivation:** Topic classification for news is a very common real-world use case (e.g., organizing articles on news websites or building simple recommendation systems). At the same time, the AG News dataset is small and clean enough that we can run experiments quickly and clearly see the difference between a hand-crafted baseline and a pretrained language model, without needing huge compute or very complex code.
- **Input / Output:** The input to our system is a single English news headline (a short piece of text). The output is one discrete label from the set {World, Sports, Business, Sci/Tech}. Internally, the baseline maps the raw string to a label using keyword rules, while the DistilBERT pipeline first tokenizes the string into subword IDs and then predicts a label with a fine-tuned classifier.
- **Success criteria:** We consider the system "good" if the DistilBERT pipeline clearly outperforms the naïve keyword baseline on the held-out AG News test set in terms of accuracy and macro-F1. In particular, we look for a large absolute improvement (around +0.4 in both

metrics in our experiments) and for balanced performance across all four classes, rather than a model that only does well on the majority class.

3 Methods

3.1 Naïve Baseline

- **Method description:** We implement a simple keyword-based text classifier for the AG News headlines. For each of the four classes (World, Sports, Business, Sci/Tech), we manually define a small list of domain-specific keywords that are likely to appear in articles of that category. Given a headline, we lowercase the text, remove punctuation, split it into tokens, and count how many keywords from each class appear in the headline. The class with the highest keyword count is returned as the prediction; if no keyword is matched at all, we fall back to the majority class in the training data (World).
- **Why naïve:** This baseline does not learn any parameters from data and does not use any pretrained model; all "knowledge" comes from our hand-written keyword lists. It treats each headline as an unordered bag of words and completely ignores word order, syntax, context, and interactions between words. The method cannot recognize synonyms, paraphrases, or more subtle semantic cues that are not explicitly included in the keyword lists. As a result, it is very easy to implement and serves as a sanity-check baseline, but it clearly underutilizes the information contained in the dataset.
- **Likely failure modes:** The keyword rules are likely to fail on ambiguous or mixed-topic headlines. For example, a story about business or technology that also mentions political events or institutions may be pushed into the World class because the World keyword list focuses on generic politics and international news, while the other lists cover a relatively narrow set of cues for sports, business, and technology. Very short or creative headlines that do not contain any predefined keywords will also be misclassified by default as World. In addition, the baseline struggles when the main topic is expressed through named entities or phrases that are not included in the lists (such as company names, product names, or sports teams), leading to systematic confusion between categories like Business and Sci/Tech or between Sports and World.

3.2 AI Pipeline

- **Models used:** We build our AI pipeline on top of the pretrained `distilbert-base-uncased` model from Hugging Face. We use the corresponding `AutoTokenizer` to convert raw headlines into token IDs, and the `AutoModelForSequenceClassification` wrapper, which attaches a randomly initialized linear classification layer on top of DistilBERT for the four AG News categories (World, Sports, Business, Sci/Tech).
- **Pipeline stages:** Conceptually, our DistilBERT-based pipeline consists of four main stages. (1) *Input preprocessing*: we take the raw headline strings from the dataset and feed them directly to the DistilBERT tokenizer, enabling truncation and padding to a fixed maximum length of 64 tokens. (2) *Text encoding*: the tokenized inputs (input IDs and attention masks) are passed through DistilBERT, which produces contextualized representations for each token and a pooled representation for the entire headline. (3) *Classification*: the pooled representation is fed into a linear classification head that outputs logits over the four classes, and we apply a softmax followed by an argmax to obtain the predicted label. (4) *Training and evaluation*: we fine-tune all model parameters using cross-entropy loss on the training split, then evaluate the trained model on the validation and test splits using accuracy and macro-F1.
- **Design choices and justification:** DistilBERT is a lightweight Transformer model that is substantially smaller and faster than BERT, while still providing strong performance on sentence-level classification tasks. This makes it a good fit for the scale of AG News and the

computational budget of this homework. We limit the maximum sequence length to 64 tokens because the dataset consists of short headlines, so longer inputs would mainly add padding without improving accuracy. We fine-tune the model for 5 epochs with a batch size of 32, a learning rate of 5×10^{-5}, weight decay of 0.01, and a warmup ratio of 0.05 to obtain stable optimization. The pipeline is implemented using the Hugging Face `Trainer` API, which keeps the training loop concise and reproducible and allows us to focus on the comparison between the naïve baseline and the pretrained Transformer model.

4 Experiments

4.1 Datasets

- **Source:** We use the AG News corpus provided through the Hugging Face `datasets` library (`"ag_news"`). Each example is an English news headline paired with one of four topic labels: World, Sports, Business, or Sci/Tech.
- **Total examples:** The original dataset contains 120,000 labeled examples in the training split and 7,600 labeled examples in the test split, for a total of 127,600 headlines.
- **Train/Test split:** We follow the official split from the `ag_news` dataset, using the provided training and test splits. From the 120,000 training examples, we further hold out 10% as a validation set using a random split with a fixed seed. This results in 108,000 examples for training, 12,000 examples for validation, and 7,600 examples for testing.
- **Preprocessing steps:** For the keyword-based baseline, we lowercase each headline, remove punctuation, and tokenize on whitespace before counting keyword matches. For the DistilBERT pipeline, we feed the raw headline strings into the pretrained tokenizer, which lowercases the text (uncased model), tokenizes it into subword units, and applies truncation and padding to a fixed maximum length of 64 tokens. The tokenizer also produces attention masks, and the original integer labels in the dataset are already encoded as {0,1,2,3}, which match the four output classes of the classifier.

4.2 Metrics

We evaluate both the keyword baseline and the DistilBERT pipeline using two standard metrics for multi-class text classification: accuracy and macro-F1. These metrics are easy to interpret and match our goal of building a system that performs well on all four AG News categories, not just on the majority class.

- **Accuracy:** The fraction of headlines in the test set for which the predicted label exactly matches the gold label. This gives a simple overall sense of how often the system is correct, and it is the metric most people think of first when they hear "classification performance."
- **Macro-F1:** The average of the per-class F1 scores, where each class (World, Sports, Business, Sci/Tech) is weighted equally. This makes it easier to see whether the model is ignoring any minority class: a system that only does well on World but often fails on Business or Sci/Tech would have a much lower macro-F1 than accuracy.

Using both accuracy and macro-F1 together lets us check that improvements are not coming only from the dominant class, but reflect better performance across all four news topics.

4.3 Results

We compare the keyword baseline and our DistilBERT-based pipeline on the AG News test set using accuracy and macro-F1 as our main metrics. The keyword baseline reaches roughly 53% accuracy and macro-F1, while the DistilBERT pipeline achieves about 94% on both metrics. This

corresponds to an absolute gain of around +0.41 in accuracy and +0.42 in macro-F1, showing that the full pipeline makes much better use of the information in the headlines.

Method	Accuracy	Macro-F1
Keyword baseline	0.5332	0.5263
DistilBERT pipeline	0.9416	0.9416

Figure 1 shows the classification report from one training run on the test set. The per-class precision/recall/F1 values there are consistent with the summary in the table above.

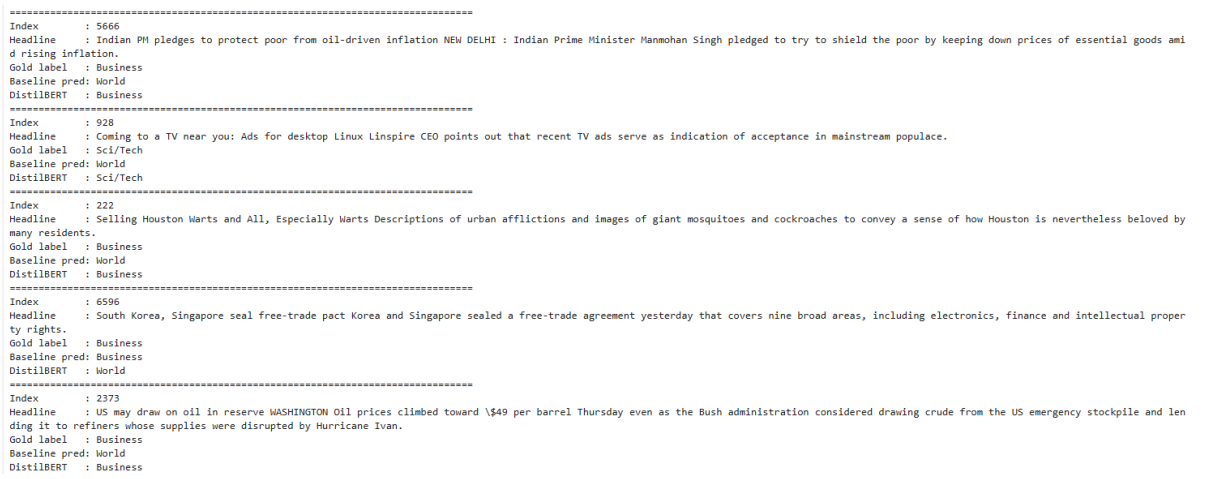

Figure 1: Classification report on the AG News test set for the DistilBERT pipeline.

Qualitative examples. We also inspected individual headlines where the two methods behave differently. Below are five short examples from our random sample:

- **Index 5666.** Headline about the Indian prime minister pledging to protect the poor from *oil-driven inflation*. Gold label: `Business`. The baseline predicts `World`, probably focusing on the country and politician, while DistilBERT correctly predicts `Business` based on the economic context (oil prices, inflation).
- **Index 928.** Headline about TV ads for a desktop Linux distribution. Gold label: `Sci/Tech`. The baseline outputs `World`, since our keyword lists miss this particular product name, but DistilBERT correctly predicts `Sci/Tech` by understanding the technology-related context.
- **Index 222.** Headline about selling Houston "warts and all" and descriptions of urban afflictions. Gold label: `Business`. The baseline again predicts `World`, whereas DistilBERT assigns `Business`, capturing that the story is about the city's image and marketing rather than general world news.
- **Index 6596.** Headline about South Korea and Singapore sealing a free-trade pact. Gold label: `Business`. Here the baseline prediction `Business` is correct, but DistilBERT predicts `World`, seemingly over-emphasizing the country names and underweighting the economic phrase *"free-trade pact"*. This illustrates that the pipeline can still be biased by prominent location words.
- **Index 2373.** Headline about using U.S. oil reserves and rising oil prices. Gold label: `Business`. The baseline predicts `World`, while DistilBERT correctly predicts `Business`, making use of financial cues such as *"oil prices"*, *"barrel"*, and supply disruptions.

Overall, the quantitative metrics and these qualitative examples tell a consistent story: the naïve keyword baseline captures some obvious patterns, but our DistilBERT pipeline is much more reliable and better aligned with the true news categories, even though it can occasionally be distracted by strong geopolitical signals.

5 Reflection and Limitations

We did not expect the DistilBERT pipeline to work as well as it did: with five training epochs and fairly standard hyperparameters, it already reached about 0.94 accuracy and macro-F1 on the AG News test set. This was a large jump over the keyword baseline, which stayed around 0.53 on both metrics but still captured some obvious patterns in sports and very typical world-news headlines. Designing that baseline turned out to be more time-consuming than we first thought, because we had to decide which keywords to include, how to handle overlaps between classes, and what to do when no rule fired. On the pipeline side, the trickiest pieces were getting the dataset splits, tokenization, and `Trainer` configuration aligned so that training and evaluation ran stably on the GPU. While we saw a few headlines that were genuinely ambiguous between Business and World or between Sci/Tech and Business, in most cases the model's mistakes were understandable once we looked closely at the wording. Accuracy and macro-F1 together gave a reasonable summary of "quality": when they improved, the qualitative examples almost always looked better as well, and macro-F1 helped ensure that no single class was being ignored. At the same time, these metrics do not differentiate between harmless confusions (e.g., Sports vs. World) and more serious ones (e.g., Sci/Tech vs. Business), so they miss some nuance that we saw in the examples. With more time or compute, we would try a slightly larger model such as BERT-base, experiment with light domain adaptation for business and technology news, and add a stronger classical baseline (e.g., TF–IDF + logistic regression) to position the gains from the transformer more clearly.

References

[1] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. URL https://arxiv.org/abs/1509.01626.

[2] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC^2) at NeurIPS*, 2019. URL https://arxiv.org/abs/1910.01108.