Mini AI Pipeline Project: From Keywords to BERT hugging-face

Seonjun Kim (2023149026)

1 Introduction

The goal of this project is to classify news headlines into four distinct categories: World, Sports, Business, and Sci/Tech. Rather than training large models from scratch, this project focuses on the process of constructing an AI workflow. I implemented two approaches: a Naïve Baseline using keyword matching rules, and an advanced AI Pipeline leveraging a pre-trained BERT model. By comparing these two methods on the AG News dataset, I aim to demonstrate the significant performance gap between heuristic approaches and modern deep learning techniques, while also reflecting on the trade-offs involving complexity and interpretability.

2 Task Definition

- **Task description:** The objective is to classify short news texts (titles and descriptions) into one of four categories: `World`, `Sports`, `Business`, or `Sci/Tech`.
- **Motivation:** Automated text classification is a fundamental task in Natural Language Processing (NLP). It is essential for organizing large volumes of information, such as news aggregation services, spam filtering, and sentiment analysis.
- **Input / Output:**
  - Input: A text string containing a news headline and a brief snippet.
  - Output: A single class label (Integer 0-3 mapped to category names).
- **Success criteria:** The system's quality is evaluated using **Accuracy** and **F1-score**. A "good" system should significantly outperform random guessing (25% accuracy).

3 Methods

This section describes the implementation of both the rule-based baseline and the deep learning-based AI pipeline.

3.1 Naïve Baseline

I implemented a simple keyword-based classifier that does not rely on machine learning parameters.

- **Method description:** I defined a dictionary of representative keywords for each class (e.g., "olympic", "score" for Sports; "stock", "market" for Business). The method tokenizes the input text and counts the occurrence of these keywords. The class with the highest count is selected. If there is a tie, a random choice is made among the top candidates.
- **Why naïve:** It ignores sentence structure, context, and semantics. For example, the word "Apple" is treated the same whether it refers to the fruit or the tech company.

- **Likely failure modes:** It fails when the text uses synonyms not present in the keyword list or relies on metaphorical language (e.g., "Bulls run on Wall Street").

3.2 AI Pipeline

I designed an improved pipeline using the **BERT** (Bidirectional Encoder Representations from Transformers) architecture.

- **Models used:** I utilized the `fabriceyhc/bert-base-uncased-ag_news` model from the Hugging Face Hub. This model is a fine-tuned version of `bert-base-uncased` specifically trained on the AG News dataset.
- **Pipeline stages:**
  1. **Preprocessing:** Tokenization (WordPiece) and truncation to fit the model's max sequence length.
  2. **Inference:** The pre-trained BERT model encodes the text and outputs logits for the 4 classes.
  3. **Post-processing:** Mapping the model's output labels (e.g., "LABEL_0") to human-readable categories.
- **Design choices:** Using a fine-tuned model allows for high accuracy without the computational cost of training from scratch. The pipeline abstraction simplifies the complexity of tensor management.

4 Experiments

4.1 Datasets

I used the **AG News** dataset, a popular benchmark for text classification.

- **Dataset source**: Hugging Face Datasets library (`ag_news`).
- **Size**: A subset of **1,000 examples** from the test split was selected for evaluation to ensure quick iteration.
- **Splits**: The experiment focuses on evaluation; thus, only the test split was used.
- **Preprocessing steps:**
  - Converted all text to lowercase.
  - Removed special characters and punctuation using regular expressions (`r'[^a-zA-Z0-9\s]'`). This retains only alphanumeric characters and whitespace.
  - Collapsed multiple consecutive spaces into a single space and stripped leading/trailing whitespace.

4.2 Metrics

I selected **Accuracy** as the primary metric because the class distribution in AG News is perfectly balanced. I also reported **Precision, Recall, and F1-score** (weighted average) to observe class-specific performance.

4.3 Results

The experimental results demonstrate a substantial performance gap between the simple keyword-based approach and the deep learning-based pipeline.

Metric	Naïve Baseline	AI Pipeline (BERT)
Accuracy	0.531	**0.897**

Table 1: Overall Performance Comparison (Accuracy)

2. Class-wise Performance (F1-Score)

Table 2 breaks down the performance by category. The difference is most prominent in the `Sci/Tech` category, where the baseline struggled significantly.

Category	Naïve Baseline			AI Pipeline (BERT)		
	Prec.	Rec.	F1	Prec.	Rec.	F1
World	0.527	0.523	0.525	0.938	0.857	0.896
Sports	0.607	0.703	0.652	0.932	**0.996**	**0.963**
Business	0.496	0.537	0.516	0.842	0.890	0.866
Sci/Tech	0.470	0.360	0.407	0.876	0.847	0.861

Table 2: Detailed Performance by Category

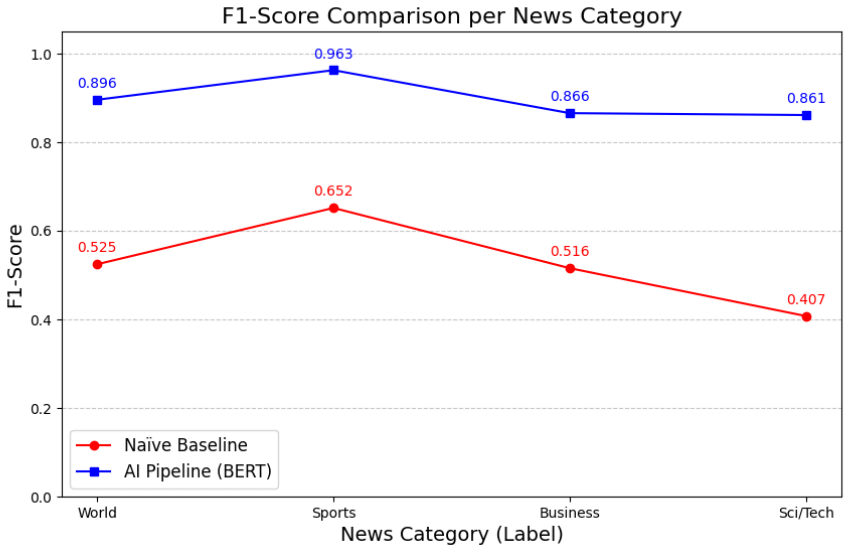

Figure 1: Comparison of F1-Scores across different news categories. The AI Pipeline shows consistent high performance, whereas the Baseline fluctuates significantly depending on the topic.

Qualitative Analysis

The quantitative results reveal several key insights:

1. **Baseline's Limitations in Sci/Tech:** The Naïve Baseline performed worst in the `Sci/Tech` category (F1: 0.407). This is likely due to the high vocabulary overlap between Tech and Business (e.g., "company", "market", "service"), which confuses a simple keyword counter. Precision (0.470) and Recall (0.360) were both the lowest among all classes.

2. **Baseline's Strength in Sports:** Interestingly, the Baseline achieved its highest performance in `Sports` (F1: 0.652). Sports vocabulary (e.g., "game", "team", "win", "cup") is distinct and less ambiguous compared to other topics, making it easier for rule-based systems to classify correctly.

3. **Robustness of AI Pipeline:** The BERT-based model maintained high F1-scores across all categories (> 0.86). It effectively resolved the ambiguity in `Sci/Tech`, improving the score from 0.407 to 0.861. This confirms that the model leverages contextual cues rather than relying solely on specific keywords.

Example Comparison Cases:

1. **Contextual Ambiguity:**
   - **Text:** "intel dauphin otellini becomes king intel 39s board has given the go ahead for the long anticipated shift in power from current ceo craig barrett to current president paul otellini come may 18 otellini will take over the chipmaker and become its fifth ever ceo"
   - **Baseline:** Predicted `World` (Failed). The model triggered on the keyword "president," associating it with political leaders/government rather than a corporate job title.
   - **AI Pipeline:** Predicted `Business` (Correct). BERT recognized entities like "Intel" and "CEO" to correctly interpret "president" in a corporate context.
2. **Keyword Overlap:**
   - **Text:** "klitschko too good for williams vitali klitschko proved too strong for danny williams as he retained his world championship crown in las vegas last night williams vowed to continue boxing despite being outclassed by klitschko"
   - **Baseline:** Predicted `World` (Failed). The explicit presence of the keyword "world" (defined as a keyword for the World category) misled the baseline.
   - **AI Pipeline:** Predicted `Sports` (Correct). The AI understood "world championship" as a modifier for a sporting event, not international news.
3. **Complex Phrasing:**
   - **Text:** "toys will be toys trust me you don 39t want to see desperate parents shopping over the holidays i was at circuit city nyse cc last week and saw a mother pleading with a sales clerk for a nintendo ds portable video game system."
   - **Baseline:** Predicted `Sports` (Failed). The baseline strictly mapped the keyword "game" to the Sports category, failing to distinguish physical sports from digital/video games.
   - **AI Pipeline:** Predicted `Sci/Tech` (Correct). The model captured the semantic meaning of "Nintendo DS" and "video game system" as technology products.

5 Reflection and Limitations

Reflection

This project highlighted the fundamental trade-off between interpretability and performance in NLP tasks. The Naïve Baseline, while transparent and easy to debug, demonstrated clear limitations in handling polysemy (e.g., "game" in tech vs. sports) and lacked robustness against distinct vocabularies. Its 53% accuracy, although better than random guessing, reveals the brittleness of heuristic rule-based systems.

In contrast, the AI Pipeline leveraging a pre-fine-tuned BERT model achieved nearly 90% accuracy. This significant leap confirms the efficacy of **Transfer Learning**, where a model pre-trained on a vast corpus effectively captures semantic nuances and contextual dependencies that simple keyword counting cannot.

Limitations

Despite the superior performance of the AI pipeline, there are inherent limitations to consider:

- **Computational Complexity:** The BERT model requires significantly more computational resources for inference compared to the lightweight baseline. For latency-sensitive applications, this heavy architecture could be a bottleneck without further optimization.
- **Interpretability Trade-off:** Unlike the baseline where errors can be directly traced to specific keywords, the BERT model operates as a "black box." Understanding exactly *why* a specific misclassification occurred is challenging without advanced interpretability tools.
- **Domain Specificity:** While the model performs well on general news, the current pipeline might struggle with highly specialized jargon or texts containing rare proper nouns that were not well-represented in the pre-training corpus.

References

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). **BERT: Pre-training of deep bidirectional transformers for language understanding**. *arXiv preprint arXiv:1810.04805*.

[2] Zhang, X., Zhao, J., & LeCun, Y. (2015). **Character-level convolutional networks for text classification**. *Advances in neural information processing systems*, 28.

[3] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). **HuggingFace's Transformers: State-of-the-art natural language processing**. *arXiv preprint arXiv:1910.03771*.