CAS2105 Homework 6: Mini AI Pipeline Project 🤗

Aliyah Tiffanie Manalo (2024190222)

1 Task Definition

- **Task description:** The task is to classify news headlines into four categories: world news, sports news, business news, and science-technology news.
- **Motivation:** This task provides a simple AI pipeline that is suitable for a beginner. News headline classification is also interesting, since classification is a core topic that is covered in another course. Choosing this task allows me to apply the knowledge that I have about AI.
- **Input / Output:** A news headline is fed into the model as an input. The output is a predicted category that belongs to one of the four categories that were mentioned above.
- **Success criteria:** The system is evaluated using accuracy, recall, precision, and a confusion matrix. A confusion matrix is plotted to compare the true and predicted labels. It also displays the model's errors, which is useful for evaluating its accuracy. The diagonal values show the model's correct predictions, while the values outside the diagonal show its wrong predictions.

2 Methods

This section includes both the naïve baseline and the improved AI pipeline.

2.1 Naïve Baseline

Implement a simple method that does not rely on heavy models. Examples include:

- Keyword-based text classification,
- Simple color/shape heuristics for image tasks,
- String-overlap–based retrieval.

In your report, explain:

- How the baseline works,
- Why it is considered naïve,
- Expected failure cases.

Baseline

- **Method description:** A keyword-based text classification is applied into the model. Each news headline is scanned to search for the predefined keywords in each category.
- **Why naïve:** It is naïve, since a fixed set of rules is used instead of learning patterns from the data.

- **Likely failure modes:** TPU codes may fail during CPU or GPU runtimes, and GPU usage limits in Colab may interrupt runs in the program. Thus, this requires modifying the original pipeline. Additionally, biases affect the model's performance. This may occur when a headline may belong to more than one category or the dataset size is reduced.

2.2 AI Pipeline

Your Pipeline

- **Models used:** DistilBert model is used as the AI pipeline model.
- **Pipeline stages:** The first stage of the pipeline is preprocessing the data. This includes loading the AG news dataset. Both datasets in training and testing are decreased to 1000 for faster runtime.

  The second stage is tokenization. At this stage, DistilBert tokenizer is defined. Using the defined tokenizer, the data is then tokenized by converting the text input into tokens.

  The third stage is training. In this stage, the model is trained using the training dataset, and hyperparameters (e.g., epoch and batch size) are adjusted.

  The next stage is testing, where the model is tested using some arbitrary data.

  Evaluation is the last stage of the pipeline. This involves visualizing the confusion matrix and calculating accuracy, precision, and recall. These are helpful for analyzing the correct and wrong predictions of the model.
- **Design choices and justification:** To execute a faster run time, the dataset was decreased to 1000 samples. Due to limited runtime and hardware in Colab, the model was executed on CPU instead of GPU.

3 Experiments

3.1 Datasets

You may use a small public dataset (e.g., from `datasets`) or construct your own. In this section, describe:

- **Dataset source**: where it comes from.
- **Size**: number of examples used.
- **Splits**: how you divided train/validation/test.
- **Preprocessing**: e.g., tokenization, resizing, truncation, normalization.

Dataset Description

- **Source: News classification using HuggingFace DistilBert by Amogh Lele:** https://www.kaggle.com/code/atechnohazard/news-classification-using-huggingface-distilbert
- **Total examples:** A total of 1000 training samples and 1000 test samples were used.
- **Train/Test split:** The train/test split from the AG news dataset was used. 1000 samples were randomly selected for training and 1000 for testing.
- **Preprocessing steps:** Each news headline was tokenized using DistilBert tokenizer. The text input was converted into tokens before being fed into the model.

3.2 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.

The model was evaluated using accuracy, recall, precision, and a confusion matrix. Accuracy measures the proportion of correctly classified news headlines out of all classifications. Recall measures how well the model identifies the true labels, while precision measures how often the predicted category is correct. The confusion matrix compares the predicted and true labels to identify the correct and incorrect classifications.

3.3 Results

Report:

- Metric values for baseline vs. pipeline,
- A results table,
- At least three qualitative examples.

Method	**Accuracy**	**Recall**	**Precision**
Baseline	0.330	0.330	0.470
AI Pipeline	0.893	0.890	0.890

The results show that the AI pipeline performed significantly better than the baseline due to its higher values in accuracy, recall, and precision. This reveals that the AI pipeline is more effective in understanding context based on news headline classification.

Qualitative Examples:
Headline: New evidence of virus risks from wildlife trade
True Label: Science-Technology News
Predicted Label: Science-Technology News

Headline: Coronavirus: Bank pumps £100bn into UK economy to aid recovery
True Label: Business News
Predicted Label: Business News

Headline: Coronavirus: Trump's bid to end Obama-era immigration policy ruled unlawful
True Label: World News
Predicted Label: World News

Overall, the model correctly classified the headlines, revealing that it understood the context behind each headline.

4 Reflection and Limitations

Reflection

The original pipeline ran 20 epochs in a dataset that contains 120,000 samples. Due to the limited runtime in Colab, the dataset was reduced to 1000 samples. Hence, the number of epochs was reduced to four to prevent the risk of overfitting. Therefore, the AI pipeline was modified and debugged to run the model smoothly. The model's accuracy was close to 0.90, suggesting that it is highly accurate. However, misclassifications may still occur. In the future, I would try to increase the dataset size to improve its performance. Additionally, using a different dataset to test the model's flexibility would be an interesting task.

5 References

Google Developers. (n.d.). Classification: Accuracy, precision, recall, and related metrics. Machine Learning. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall#recall_or_true_positive_rate