Professional Report Summary

This study and analysis employed the Reuters news dataset sourced from the NLTK corpus. The dataset encompassed 90 distinct categories, with each document potentially being associated with multiple classes, thereby constituting a multilabel classification task. Notably, upon meticulous scrutiny, it was revealed that approximately 85 percent of cases exclusively comprised a sole label. Consequently, two distinct methodologies were pursued to address this situation.

The first approach involved treating the multilabel records as instances of single-label multiclass classification. This entailed converting the multilabel instances into the singular label that occurred most frequently. Additional alternative approaches have been detailed within the accompanying Jupyter notebooks.

The second approach entailed constructing a Multi Label Classification model.

For the single-label classification task, two strategies were explored: the bag of words approach and the Spacy pipeline method. Prior to implementation, the dataset underwent thorough cleansing and preprocessing. It was deemed appropriate to exclude labels with fewer than 50 instances from consideration, a criterion applied to both the single-class and multiclass model datasets.

The Spacy pipeline text classification model yielded promising results, achieving a weighted average F1 score of 0.92 in the context of single-class classification. In the realm of multilabel classification, the Hamming loss metric was adopted, and the model demonstrated a commendable hamming loss of 0.0085.

Moving forward, there are avenues for enhancing the model's performance. Leveraging transformer embeddings and GPU acceleration, both supported by Spacy, holds promise. Furthermore, Spacy accommodates recent advancements in Large Language Models (LLMs) and offers compatibility with third-party adapters capable of harnessing the capabilities of these expansive models. However, owing to present constraints pertaining to time and resources, the implementation of these steps remains pending for this particular submission.

Appendix

- The data directory houses the Spacy models and dataset files.
- Configuration files responsible for model construction can be found within the config directory.
- A requirements.txt file has been provided with this submission to facilitate the replication of results.