[NLP] News_articles_classif (RNN) - Project Report

1. Project Overview

This project addresses the problem of automatically classifying news articles into one of 41 categories based on their headlines and descriptions.

The goal is to build a deep learning model that reporters can use to label articles in real time.

Two main approaches are explored:

- A Simple RNN-based model

- A Transformer-based model using DistilBERT from Hugging Face

Dataset used: 'News Category Dataset' from Kaggle (/kaggle/input/news-category-dataset)

2. Model Evaluation

The models were evaluated using precision, recall, F1-score, and accuracy.

Below is a summary of the performance metrics from the best model (DistilBERT):

- Accuracy: 62%

- Macro F1-score: 0.51

- Weighted F1-score: 0.61

- Test size: 20,953 samples

- Number of categories: 41

A detailed classification report is available in the source files and README.

[NLP] News_articles_classif (RNN) - Project Report

3. Technologies Used

- Python
- PyTorch
- Hugging Face Transformers
- Simple RNN (Keras or PyTorch)
- Jupyter Notebooks
- Scikit-learn
- Pandas, NumPy, Matplotlib, Seaborn

4. Deployment & Usage

This project is structured to be easily deployed and maintained. A Streamlit or FastAPI interface can be used for reporters to input text and receive predicted categories.

A `requirements.txt` file is included to set up the environment.

For more details, refer to the README.md file on GitHub.