Political Bias in News Media

MAIS 202 Winter 2025
Roman Sejnoha
roman.sejnoha@mail.mcgill.ca

Approach

This project aims to develop a model that can accurately gauge the political bias of a given piece of news media. To do this, we used a heavily-processed version of this dataset found on Kaggle.

Table structure

The structure of the table is shown below.

Title	Link	Text	Source	Bias	Processed_Text	BERT_Embeddings

The key value here is the *Bias*, as it is the core of what we are hoping to predict. Additionally, we have tokenized the data found in *Text* and stored it in *BERT_EMBEDDINGS*. After cleaning the data and removing blank or incomplete rows, the resulting dataset contains roughly 3400 entries, which has served as a suitably large sample for training and testing.

Modeling work

Upon finishing our preliminary data cleaning and analysis, we used the sklearn train_test_split function with a 70/15/15 training/validation/test split and began testing a number of models to gauge their potential efficacy for the project. We found that a bert-base-uncased model, implemented through our BERT_Arch class, produced an MSE of 2.67, which, in the case of this model, speaks to a relatively accurate, robust model. Additionally, the f1-score calculated for the model, though inconsistent, generally ranges between 0.55 and 0.71, indicating a generally precise model. Given these promising results, we've made only minor alterations since deliverable 2.

Summary: Full Stack in Use

- **Data:** pandas, scikit-learn
- **Tokenizer & Model Backbone:** HuggingFace's transformers (BERT)
- **Neural Network:** PyTorch (nn.Module, layers, optimizers, losses)
- **Training Utilities:** PyTorch DataLoader, TensorDataset, sampling

- **Evaluation:** scikit-learn metrics
- **Device Management:** torch.device

In the end, we are left with a model that is generally capable of classifying the political lean of a given document between left, lean left, center, lean right, and right. Moving forward, were we to continue working on this project, we would work to test other models and to develop more bespoke models for this task that might be even more accurate.