**AG NEWS TEXT CLASSIFICATION REPORT**

1. **Introduction**

The task of this project is to classify news articles from the AG News dataset into one of four categories: World, Sports, Business, and Science/Technology. The dataset consists of over 120,000 (50000 rows are used in our model 40000 for training 10000 for validation ) news articles that are split into a training set and a test set 7600.

Dataset Link:- [https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset](%20https:/www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset)

1. **Preprocessing**

The first step in preprocessing the data is to read in the CSV files using Pandas. We then use the Hugging Face Transformers library to tokenize the text data and convert it into numerical form. We used the BERT model for tokenization, which involves breaking the text into subwords and mapping each subword to a unique numerical representation.

1. **Model Architecture and Fine-tuning**

We used the BERT model for text classification, which is a smaller and faster version of the BERT model. We fine-tuned the model on the AG News dataset using the AdamW optimizer with a learning rate of 2e-5 and trained it for 5 epochs. We used the BertForSequenceClassification class from the Hugging Face Transformers library to create the model.

1. **Evaluation Metrics and Results**

We evaluated the performance of the model on the test set using the accuracy metric. The accuracy is the number of correct predictions divided by the total number of predictions. Our model achieved an F1 score of 93.18% on the test set.

1. **Discussion**

Our model achieved a high F1 score on the test set, indicating that it is able to classify news articles into the correct category with a high degree of accuracy. One possible way to improve the model is to use a larger and more powerful pre-trained language model such as BERT or RoBERTa. Another way to improve the model is to use an ensemble of models, where multiple models are trained on different subsets of the data and their predictions are combined to produce a final prediction also we can improve using train model on whole dataset it will give better results.

1. **Sample Predictions**

Here are some sample predictions made by the model on the test set:

* "Apple unveils new iPhone": Science/Technology
* " virat kohli scored 100 in test match": Sports
* "Oil prices surge as tensions rise in the Middle East": Business

1. **Code and Dataset**

The code and dataset used for this project can be found on

Dataset:-  [https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset](%20https:/www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset)

GitHub: <https://github.com/Santoshuske/Text-Classification->