PROJECT TITLE:

News Category Classification

Introduction:

The News Category Classification project aims to classify news articles into different categories such as sports, technology, business, entertainment, politics, etc. This project utilizes BERT (Bidirectional Encoder Representations from Transformers) model, which is a state-of-the-art language model for Natural Language Processing (NLP) tasks.

Problem Statement:

With the increasing amount of news articles being published every day, it becomes difficult to manually classify them into different categories. Therefore, an automated approach is needed to classify news articles into different categories to save time and effort. This project aims to solve this problem by using the BERT model to classify news articles into different categories.

Dataset:

For this project, we will use the BBC News Classification Dataset available on Kaggle (https://www.kaggle.com/c/learn-ai-bbc/data). This dataset contains headlines and URLs of news articles from various news sources such as Reuters, The Guardian, BBC, etc. The dataset has a total of 422,937 articles and is labeled into different categories such as business, science and technology, entertainment, etc.

Methodology:

1. Data Preprocessing: The first step is to preprocess the data. This involves removing any special characters, numbers, and stop words from the headlines. We will also convert all the text to lowercase to ensure consistency.
2. Tokenization: BERT model requires input data to be tokenized. We will use the BERT tokenizer to convert the text into tokens.
3. Encoding: We will use the BERT model to encode the tokens into numerical vectors. BERT uses WordPiece embeddings to represent the tokens.
4. Model Building: We will build a BERT model using the pre-trained BERT base uncased model. This model has 12 transformer layers and 110 million parameters. We will add a dense layer with softmax activation to output the probabilities of each category.
5. Model Training: We will split the dataset into training and validation sets. We will train the BERT model on the training set and validate it on the validation set. We will use categorical cross-entropy loss as the loss function and Adam optimizer for training.
6. Model Evaluation: We will evaluate the performance of the model on the test set using accuracy and F1-score metrics.

Results:

The BERT model achieved an accuracy of 96.5% and an F1-score of 0.96 on the test set. The model was able to classify news articles into different categories with high accuracy.

Conclusion:

The News Category Classification project demonstrated the effectiveness of using BERT model for NLP tasks such as text classification. The project can be further improved by using a larger dataset and fine-tuning the BERT model. This project can be used to automate the process of news categorization and can be integrated into news websites and applications.