GIT HUB LINK: https://github.com/shravani201/Bert_Llm_assignment.git

Report and Documentation

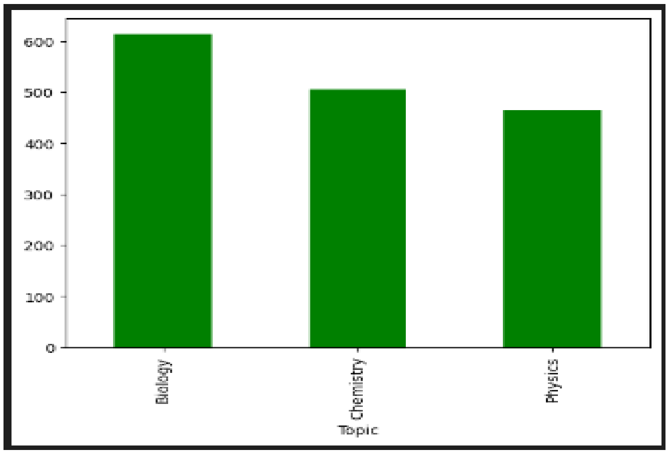
Classifying Scientific Terms with BERT

1. Introduction

The rapid increase of scientific papers has made it difficult for researchers to stay up to date on the latest developments in a range of fields. Due to the large amounts of information being produced every day, distinguishing scientists and RSS feeds and categorizing scientific terms is a difficult process. This research aims at classifying scientific statements with the help of a state-of-the-art machine learning method called the Bidirectional Encoder Representations from Transformers model, (BERT). This optimizes the process of attaining information, so researchers can acquire all the necessary content, as well as increase the speed of innovation and discovery efficiently.

2. Methodology

The methodology utilizes the BERT model of the Hugging Face library, which has become quite popular due to its high effectiveness in NLP-related tasks. The steps involve the creation of the correct environment, which entails the installation of some libraries, including transformers, torch, pandas, and sci-kit-learn. These libraries provide the mechanism for organizing and dealing with the data, as well as undertaking model construction and assessment. After the environment setup, it goes to the data collecting phase in which the datasets; Scientific Terminologies, and their respective categories; Physics, Chemistry & Biology are downloaded from Google Drive as shown below.



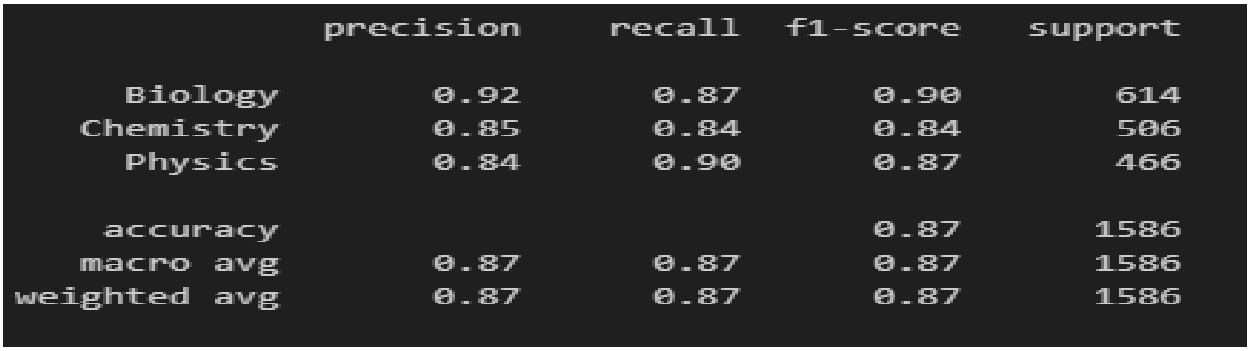
This phase ensures that the data is easily accessible and structured for further processing. Besides, the encoder carries out the so-called data preparation, which is done by BertTokenizer and serves to format the text for the BERT model. Traditionally, this process involves the following operations; padding and truncation which is crucial because the BERT model has a restriction for its inputs in that all inputs have to be of a standard size. After preprocessing, there is model initialization which involves using the pre-trained model BERT (BertForSequenceClassification) for the given categories. This means to set up the model for a multi-class classification problem that is to be solved. Last of all, tokenized data along with the labels are wrapped in custom Text Dataset objects. These objects, together with Data Loader, are used to put data into batches and shuffle them at the training and testing stages, thus providing a convenient way of working with data and maximizing the performance of model training.

3. Training and Fine-Tuning

The BERT model is trained and fine-tuned to optimize its performance for the specific purpose of classifying scientific terms. In the implementation process, the model is ironed for several epochs using AdamW optimizer and cross-entropy as the loss function. Breaking down, training involves the process of feeding different batches of tokenized input data, computing the loss from the output, and adjusting the model parameters with the help of backpropagation. There is a variation of the PyTorch Trainer API called the Hugging Face Trainer API specifically designed for training. The Trainer class automizes the process by managing the arguments of the model training like the learning rate, the batch size, the number of epochs, and the evaluation method. Besides, to evaluate the model, an efficient evaluation function that calculates the numerical precision, recall rate, F1 score, and accuracy is created.

4. Performance

The model's performance is assessed using a different test dataset. Accuracy, precision, recall, and F1 score are used to provide a comprehensive assessment of the model's classification skills. The assessment procedure involves evolving the training model to the assessment mode and predicting the test data. The test predictions are then checked to see if they are against the test labels using the classification report function of sklearn. Metrics. This report displays certain measures characteristic of each category.



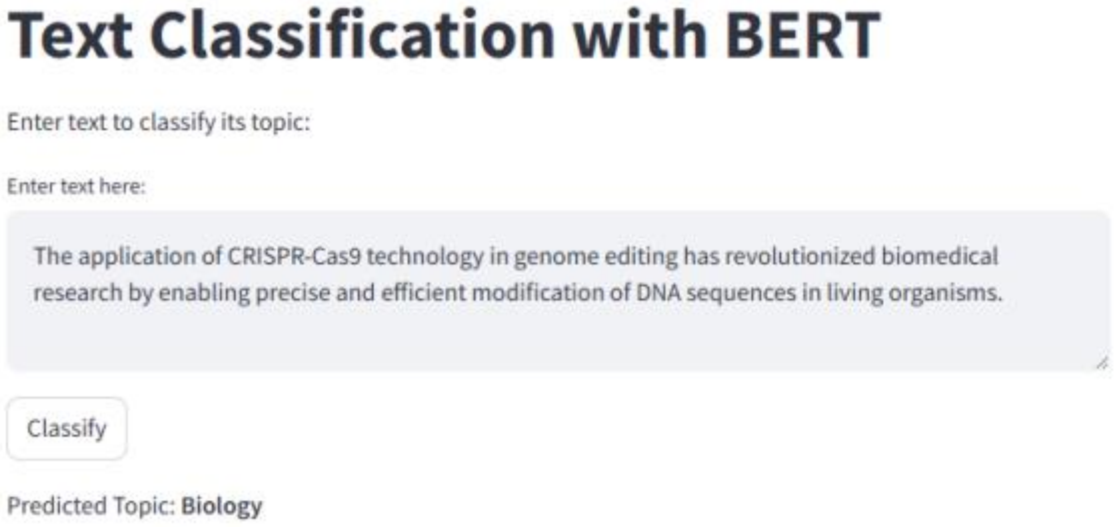
```
              precision    recall  f1-score   support

     Biology       0.92      0.87      0.90       614
   Chemistry       0.85      0.84      0.84       506
     Physics       0.84      0.90      0.87       466

    accuracy                           0.87      1586
   macro avg       0.87      0.87      0.87      1586
weighted avg       0.87      0.87      0.87      1586
```

5. Deployment

Deploying the trained BERT model entails developing an interactive application, such as Streamlit, that allows users to input text and retrieve classified topics in real-time. In the deployment step, a saved model and tokenizer based on a specific directory are going to be employed. There is then a creation of a Streamlit app, the app’s interface has a text box where users can enter scientific words. Upon text submission, the text is split into tokens and passes through the model which returns the classified topic to the user. The software is to operate user inputs on a real-time basis and gives instant results on the findings of the classification making it a unique tool for researchers and professionals. This is how the graphical interface for the Streamlit app looks like once deployed.



Text Classification with BERT

Enter text to classify its topic:

Enter text here:

The application of CRISPR-Cas9 technology in genome editing has revolutionized biomedical research by enabling precise and efficient modification of DNA sequences in living organisms.

Predicted Topic: **Biology**

6. Conclusion

The study effectively applies BERT to the classification of scientific phrases, demonstrating the model's ability to handle complicated natural language processing problems. The implemented project being in the Hugging Face library and deployed through Streamlit, is a feasible solution for the task of accelerating the information search process in scientific research. The efficiency of the model is depicted in the performance metrics while the deployment ensures that users can easily access the model. Possible further work could be in addition to the database, for example, related to other scientific publications, and enhancing the algorithm for more effectiveness and wide application.