ASSIGNMENT-1

BERT + Explainability Resources

Please refer the following resources for implementation:

1. https://www.mzes.uni-mannheim.de/socialsciencedatalab/article/bert-explainable-ai/
2. https://towardsdatascience.com/bert-for-dummies-step-by-step-tutorial-fb90890ffe03
3. https://medium.com/@kalia_65609/interpreting-an-nlp-model-with-lime-and-shap-834ccfa124e4
4. https://github.com/KaliaBarkai/KaggleDisasterTweets

The objective of this assignment is to explore the application of BERT (Bidirectional Encoder Representations from Transformers) for text classification tasks and to gain insights into the model's predictions through explainability techniques.

Tasks:

- Select a suitable text classification dataset (e.g., sentiment analysis, topic classification, etc.) and provide a brief description of the dataset.
- Preprocess the data, including tasks like tokenization, padding, and handling of labels.
- Use a pre-trained BERT model (e.g., `bert-base-uncased`) from a library like Hugging Face Transformers.
- Fine-tune the BERT model on the selected dataset for the chosen classification task.
- Evaluate the performance of the fine-tuned BERT model on a test set using appropriate metrics (e.g., accuracy, precision, recall, F1-score).
- Introduce explainability techniques (e.g., LIME, SHAP, attention mechanisms) and explain their importance in gaining insights into the model's predictions.
- Apply at least two explainability techniques to analyze the predictions made by the BERT model on the test set.
- Provide visualizations and interpretations of the explanations.
- Compare the explanations generated by different techniques and discuss any patterns or insights gained from the explanations.

Deliverables:

- Python script containing the code for data preprocessing, BERT implementation, and explainability techniques application.
- A report summarizing the methodology, results, and insights obtained from the experiment.