MRP: RESULTS

Abobakr Al-kaf

JULY 26, 2024
TORONTO METROPOLITAN UNIVERSITY

Exploratory Analysis Results

The MedQuAD dataset is a complex and diverse medical question-answer dataset. It contains questions categorized into 5126 unique focus areas, indicating a highly granular classification task.

Key Observations:

- The dataset is highly imbalanced, with many focus areas having very few samples.
- The questions are varied in length and complexity, requiring careful preprocessing and tokenization.

Experiment 1: Logistic Regression Model

Objective:

- To establish a baseline for classification using a simple Logistic Regression model.

Actions:

- Preprocessed the data using basic text cleaning and tokenization.
- Used TF-IDF vectors as features for the Logistic Regression classifier.
- Trained the model and evaluated its performance on a validation set.

Results:

- The logistic regression model struggled with the high dimensionality and imbalance of the dataset.
- Achieved low precision, recall, and F1-scores across most categories.
- The model could not handle the complexity and variability of the data effectively.

Experiment 2: DistilBERT Model

Objective:

- To improve classification performance using a pre-trained DistilBERT model.

Actions:

- Used the DistilBERT tokenizer for text preprocessing and tokenization.
- Fine-tuned the DistilBERT model on the MedQuAD dataset for 3 epochs.
- Evaluated model performance at each epoch.

Results:

- DistilBERT significantly outperformed the Logistic Regression model, showing higher precision, recall, and F1-scores.
- However, the model still faced challenges with the imbalance and diversity of the dataset.
- Performance improved with each epoch, indicating that further fine-tuning could yield even better results.

Experiment 3: LSTM Model

Objective:

- To further explore the effectiveness of a neural network approach using an LSTM model.

Actions:

- Tokenized the text data and built a vocabulary.
- Encoded the text and labels for input to the LSTM model.
- Built and trained an LSTM model with padding and packing of sequences for efficient training.
- Evaluated model performance at each epoch.

Results:

- The LSTM model showed improvement over Logistic Regression but did not outperform DistilBERT.
- The model faced difficulties in handling the highly imbalanced data.
- Performance metrics indicated that while the model learned some patterns, it struggled with generalizing to the validation set.

Discussion

- **Experiment 1**: The Logistic Regression model provided a weak baseline, unable to handle the complexity of the MedQuAD dataset.
- **Experiment 2**: DistilBERT showed significant improvement, leveraging pre-trained knowledge and fine-tuning to adapt to the dataset. It managed to capture more patterns and nuances in the data.
- **Experiment 3**: The LSTM model, while better than Logistic Regression, did not match the performance of DistilBERT. The LSTM's ability to handle sequences was beneficial, but the imbalance in the dataset posed a challenge.

Overall, pre-trained transformer models like DistilBERT demonstrated superior performance in handling large and complex medical datasets compared to traditional machine learning models and even some deep learning models like LSTMs. Future work should focus on addressing class imbalance and further fine-tuning transformer models to achieve better performance.