Report on Enhanced Strategies for Scraping and Analyzing Reddit Data for Sentiment Analysis

https://github.com/srivikas777/Sentiment-Analysis-on-Reddit-Data-using-ClinicalBERT.git

Introduction

Sentiment analysis plays a crucial role in understanding public opinion and perception, especially in online forums like Reddit. However, accurately gauging sentiment becomes challenging when analyzing comments independently of their parent posts, leading to a loss of context. Moreover, inadequate data collection often results in overfitting, diminishing the reliability of sentiment analysis results.

Issue Identification

The primary issue identified was the loss of context during sentiment analysis due to the separation of comments from their parent posts. Contextual cues from the post are essential for interpreting the sentiment expressed in comments accurately. Additionally, insufficient data collection led to significant overfitting, compromising the reliability of sentiment analysis results.

Brainstorming Solutions

To address these issues, various solutions were considered:

- Thread Aggregation: Combining comments with their parent posts to preserve context.

- Contextual Embedding: Utilizing NLP techniques to capture contextual information.

- Feature Engineering: Incorporating additional features representing context.

- Hierarchical Models: Implementing models capable of handling hierarchical data structures.

- Sentiment Cue Extraction: Developing a framework to extract sentiment cues.

- Negation Handling: Improving negation handling in sentiment analysis.

- Testing other models.

- Experimenting with different hyperparameters and settings.

Approach Evolution

The approach evolved towards non-model-based techniques to address sentiment analysis challenges. Key changes implemented include:

1. Lexicon Expansion: Incorporating domain-specific terms related to clinical trials into the lexicon.

2. Data Structure Modification: Nesting comments under their respective posts to maintain context.

3. Code Refinement: Adjusting code snippets for accurate data processing.

Exploration of Models

Models like ClinicalBERT and Bio_ClinicalBERT were experimented with to enhance sentiment analysis accuracy.

Hyperparameter Tuning

Hyperparameters were adjusted to optimize model performance:

- Learning Rate Adjustment: Exploring different rates to optimize convergence.

- Batch Size Optimization: Varying batch sizes considering memory constraints.

- Increased Number of Epochs: Extending training epochs for better learning.

Addressing Class Imbalance

Techniques like dataset rebalancing, class weights assignment, and resampling were employed to mitigate class imbalance effects.

Early Stopping and Model Checkpoints

Early stopping mechanisms and periodic model checkpoints were implemented to prevent overfitting and ensure progress retention during training.

Conclusion

Despite efforts to enhance model performance, unexpected increases in validation loss were observed post-modifications. Further analysis is being conducted to understand and address this issue. Continuous experimentation and fine-tuning are being undertaken to improve the effectiveness and accuracy of the sentiment analysis system. However, it's noted that the model still exhibits tendencies to overfit exponentially. Limited data and the complexity of the model are identified as major concerns in this regard. To mitigate these issues, the best possible settings are being utilized, considering all available resources.

The revised approach aims to preserve context and refine data processing techniques to capture sentiment nuances accurately. This report outlines key issues, brainstormed solutions, and specific changes made to improve sentiment analysis on Reddit data. The focus remains on enhancing sentiment analysis without solely relying on machine learning models. Continued vigilance and adaptation are crucial for achieving desired improvements.