Sentiment Analysis

Introduction

This project explores sentiment analysis using a BERT-based classifier model on a custom dataset generated with ChatGPT. The goal was to evaluate how well BERT performs in classifying sentiment into positive, negative, and neutral categories and to analyze challenges that arise from using synthetic data and model limitations.

Methodology

- **Model**: We used the BERT base model from Hugging Face Transformers as the backbone for sentiment classification.

- **Data**: The dataset was synthetically generated using ChatGPT prompts to simulate social media posts. Each post was labeled with a sentiment category.

- **Training**: The model was trained for 20 epochs using a standard training loop in PyTorch. Training and test accuracy were logged at regular intervals.

- **Evaluation**: Evaluation was based on accuracy across training and testing sets, and additional qualitative analysis of misclassifications was conducted.

Results

- **Training Accuracy**: The model achieved high training accuracy, exceeding 90% by epoch 20.

- **Test Accuracy**: The test accuracy plateaued and started to decline slightly, indicating potential overfitting.

- **Error Analysis**:

    - Many neutral classifications were mislabeled due to subtle cues in language not being effectively captured.

    - Positive and negative sentiments were sometimes incorrectly classified as neutral.

    - Certain words typically associated with strong sentiment were overlooked due to context sensitivity.

Main Ideas and Insights

- **ChatGPT-generated data**: While convenient, the dataset's reliability is questionable. Synthetic posts may not capture the nuanced linguistic features of real-world sentiment expression.

- **Label Noise**: Errors in sentiment labeling (especially subtle or sarcastic content) led to performance bottlenecks.

- **Model Limitations**: Despite fine-tuning, BERT occasionally failed to disambiguate context-driven sentiment cues.

- **Overfitting**: The divergence between training and test accuracy in later epochs indicates that regularization or early stopping may be necessary.

Comparison with Other Research

Compared to benchmarks using real-world datasets like IMDb or SST-2:

- **Accuracy**: Our model's test accuracy was lower than models trained on cleaner, well-labeled datasets.

- **Data Quality**: Synthetic data lacks the diversity and unpredictability of real user-generated content, impacting generalization.

- **Model Behavior**: Other studies often implement ensemble methods or larger models (e.g., RoBERTa, XLNet) with better handling of context and subtleties.

Conclusion

This project highlights the trade-offs between dataset accessibility and model performance. While synthetic data enables rapid prototyping, it introduces reliability concerns that must be addressed for production-level sentiment analysis. Improvements could include using real-world datasets, implementing data augmentation, refining labeling strategies, or employing ensemble models for robustness.