[Applied Analytics Project Week 13 Assignment]

Major: Applied Analytics

Name: Heng Bai, Yuang Guo

Narrative Documentation for Sentiment Analysis Project

Introduction

Our project focuses on classifying Twitter sentiments into four categories: Positive, Negative, Neutral, and Irrelevant. By leveraging TF-IDF for feature extraction and a neural network for classification, we have built a robust pipeline for analyzing and predicting sentiment in textual data. Throughout the term, we experimented extensively with various models and embedding techniques, striving to apply our NLP knowledge to achieve the best possible performance. This journey deepened our understanding of the mathematical representation of words and provided invaluable insights, making this project an enriching learning experience from start to finish.

Workflow Overview

1. **Setup and Initialization**:
    - The environment is configured by Google Drive for data access and importing necessary libraries for data manipulation, NLP, and modeling.

2. **Data Loading and Preprocessing**:
    - Datasets are loaded and prepared with sentences in string format.
    - The preprocessing pipeline includes:
        - Text cleaning (removing noise like punctuation and stopwords).
        - Lemmatization for standardizing words.
        - Feature extraction using TF-IDF, capturing the importance of terms across documents.

3. **Model Architecture**:
    - The model is built using TensorFlow's Sequential API with the following layers:
        - **Input Layer**: Accepts TF-IDF features.
        - **Hidden Layers**: Three dense layers with ReLU activations and dropout for regularization.
        - **Output Layer**: A softmax activation layer outputs probabilities for four classes.

4. **Training and Evaluation**:
    - The model is trained on preprocessed data with early stopping to prevent overfitting.

    - Metrics like accuracy are calculated on test data to evaluate performance. Visualizations of metrics in seaborn of confusion matrix and classification report at last.

Highlights and Strengths

- **Effective Feature Engineering**: The use of TF-IDF ensures strong representation of text data, critical for sentiment analysis tasks.
- **Model Regularization**: Dropout layers minimize overfitting risks, ensuring better generalization.
- **Multi-Class Classification Capability**: The architecture is well-suited for handling multiple sentiment categories.

Suggestions for Improvement

Our project focuses on gaming-related topics on Twitter, which often include new terms and ambiguous language. Keeping up with the rapidly evolving vocabulary and context of the gaming community presents significant challenges. To address this, we could consider methods such as dynamic vocabulary updates using pre-trained embeddings (e.g., GloVe or Word2Vec), fine-tuning transformer models like BERT on domain-specific data, or employing contextual word representations to capture nuanced meanings. These approaches would help us adapt to the dynamic nature of gaming discourse and improve the accuracy of our sentiment analysis.

Conclusion

Our project effectively demonstrates a full pipeline for sentiment analysis using TF-IDF and a neural network. With slight refinements in metrics, explainability, and dataset handling, this project could serve as an excellent foundation for practical applications in social media analytics. Appreciate to Savas and Elifcan for patience and guidance in this term and all colleagues for reading and commenting on our project.