Sarcasm Detection Using Fine-Tuned Transformer Models

Introduction:

Sarcasm detection is a challenging Natural Language Processing (NLP) task, which involves NLP tools guessing whether the intended meaning of a sentence differs from its literal interpretation. Sarcasm detection asks NLP tools to understand contextual cues, sentiment contrast, and pragmatic knowledge, which is a huge leap from just grasping the meaning of each word in a sentence. Earlier approaches try to overcome the problem by using static word embeddings, such as Glove or Word2Vec, with classical machine learning classifiers. However, they become outdated with the advent of transformer-based architectures.

The transformer-based architectures use self-attention mechanisms to model contextual dependencies across an entire input sequence. Therefore, it can understand the context of a sentence better to determine whether it's sarcasm. In this project, I'm going to train and evaluate a transformer-based sarcasm detection model and compare its performance against classical baselines. Through quantitative evaluation and error analysis, the project seeks to demonstrate the advantages and limitations of transformer-based models for identifying sarcastic language.

Data Source:

The project will use the **Sarcasm Headlines Dataset** from Kaggle as the training and testing dataset. The dataset contains news headlines annotated as sarcastic or non-sarcastic, and it's chosen because it's clean, balanced, and widely used as a benchmark.

Methods, Techniques, and Technology

Traditional Baseline Methods
- Implement static word embeddings with Glove and Word2Vec.
- Train the model with a simple neural network.
- The model will serve as a baseline

Transform-Based Methods
- Select a pretrained transformer model from HuggingFace (bert-base-uncased or roberta-base)
- Fine-tune the model for binary sarcasm classification using labeled training data.

Evaluation

- Compare F1-score, accuracy, precision, and recall between the baseline model and the transform-based model.

Technology
- HuggingFace Transformers, static word embeddings, neural network building and training.

Products to Be Delivered

The product of this project is a fine-tuned transformer-based sarcasm detection model, trained on a labeled sarcasm dataset.

The trained model will demonstrate the following achieved goals:
- Improved contextual understanding of sarcastic language compared to static embedding-based classifiers.
- Higher classification performance on standard evaluation metrics.
- Practical applicability for sarcasm detection tasks without reliance on paid large language model APIs.

The performance characteristics and limitations of the trained model will be documented and analyzed as part of the final submission.