Project Proposal
Sarcasm Detection in News Headlines

Problem Selection

The focus is on sarcasm detection within news headlines, a problem with significant implications in natural language processing (NLP). Sarcasm can skew the sentiment analysis and automated information extraction if not correctly identified. This project aims to mitigate such issues by accurately classifying sarcastic statements. This problem is especially pertinent in today's social media-heavy landscape, where textual content is a primary communication medium.

Dataset

The dataset for this project will be sourced from Kaggle: "News Headlines Dataset for Sarcasm Detection". It contains a compilation of news headlines from two sources: The Onion, which is known for sarcastic headlines, and HuffPost, which provides genuine headlines. This dataset provides a balanced ground for training and testing sarcasm detection models.

NLP Methods:

For NLP methods, the proposal includes:
- Baseline Model: A classical NLP approach using a Bag-of-Words (BoW) model combined with a Naive Bayes classifier will be established first.
- Advanced Model: Post the baseline, a deep learning model using BERT will be implemented. BERT's pre-trained context-aware embeddings are expected to significantly enhance sarcasm detection accuracy.

Packages:

The following Python packages will be utilized:
- `NLTK`: For natural language tasks like tokenization.
- `Scikit-learn`: For implementing the baseline model and various preprocessing techniques.
- `TensorFlow` and `Transformers` by Hugging Face: For leveraging pre-trained BERT models and developing custom layers if necessary.

These are chosen for their robustness, ease of use, and comprehensive functionalities, which are conducive to the project's success.

NLP Tasks:

In-depth NLP tasks will involve:
- Preprocessing: Employing techniques like lowercasing, lemmatization, and removal of stop words and punctuation.
- Vectorization: Using TF-IDF for the baseline and token embeddings for the BERT model.
- Model Development: Iteratively designing and fine-tuning the architecture.
- Validation: Employing cross-validation techniques to ensure the model's generalizability.

Performance Metrics:
A detailed evaluation framework will include:
- Accuracy: As a primary metric for a high-level model performance overview.
- Precision, Recall, F1-Score: To thoroughly evaluate model performance, especially in handling class imbalances.
- AUC-ROC Curve: To assess the model's ability to discriminate between the classes.
- Loss Metrics: To monitor and guide the training process effectively.

Project Schedule:
The schedule will be detailed to ensure accountability at each phase:
- Weeks 1-3: Detailed dataset analysis, including sarcasm pattern identification,Baseline model development
- Weeks 3-5: Implementation and fine-tuning of the BERT model.
- Week 6: Comprehensive analysis using established metrics.
- Week 7: Finalization of the report, creation of a presentation, and a rehearsal for the project defense.
This more detailed schedule is designed to ensure thoroughness and quality in the project's execution.