Project Proposal

Sarcasm Detection in News Headlines

Team Members
Namratha Prakash
Shashank Shivakumar

Introduction

Sarcasm detection within news headlines is crucial for improving NLP tasks such as sentiment analysis, information extraction, and media reliability assessment. Misinterpreting sarcasm can result in skewed analysis, especially on platforms like social media where text is a primary communication medium. This project seeks to accurately classify sarcastic headlines, enhancing the precision of sentiment models and benefiting applications in content moderation, automated sentiment monitoring, and public sentiment analysis.

Objective

The main objective of this project is to develop and evaluate advanced NLP models to detect sarcasm in news headlines more effectively than conventional approaches. This will involve a comparative analysis between traditional models and state-of-the-art neural network techniques to identify the optimal approach for sarcasm detection in news media.

Data Source

This project will use the *News Headlines Dataset for Sarcasm Detection*, available on Kaggle, which includes headlines from *The Onion* (sarcastic) and *HuffPost* (non-sarcastic). This balanced dataset is ideal for training and evaluating models due to the clear distinction between sarcastic and genuine content.

Methodology

1. **Data Preprocessing**:
   Initial text cleaning will be conducted to handle elements like punctuation and stopwords, followed by tokenization and lemmatization for standardized text processing.

2. **Feature Engineering**:

   - **Textual Features**: Extract basic text features like word counts, character counts, and n-grams.
   - **Semantic Features**: Use TF-IDF and word embeddings (e.g., BERT embeddings) to capture context-aware semantics for sarcasm detection.

3. **Model Development**:

   - **Baseline Model**: Implement a Bag-of-Words (BoW) model combined with a Naive Bayes classifier as a benchmark.
   - **Advanced Models**: Implement BERT-based neural network models, leveraging pre-trained embeddings to capture the nuances in sarcastic language.

4. **Evaluation**:

   - Use metrics like accuracy, precision, recall, F1-score, and the AUC-ROC curve for performance assessment.
   - Cross-validation will ensure model robustness and generalizability across datasets.

Tools and Technologies

- **Programming Language**: Python
- **Libraries**:
    - **NLTK**: For text preprocessing (tokenization and lemmatization).
    - **Scikit-learn**: For implementing and evaluating the baseline model.
    - **TensorFlow & Hugging Face Transformers**: For building, fine-tuning, and implementing the BERT model.
- **Cloud Platforms**: Consider using AWS or Google Cloud for potential large-scale deployment.

Expected Outcomes

- **Model Performance**: Enhanced sarcasm detection accuracy over baseline methods, enabling more accurate sentiment analysis.
- **Insights**: Recommendations for using sarcasm detection to refine content moderation strategies, improve automated sentiment analyses, and support NLP applications for social media and news platforms.
- **NLP Contribution**: Findings on the effectiveness of neural embeddings in detecting sarcasm across diverse news headlines.

Project Schedule

- **Weeks 1-3**: Dataset analysis and baseline model development.
- **Weeks 3-5**: Implementation and fine-tuning of the BERT model.
- **Week 6**: Model evaluation and analysis using selected metrics.
- **Week 7**: Final report preparation, presentation creation, and rehearsal for the project defense.

Conclusion

This project aligns with current NLP research objectives by applying advanced techniques in sarcasm detection to a real-world dataset. The findings have the potential to enhance media analysis accuracy and contribute valuable insights into the field of sarcasm detection in textual data.

Dataset Link: https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection/data