NLP Project

Charan Reddy Kumar, Tenneti Srinivas Saiteja

Project Proposal: Duplicate Question Detection on Quora Using Advanced NLP Techniques

Introduction

Quora serves as a global platform where over 100 million users engage in asking and answering questions across diverse topics. A prevalent challenge on the platform is the redundancy of questions with similar intents but varying phrasings. This not only dilutes the quality of content but also hampers users' ability to find the most relevant answers efficiently. Currently, Quora employs a Random Forest model to identify and merge duplicate questions. However, with advancements in Natural Language Processing (NLP), there is potential to improve this system significantly.

Objective

The primary goal of this project is to develop and evaluate advanced NLP models to detect duplicate questions on Quora more effectively than the existing Random Forest approach. By leveraging state-of-the-art techniques introduced in our course, such as word embeddings, neural networks, and deep learning algorithms, the project aims to enhance the accuracy of duplicate detection, thereby improving user experience on the platform.

Data Source

The analysis will be based on the "Quora Question Pairs" dataset available from the Kaggle competition. This dataset comprises over 400,000 pairs of questions with annotations indicating whether the question pairs are duplicates. The rich and sizable dataset provides an excellent foundation for training and evaluating complex models.

https://www.kaggle.com/competitions/quora-question-pairs/overview

Methodology

1. **Data Preprocessing**: Clean and normalize the text data to handle issues like misspellings, punctuation, and stopwords. Techniques such as tokenization, lemmatization, and stemming will be applied.
2. **Feature Engineering**:
    - **Textual Features**: Extract features like word counts, character counts, and n-grams.
    - **Semantic Features**: Utilize word embeddings (e.g., Word2Vec, GloVe, Google's Universal Sentence encoder) to capture semantic similarities between questions.
    - **Distance Metrics**: Compute similarity measures such as cosine similarity and Jaccard distance between question pairs.
3. **Model Development**:
    - **Baseline Model**: Replicate the Random Forest model for benchmarking.
    - **Advanced Models**:
        - **Deep Neural Networks**: Implement models like Siamese LSTM networks to capture sequential dependencies in text.

        - **Transformer-Based Models**: Explore the use of BERT or RoBERTa for contextual embeddings.
        - **Ensemble Methods**: Combine multiple models to improve overall performance.
4. **Evaluation**:
    - Use metrics like accuracy, F1-score, precision, and recall to evaluate model performance.
    - Perform cross-validation to ensure the robustness of results.
5. **Deployment Considerations**:
    - Discuss scalability and computational efficiency for real-world application.
    - Consider the implementation of the model on a cloud platform to leverage distributed computing resources.

Tools and Technologies

- **Programming Languages**: Python
- **Libraries**: TensorFlow, PyTorch, scikit-learn, NLTK, spaCy
- **Cloud Platform**: AWS or Google Cloud for scalable computing and storage

Expected Outcomes

- **Improved Model Performance**: Demonstrate enhanced accuracy and reliability over the existing Random Forest model.
- **Actionable Insights**: Provide recommendations for Quora to implement more sophisticated duplicate detection mechanisms.
- **Contribution to NLP**: Offer valuable findings on the effectiveness of advanced NLP techniques in semantic similarity tasks.

Potential Impact

By improving duplicate question detection, the project can significantly enhance the user experience on Quora. Seekers will find high-quality answers more efficiently, and contributors can focus on providing unique insights without redundancy. The methodologies developed could also be applicable to other platforms facing similar challenges, such as forums and customer support systems.

Conclusion

This project aligns with the course objectives by applying advanced NLP methods to a real-world dataset. It offers an opportunity to delve deep into machine learning algorithms and contribute meaningfully to the field of question-answering systems. All analyses will be conducted on a cloud platform to ensure scalability and reproducibility.