Hallucination Detection with Logistic Regression

Lightweight text classifier for factuality analysis in summaries

Overview

This project aims to detect hallucinations in automatically generated text summaries by classifying them as **factual** or **hallucinated**. It utilizes a custom logistic regression model implemented from scratch, and employs the Bag of Words (BoW) technique for feature extraction. The approach demonstrates that even simple interpretable models can perform competitively on factuality detection tasks.

Objectives

- Detect hallucinated summaries using binary classification.
- Build a custom logistic regression model without external ML libraries.
- Extract textual features using the Bag of Words approach.
- Evaluate model performance using standard metrics.
- Analyze errors for qualitative understanding.

Process

1. **Data Loading:** Imported the XSum Hallucination dataset using `pandas`, selecting `summary` and `is_factual` fields.
2. **Text Cleaning:** Performed basic preprocessing and tokenization.
3. **Vectorization:** Constructed the BoW vocabulary and transformed summaries into numerical feature vectors.
4. **Model Development:** Implemented logistic regression using NumPy with sigmoid activation and gradient descent.
5. **Training & Evaluation:** Trained the model and validated using accuracy, precision, recall, and F1-score.
6. **Cross-Validation:** Applied k-fold cross-validation to ensure generalizability.

Methodology

Binary Classification: Each summary was mapped to a binary label indicating whether it was factual (`1`) or hallucinated (`0`).

Model Structure: A scratch-built logistic regression model was trained using cross-entropy loss and optimized via gradient descent.

Evaluation Metrics: Accuracy, precision, recall, and F1-score were calculated using `scikit-learn`.

Error Analysis: Misclassified summaries were logged and analyzed to identify recurring hallucination patterns and edge cases.

Results

The model demonstrated strong performance despite its simplicity. Accuracy and F1 scores were consistent across validation folds. Analysis of errors revealed common hallucination indicators such as vague terminology, incorrect entities, or abstract generalizations.

Key Learnings

- Simpler models like logistic regression can yield meaningful insights and results in NLP tasks.
- Bag of Words remains a strong baseline for text classification.
- Implementing from scratch improves understanding of optimization and model behavior.
- Error patterns can guide further improvements in preprocessing and feature engineering.

Dataset

XSum Hallucination Annotations Dataset
Each data point consists of:

- `summary`: Text summary (input).
- `is_factual`: Binary label (0 or 1).

Tools and Technologies

- **Language:** Python
- **Libraries:** pandas, NumPy, scikit-learn (for evaluation only)

Contribution

This project is open to collaboration. Feature suggestions and contributions are welcome. Feel free to open issues or submit pull requests.