IMDB Reviews: Sentiment Analysis

Team 25

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Nikola Dobricic | Milan Stankovic  (in front of a random forest) | Ioana Popescu |
| * LSTM Model * Embedding methods * Bert Model | * Model * Model Performance * Random Forests | * Data Analysis * Bow Model * Model Performance |

**Project Overview: Predicting Movie Review Ratings**

**Objective:** Develop and compare various models to predict numerical movie review ratings (from 1 to 10) from their textual content.

**Models Compared:**

1. **Linear Regression with Bag of Words**
2. **Random Forests with GloVe Embeddings**
3. **LSTMs**
4. **BERT**

**Key Steps:**

* **Data Preprocessing:**
  + Cleaning text (removing HTML tags, punctuation, special characters).
  + Tokenization and removal of stop words.
* **Feature Extraction:**
  + Bag of Words and GloVe Embeddings.
* **Model Training:**
  + Linear Regression, Random Forest, LSTM, and BERT.
* **Model Evaluation:**
  + Metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), Precision, Recall.

**Findings:**

* **Bag of Words:** Simple but ineffective, lacking context understanding.
* **Random Forest with GloVe:** Balanced performance, but limited by lack of context comprehension.
* **LSTMs:** Superior in capturing sequential dependencies and context, handling varying input lengths.
* **BERT:** Best performance, leveraging deep contextual understanding and advanced pre-trained architecture.

**Dataset:**

* 50,000 IMDB movie reviews from Stanford Large Movie Review dataset, used for extending binary sentiment classification to continuous rating prediction.

**Applications:**

* Useful for marketing and film companies to analyze public opinion on movies from social media reviews, aiding in sentiment analysis and market research.

**Future Steps:**

* Fine-tune models on social media comments.
* Pre-train LSTMs on large datasets for improved performance.

**Limitations:**

* Dataset specificity may limit generalizability.
* Did not explore review length or reviewer credibility impacts.
* Omitted detailed hyperparameter tuning and model interpretability analysis.

**Project Repository:** [GitHub](https://github.com/NDobricic/imdb-sentiment-analysis)