Warming up the Mic: Build your own speech to text engine

Week–2 Coding Task Assignment Report

Contents

Mentor: Aditiya Sanapala
Mentee: Jayent Dev
Roll Number: 24B1232

Abstract

This report presents an experimental study of classical and deep learning approaches for sentiment analysis on the IMDB movie review dataset. We compare a traditional TF-IDF feature-based Logistic Regression classifier with a Long Short-Term Memory (LSTM) neural network that operates on word embeddings. The objective is to understand the strengths and limitations of sparse representations versus dense, sequence-aware models in Natural Language Processing (NLP).

1 Introduction

Natural Language Processing (NLP) aims to enable machines to understand and generate human language. Early NLP systems relied heavily on hand-crafted features and frequency-based representations such as Bag-of-Words and TF-IDF. While effective, these methods fail to capture semantic relationships and word order.

Recent advances in deep learning have introduced dense word embeddings and sequence models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. These models capture contextual information and long-range dependencies in text. This report evaluates and compares these two paradigms through a sentiment classification task.

2 Dataset

The IMDB movie review dataset was used for all experiments. It consists of:

- 25,000 training reviews
- 25,000 test reviews
- Binary sentiment labels: positive (1) and negative (0)

The dataset was downloaded using the Hugging Face `datasets` library and stored locally in CSV format for compatibility with different modeling pipelines.

3 Word Embeddings

Word embeddings map discrete words into continuous vector spaces where semantic relationships are preserved geometrically.

3.1 Word2Vec

Word2Vec learns word representations by predicting context words given a target word (Skip-gram) or predicting a target word from its context (CBOW). Words that appear in similar contexts obtain similar vector representations.

These embeddings were trained on the IMDB training corpus and later used for visualization and as input representations for neural models.

3.2 Embedding Visualization

To qualitatively analyze the learned embeddings, dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-SNE were applied. The resulting plots demonstrate semantic clustering of related words, validating the effectiveness of the learned representations.

4 Models

4.1 TF-IDF + Logistic Regression

TF-IDF represents documents as sparse vectors that encode word importance based on term frequency and inverse document frequency. Logistic Regression is a linear classifier trained on these vectors to predict sentiment labels.

This model serves as a strong and fast baseline.

4.2 LSTM Sentiment Classifier

The LSTM model processes reviews as sequences of word indices. An embedding layer maps words to dense vectors, which are then fed into an LSTM network. The final hidden state is passed through a fully connected layer with a sigmoid activation to produce a sentiment probability.

Unlike TF-IDF, the LSTM explicitly models word order and long-range dependencies.

5 Experimental Setup

- Optimizer: Adam
- Loss Function: Binary Cross-Entropy
- Evaluation Metrics: Accuracy, F1-score

Both models were evaluated on the same test set to ensure fair comparison.

6 Results

Table 1: Performance Comparison of Models

Model	**Accuracy**	**F1-Score**	**Training Time**
TF-IDF + Logistic Regression	0.88376	0.88361	Very Fast
LSTM Classifier	0.51656	0.65105	Slow

The LSTM model achieves higher accuracy and F1-score due to its ability to model sequential dependencies and semantic relationships between words. However, this improvement comes at the cost of increased training time and computational complexity.

7 Discussion

The results highlight a fundamental trade-off in NLP systems. Classical models such as TF-IDF combined with linear classifiers are computationally efficient and surprisingly effective. However, they lack the ability to understand word order and deeper semantic structure.

LSTM-based models, although more expensive to train, provide better performance by leveraging dense representations and temporal modeling. This makes them more suitable for tasks requiring contextual understanding.

8 Conclusion

In this work, we compared traditional and deep learning approaches for sentiment analysis. While TF-IDF with Logistic Regression provides a strong and efficient baseline, LSTM-based models demonstrate superior performance by capturing contextual information.

This study motivates the use of more advanced architectures such as attention mechanisms and transformers, which further improve performance by addressing the limitations of recurrent models.