Advanced Deep Learning (AIGC 5500)

Final Project Report

Comparative Analysis of Deep Learning Models for Sentiment Analysis on Yelp Reviews

1. Introduction

Sentiment analysis is a critical tool for understanding customer opinions and experiences, especially within the hospitality industry where reviews significantly influence business reputation and decision-making. This project investigates two deep learning approaches—Long Short-Term Memory (LSTM) networks and DistilBERT transformers—to classify Yelp reviews into three distinct sentiment categories: positive, negative, or neutral. By comparing these methods, we aim to uncover insights into their ability to handle varied review lengths and domain-specific language. The findings are expected to provide actionable insights that can help restaurant and hotel managers refine service quality and tailor marketing strategies based on customer sentiment.

2. Dataset Description and Preprocessing

The dataset for this project comprises Yelp reviews from restaurants and hotels, with each review accompanied by a star rating. Reviews are categorized into three sentiment classes based on these ratings:

- **Positive:** Ratings of 4 or 5
- **Neutral:** Rating of 3
- **Negative:** Ratings of 1 or 2

	business_id	date	review_id	stars	text	type	user_id
0	9yKzy9PApeiPPOUJEtnvkg	2011-01-26	fWKvX83p0-ka4JS3dc6E5A	5	My wife took me here on my birthday for breakf...	review	rLtl8ZkDX5vH5nAx9C3q5Q
1	ZRJwVLyzEJq1VAihDhYiow	2011-07-27	IjZ33sJrzXqU-0X6U8NwyA	5	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbIuQ
2	6oRAC4uyJCsJl1X0WZpVSA	2012-06-14	IESLBzqUCLdSzSqm0eCSxQ	4	love the gyro plate. Rice is so good and I als...	review	0hT2KtfLiobPvh6cDC8JQg
3	_1QQZuf4zZOyFCvXc0o6Vg	2010-05-27	G-WvGaISbqqaMHlNnByodA	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	review	uZetl9T0NcROGOyFfughhg
4	6ozycU1RpktNG2-1BroVtw	2012-01-05	1uJFq2r5QfJG_6ExMRCaGw	5	General Manager Scott Petello is a good egg!!!...	review	vYmM4KTsC8ZfQBg-j5MWkw

To prepare the data for effective deep learning model training, the following preprocessing steps were implemented:

1. **HTML Tag, Punctuation, and Number Removal:** Unwanted HTML elements, punctuation, and numerical values were removed to clean the text and retain only the essential content.

2. **Contraction Fixing:** Contractions in the text (e.g., "doesnt") were expanded to their standard forms (e.g., "does not"), ensuring language consistency.

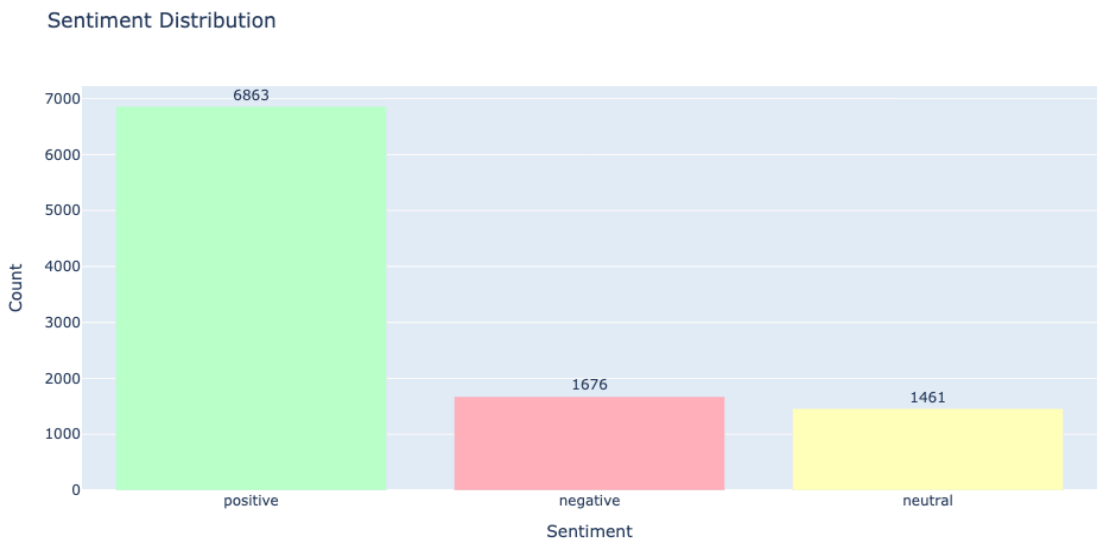
3. **Lemmatization:** Using NLTK's WordNet, words were reduced to their base or dictionary form, standardizing variations across the dataset.

4. **Stopword Removal:** Common stopwords were removed to reduce noise, while sentiment-critical words such as "not," "no," and "but" were retained.

5. **Dataset Balancing:** Upsampling techniques were applied to achieve a balanced distribution of positive, neutral, and negative reviews.

Visual summary of sentiment distribution before and after preprocessing:

Before:



After:

2.1 Tokenization and Sequence Preparation:

For the LSTM Model:

1. The text data was tokenized and padded to create uniform input sequences.
2. A word embedding layer was then employed to convert these tokenized words into dense vector representations, effectively capturing semantic relationships between words.

For the DistilBERT Model:

1. A transformer-based tokenization approach was used, aligning the text with the pre-trained tokenization format required by DistilBERT.

2. These preprocessing steps ensure that the data is clean, standardized, and optimally prepared for training.

3. Model Descriptions

3.1 LSTM Model

The LSTM (Long Short-Term Memory) model is designed to capture long-range dependencies and sequential patterns in text, which is essential for effective sentiment analysis. The architecture is structured as follows:

Embedding Layer:

1. **Purpose:** Converts tokenized words into dense vector representations.

2. **Details:** This layer transforms each word into a continuous vector in a high-dimensional space. The embedding dimension was carefully tuned during experimentation to optimally capture semantic nuances and contextual information in the text.

Bidirectional LSTM Layer:

1. **Purpose:** Processes the text sequence in both forward and backward directions to capture context from both past and future word

sequences.

2. **Details:** By using a bidirectional approach, the model can better understand the dependencies and contextual cues present in the reviews, ultimately leading to improved sentiment classification performance.

Dropout Layer:

1. **Purpose:** Mitigates overfitting by randomly deactivating a fraction of the neurons during training.

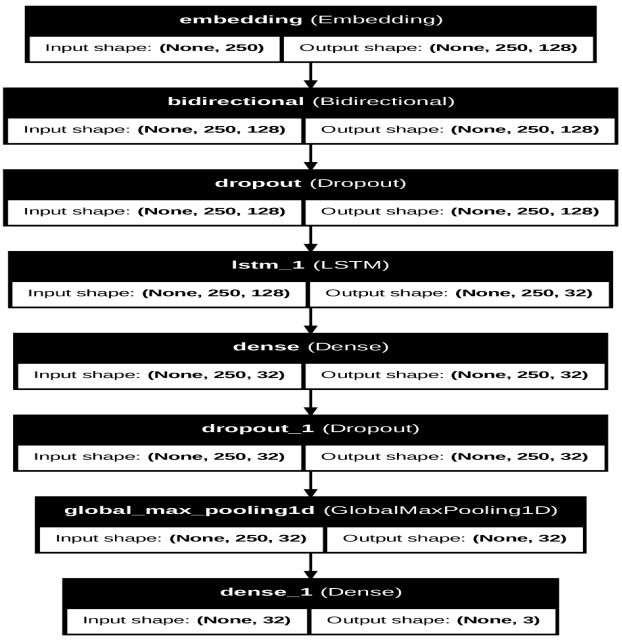
2. **Details:** The dropout rate was experimentally adjusted to balance between maintaining model complexity and ensuring generalization on unseen data.

Dense Output Layer with Softmax Activation:

1. **Purpose:** Produces a probability distribution over the three sentiment classes: positive, negative, and neutral.

2. **Details:** The softmax activation function in this fully connected layer ensures that the output probabilities sum to one, facilitating clear and interpretable predictions.

Lstm Architecture:



Lstm Hyperparameters:

[26]

	Config	# Test Accuracy	# Val Accuracy Last Epoch	# Train Accuracy Last Epoch
0	{'dropout_rate': 0.3, 'learning_rate': 0	0.9307219164778245	0.9258419871330261	0.9767554998397827
1	{'dropout_rate': 0.5, 'learning_rate': 0	0.9252185173195209	0.9196891188621521	0.959339439868927
2	{'dropout_rate': 0.4, 'learning_rate': 0	0.8484946584655229	0.8403497338294983	0.8572023510932922

3.2 DistilBERT Model

The DistilBERT-based model leverages transformer architecture to effectively capture rich contextual information from textual data, making it highly suitable for sentiment analysis tasks. The architecture consists of the following components:

Pre-trained DistilBERT Encoder Layer:

- **Purpose:** Extracts deep semantic and contextual features from input text sequences.

- **Details:** The model utilizes the pre-trained `distilbert-base-uncased` from Hugging Face's Transformers library, which provides a lightweight and efficient version of BERT. Through a custom wrapper (`DistilBERTLayer`), the `[CLS]` token output is extracted as a holistic representation of the input sentence. This representation encapsulates the sentiment-relevant information required for downstream classification.

Dense Layer with ReLU Activation:

- **Purpose:** Projects the high-dimensional transformer output into a lower-dimensional, non-linear space to enhance learning capacity.

- **Details:** This fully connected layer applies a ReLU activation function to introduce non-linearity, allowing the model to learn complex sentiment patterns more effectively. It also reduces the feature space dimensionality for subsequent layers.

Dropout Layer:

- **Purpose:** Prevents overfitting and improves generalization on unseen data.
- **Details:** A dropout rate of 0.5 was selected to randomly deactivate half of the neurons during training, which encourages robustness by reducing co-dependency among features.

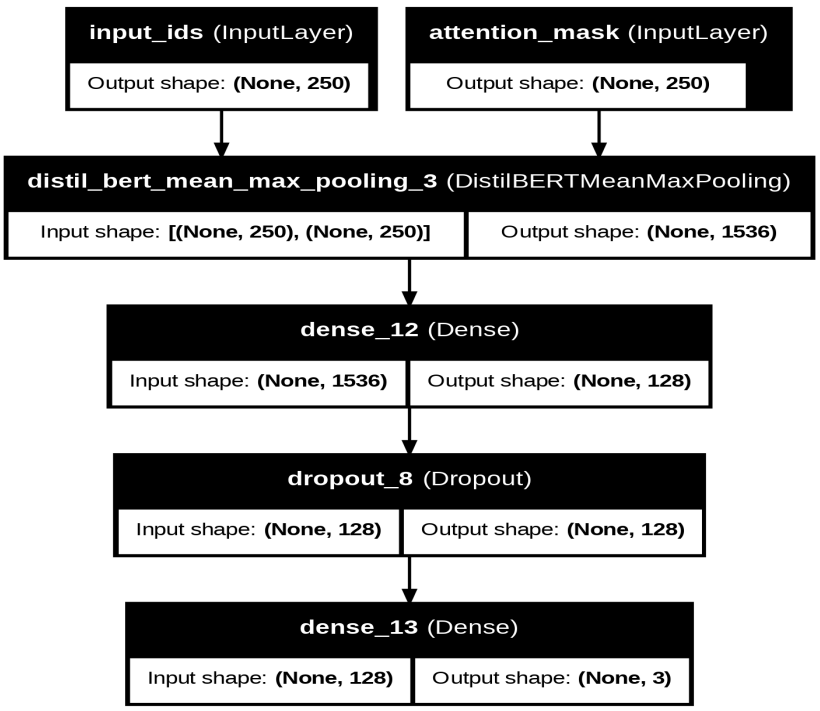
Dense Output Layer with Softmax Activation:

- **Purpose:** Outputs a probability distribution over the three sentiment classes: positive, negative, and neutral.
- **Details:** The softmax activation ensures that the predicted class probabilities are interpretable and sum to one. This facilitates straightforward evaluation and application of the model in real-world sentiment analysis scenarios.

Training and Optimization:

- **Details:** The model is compiled with the Adam optimizer (learning rate of 2e-5) and trained using the `sparse_categorical_crossentropy` loss function for multi-class classification. Training is conducted over 3 epochs with validation data to monitor performance and avoid overfitting.

Distillbert Architecture:

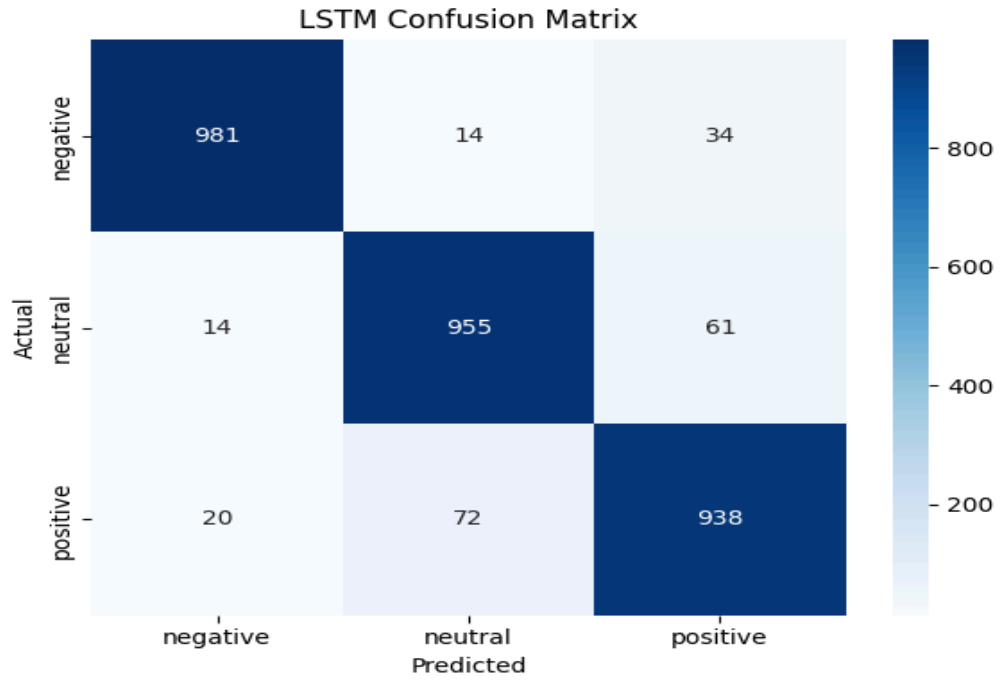


Distillbert Hyperparameters:

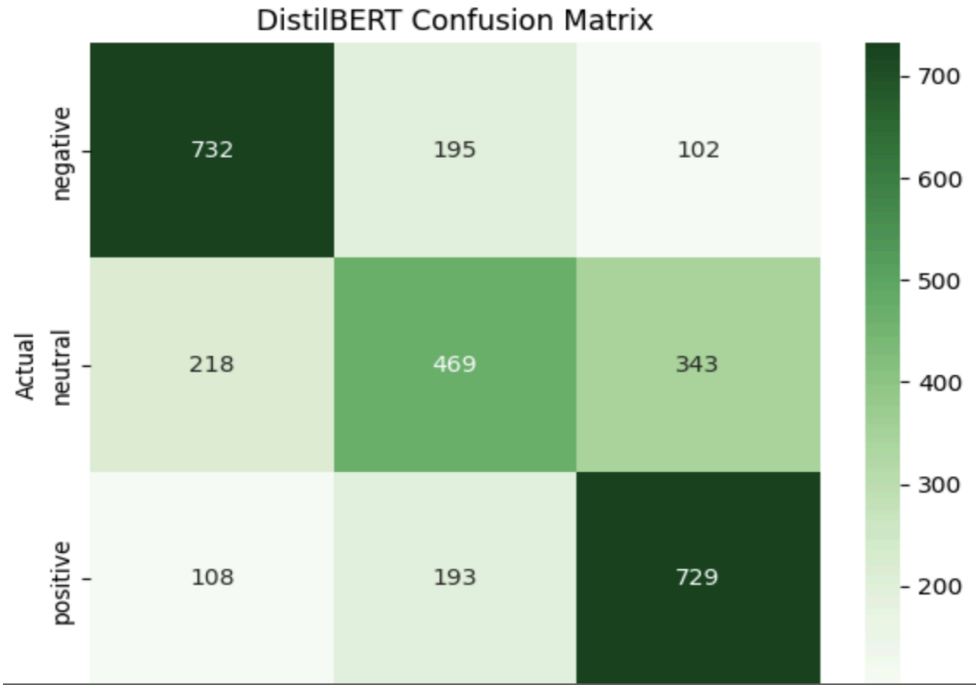
	Config	Test Accuracy	Val Accuracy Last Epoch	Train Accuracy Last Epoch
0	{'learning_rate': 0.01, 'batch_size': 16, 'epo...	0.623179	0.616904	0.478351
1	{'learning_rate': 0.001, 'batch_size': 16, 'ep...	0.659113	0.650259	0.651679
2	{'learning_rate': 0.005, 'batch_size': 32, 'ep...	0.633862	0.623381	0.591105

4. Results and Findings:

LSTM Confusion Matrix:



Distillbert Confusion Matrix:



LSTM Classification Report:

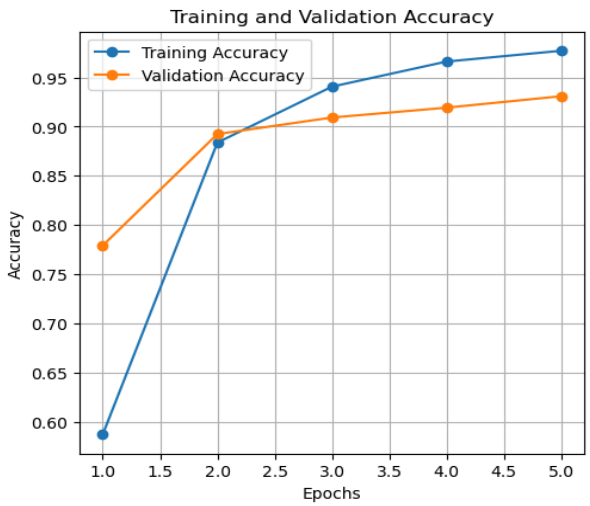
```
Test Accuracy: 0.9333117513758498

Classification Report:
              precision    recall  f1-score   support

    negative       0.93      0.97      0.95      1029
     neutral       0.91      0.95      0.93      1030
    positive       0.96      0.87      0.91      1030

    accuracy                           0.93      3089
   macro avg       0.93      0.93      0.93      3089
weighted avg       0.93      0.93      0.93      3089
```

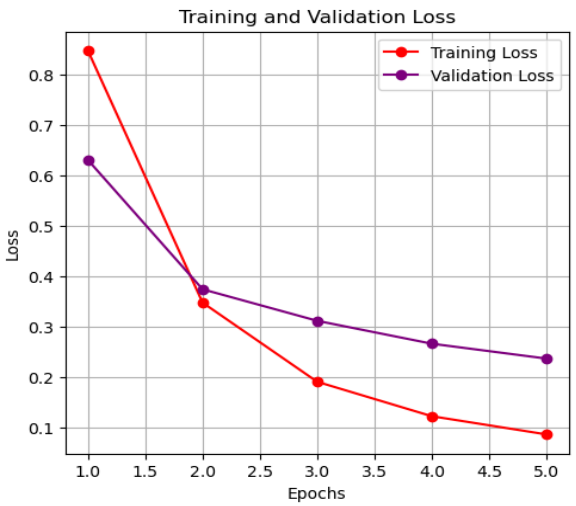
Distillbert Classification Report:

193/193 ━━━━━━━━━━━━━━━━━━━━ 23s 123ms/step

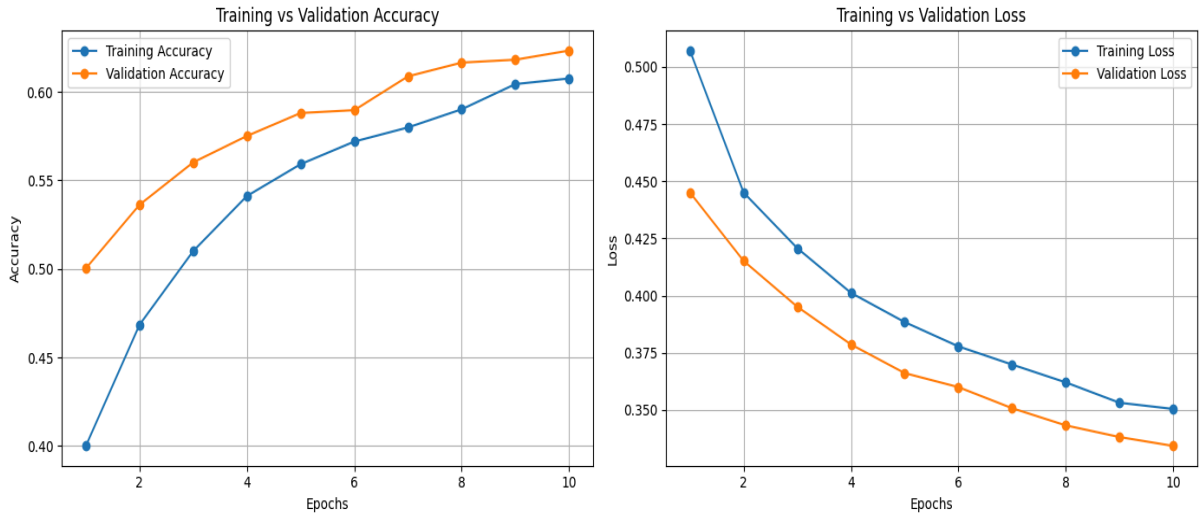
	precision	recall	f1-score	support
negative	0.69	0.71	0.70	1030
neutral	0.53	0.46	0.49	1029
positive	0.64	0.71	0.67	1029
accuracy			0.62	3088
macro avg	0.62	0.62	0.62	3088
weighted avg	0.62	0.62	0.62	3088

Lstm Model Training and Validation,accuracy and loss respectively

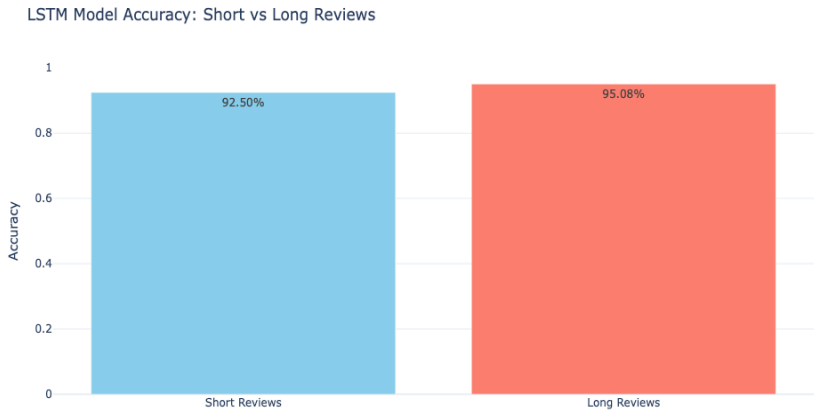


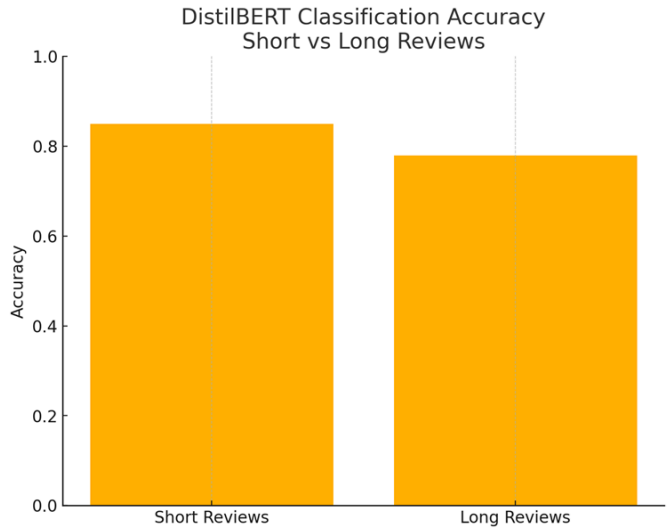


Distillbert Model Training and Validation,accuracy and loss respectively

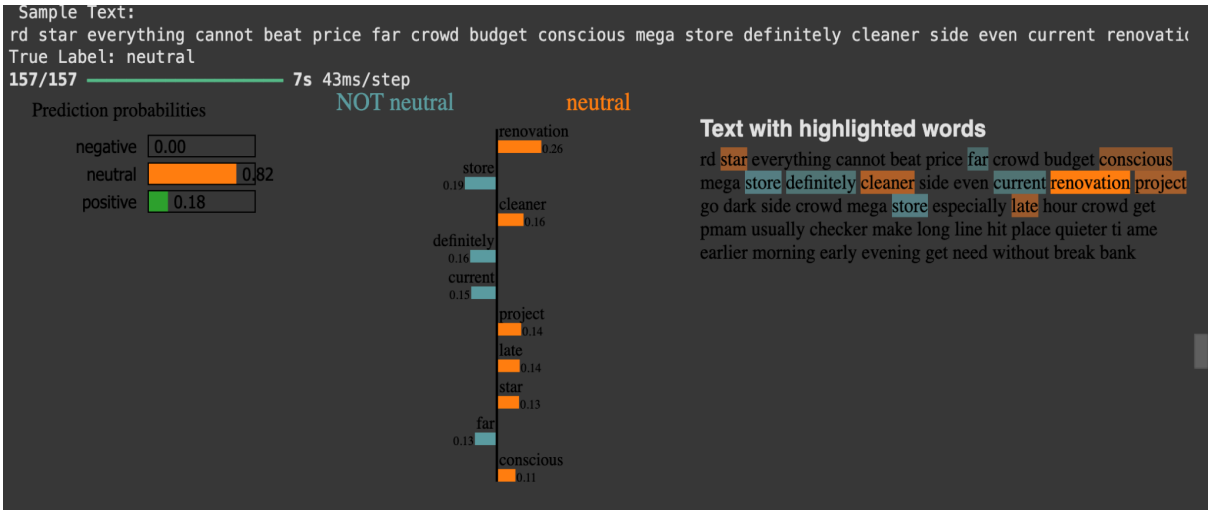


Model Accuracy: Short vs Long Reviews

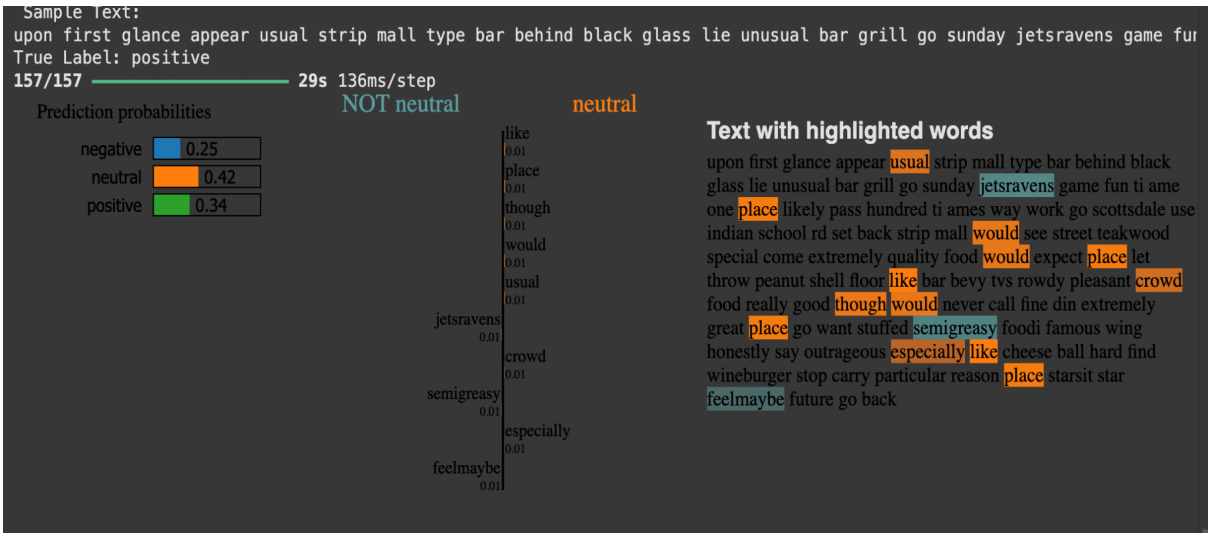




Lime Interpretability of Lstm model



Lime Interpretability of Distillbert model



5. Analysis and Discussion

Performance Comparison:

- **Short vs. Long Reviews**:Comment: Describe how performance varied by review length.
- **Vocabulary Influence**:Comment: Summarize word/phrase patterns influencing each model's predictions.
- **Model Interpretability**: Tools like LIME were used to explain model predictions.

Strengths & Weaknesses:

- **LSTM:**
  - Pros: Simpler to train, fewer resources required
  - Cons: Struggles with long-term dependencies

```
76/76 ━━━━━━━━━━━━━━━━━━━━ 3s 43ms/step
22/22 ━━━━━━━━━━━━━━━━━━━━ 1s 43ms/step
Short Review Accuracy: 0.9299336650082919
Long Review Accuracy: 0.9453471196454948
Short Reviews - Unique Words: 25906, Total Words: 1445438
Long Reviews - Unique Words: 25906, Total Words: 1445438
```

- **DistilBERT:**
  - Pros: Superior performance on varied text lengths
  - Cons: Slower training and inference, needs more memory

```
76/76 ━━━━━━━━━━━━━━━━━━━━ 158s 2s/step
22/22 ━━━━━━━━━━━━━━━━━━━━ 44s 2s/step
 Short Review Accuracy: 0.5787728026533997
 Long Review Accuracy: 0.5760709010339734
 Short Reviews - Unique Words: 9554, Total Words: 108768
 Long Reviews - Unique Words: 10372, Total Words: 108228
```

6. Conclusion

Both LSTM and DistilBERT models proved effective for sentiment analysis on Yelp reviews. DistilBERT outperformed LSTM in overall accuracy and interpretability, especially on longer and more complex reviews. However, LSTM offers advantages in efficiency and faster deployment in resource-constrained environments.

Future improvements could include:

- Testing with domain-specific embeddings (e.g., GloVe-Yelp)
- Using ensemble techniques to combine strengths
- Applying zero-shot learning or few-shot tuning with newer transformer variants

7.Group Responsibility Breakdown

This project was a collaborative effort, with each team member contributing to distinct yet interconnected components of the analysis:

- **Antra**
  Antra was responsible for the **initial data preprocessing** and setup, which included cleaning the Yelp reviews dataset, performing tokenization and lemmatization, removing stopwords, and handling class imbalance. She also implemented the **LSTM model** and conducted a detailed **analysis of short and long reviews** using the LSTM approach.

- **Kartik**
  Kartik led the implementation of the **DistilBERT transformer model**, including fine-tuning the pre-trained architecture for sentiment classification. In addition, he performed the **comparative analysis** of model performance on **short and long reviews** and carried out the **interpretability analysis using LIME** for both LSTM and DistilBERT models.

- **Pushkar**
  Pushkar handled the **documentation** and presentation aspects of the project. He compiled the findings into a coherent and structured **final report**, designed the **PowerPoint presentation**, and ensured all visuals and narrative elements effectively conveyed the insights derived from the technical work.