Movie Review Classification: Sentiment Analysis

This presentation explores the development and deployment of a movie review classification model leveraging sentiment analysis, a critical aspect of natural language processing.

Presented by:

- Sarah Osama Mohamed Elfishawy.
- Mahmoud Attya Mohamed Attya.
- Yasmin Hamdy Mahmoud.
- Lilian Stephanos Yonan.
- Osama Hafez Fathy Hafez.

Problem definition

The project's objective is to create a sentiment analysis model that can effectively classify movie reviews from the IMDb dataset as positive or negative.

1. **Sentiment Analysis**

   Sentiment analysis plays a crucial role in various applications, including market analysis, brand monitoring, and gauging consumer opinions.

2. **IMDb Dataset**

   The IMDb dataset, containing a vast collection of movie reviews, serves as a valuable resource for training and evaluating the model.

Movie Classification Model Flowchart

1 Data Collection

The initial stage involves gathering movie reviews from IMDb, including both the textual content of the review and the associated sentiment label (positive or negative).

2 Data Preprocessing

Data preprocessing is essential for preparing the data for model training. This step involves cleaning the data by removing noise, such as punctuation, stop words, and irrelevant characters, and converting the text into a numerical format.

3 Model Training

The model is trained on the preprocessed data, utilizing machine learning algorithms to learn patterns and relationships between the text and sentiment labels.

4 Model Evaluation

The trained model is then evaluated on unseen data to assess its accuracy and generalizability.

5 Deployment

Once the model is deemed satisfactory, it is deployed, making it available for real-time predictions on new movie reviews.

Dataset Description

The project utilizes the IMDb Reviews dataset, a widely used resource for sentiment analysis tasks.

IMDb Reviews Dataset

The dataset contains 50,000 movie reviews, each labeled as either positive or negative.

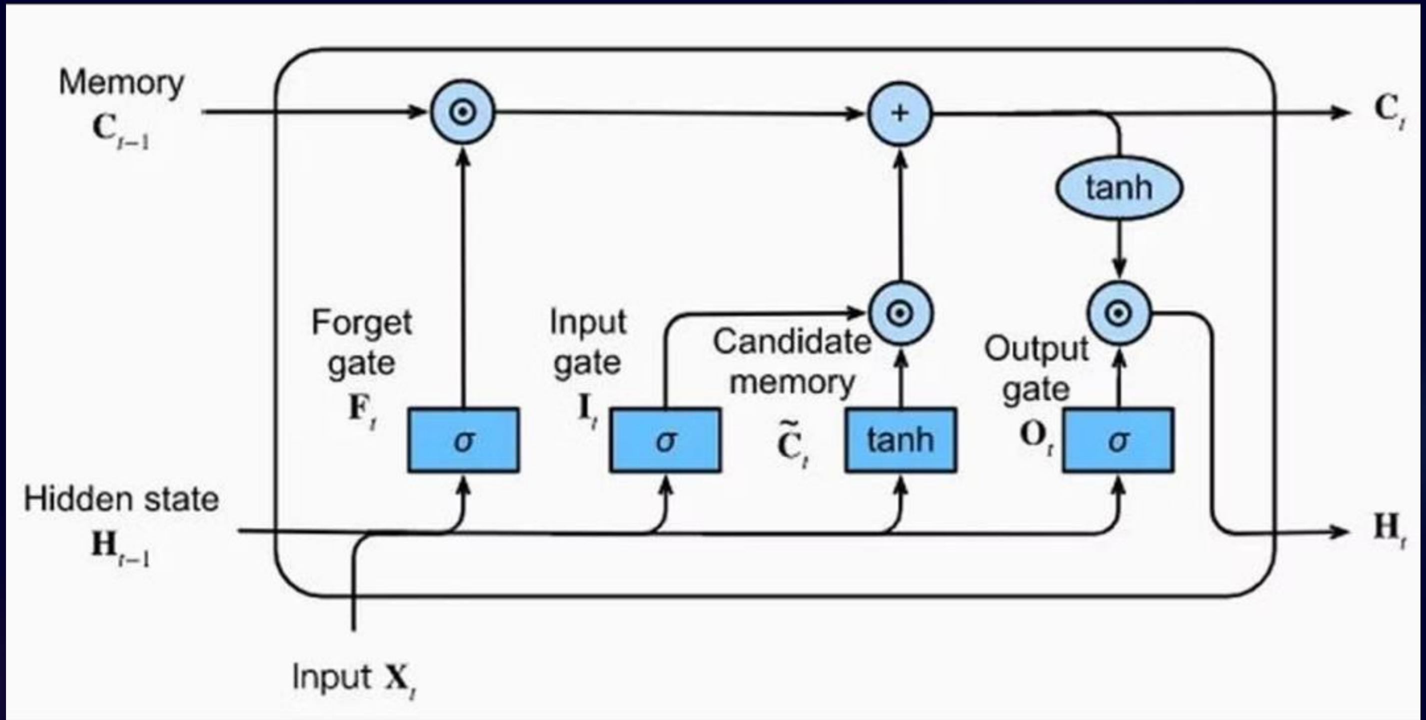
Data Split

The dataset is divided into two subsets: a training set of 25,000 reviews and a test set of 25,000 reviews. This split allows for training the model on one set and evaluating its performance on an independent set.

Textual Content & Labels

Each review consists of free-form text expressing opinions about a specific movie. The labels associated with each review indicate whether the sentiment expressed is positive (1) or negative (0).

Model Architecture



1

Embedding Layer

This layer transforms integer-encoded words into dense vector representations, enabling the model to learn semantic relationships between words.

2

LSTM Layer

Long Short-Term Memory (LSTM) layers are designed to handle sequential data, such as text, by capturing long-term dependencies.

3

Dropout Layers

Dropout layers prevent overfitting by randomly setting a fraction of the input units to zero during training, forcing the model to rely on a variety of features.

4

Dense Layer & Output Layer

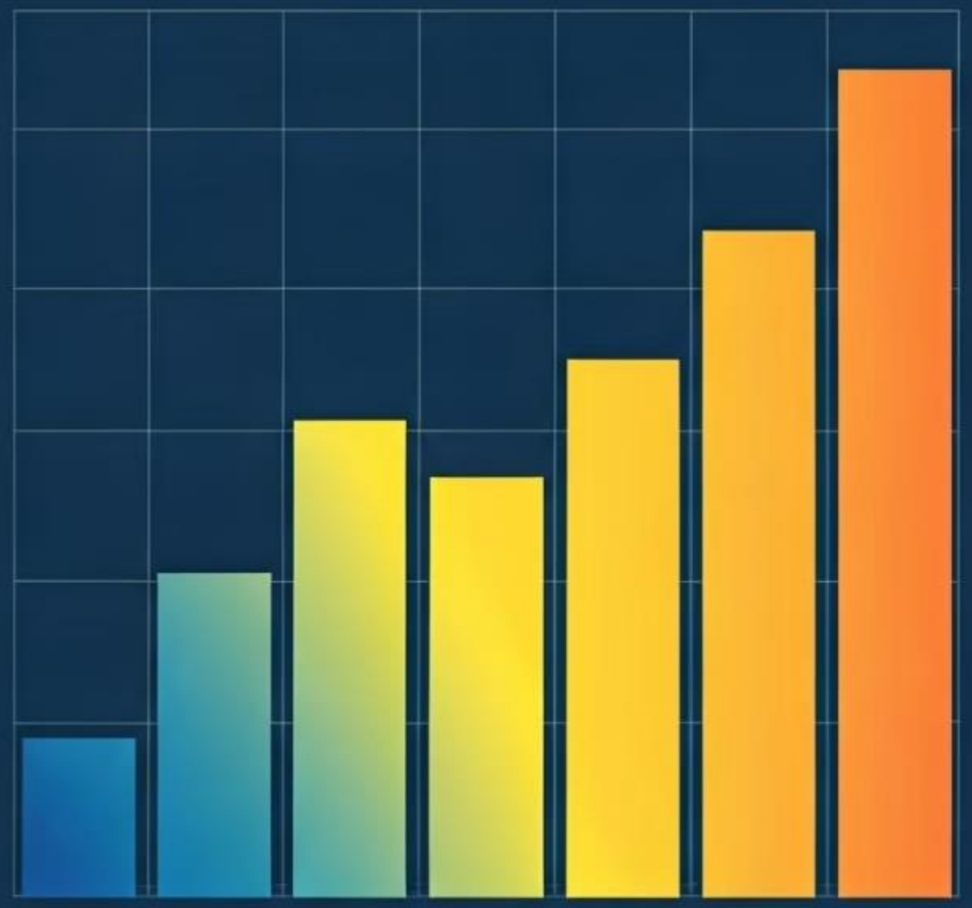
The dense layer adds a fully connected layer to learn complex representations, and the output layer provides the final classification output, predicting whether a review is positive or negative.

Training

The model is trained using a set of carefully chosen parameters to optimize its performance.

Loss Function	Binary Cross-entropy is suitable for binary classification tasks.
Optimizer	Adam optimizer, which adapts the learning rate during training.
Metrics	Accuracy to evaluate the model's performance.
Training Procedure	The model is trained for 5 epochs. The model is validated on a portion of the training data to monitor overfitting.

The model is trained for 5 epochs, and validation is performed on a portion of the training data to monitor for overfitting.



Evaluation

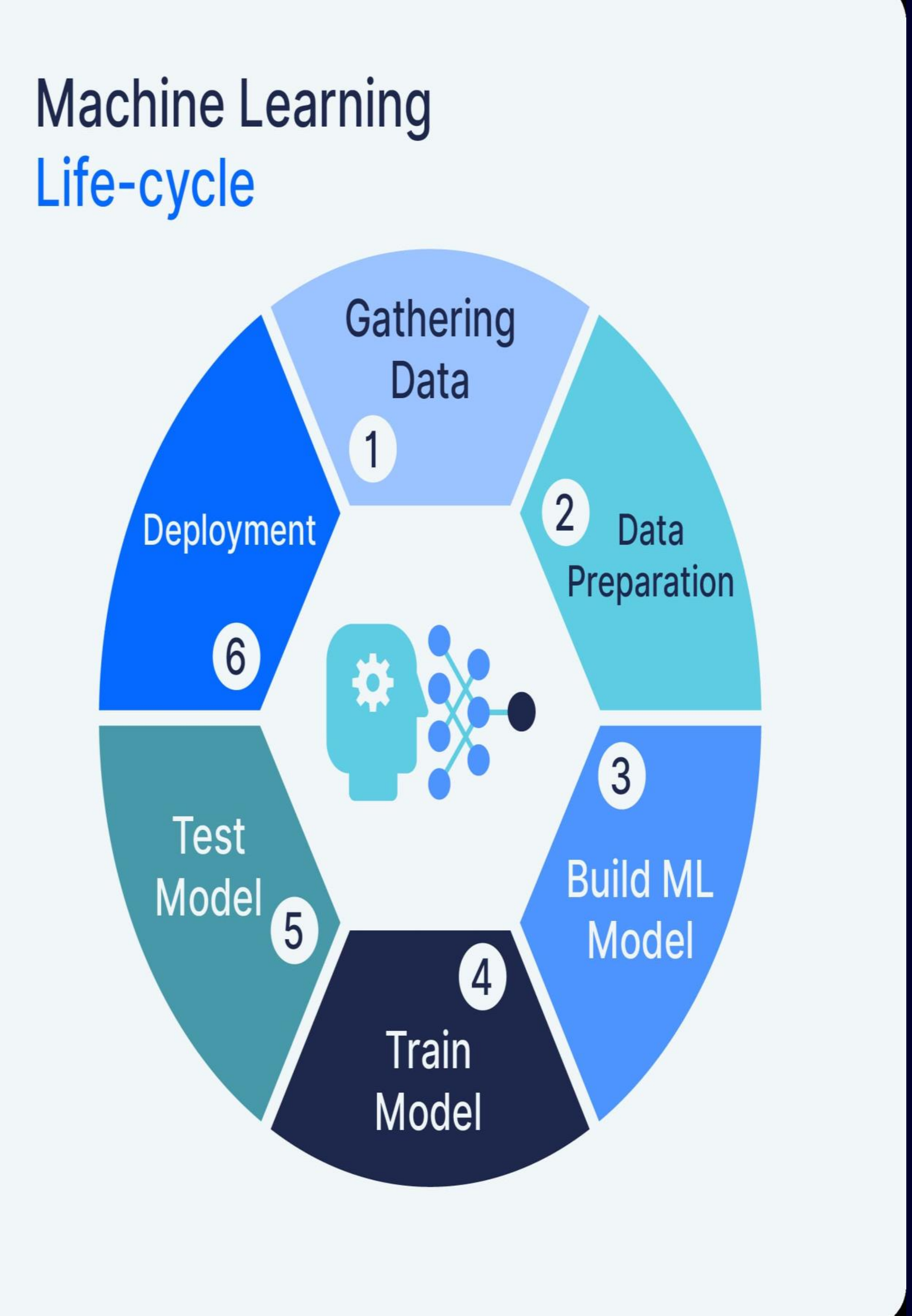
The trained model is evaluated on unseen data to assess its generalizability and performance.

Train Accuracy

The model achieves an impressive train accuracy of 99.99%.

Test Accuracy

The model demonstrates a test accuracy of 92%, indicating its ability to generalize well to unseen data.



Deployment

The trained model is deployed to make it accessible for real-time predictions.

1

Function Definition

A function, predict_sentiment, is defined to preprocess input text and make predictions using the trained model.

2

Gradio Interface Creation

A user-friendly Gradio interface is created, providing text input for movie reviews and displaying the predicted sentiment as output.

3

Launching the Interface

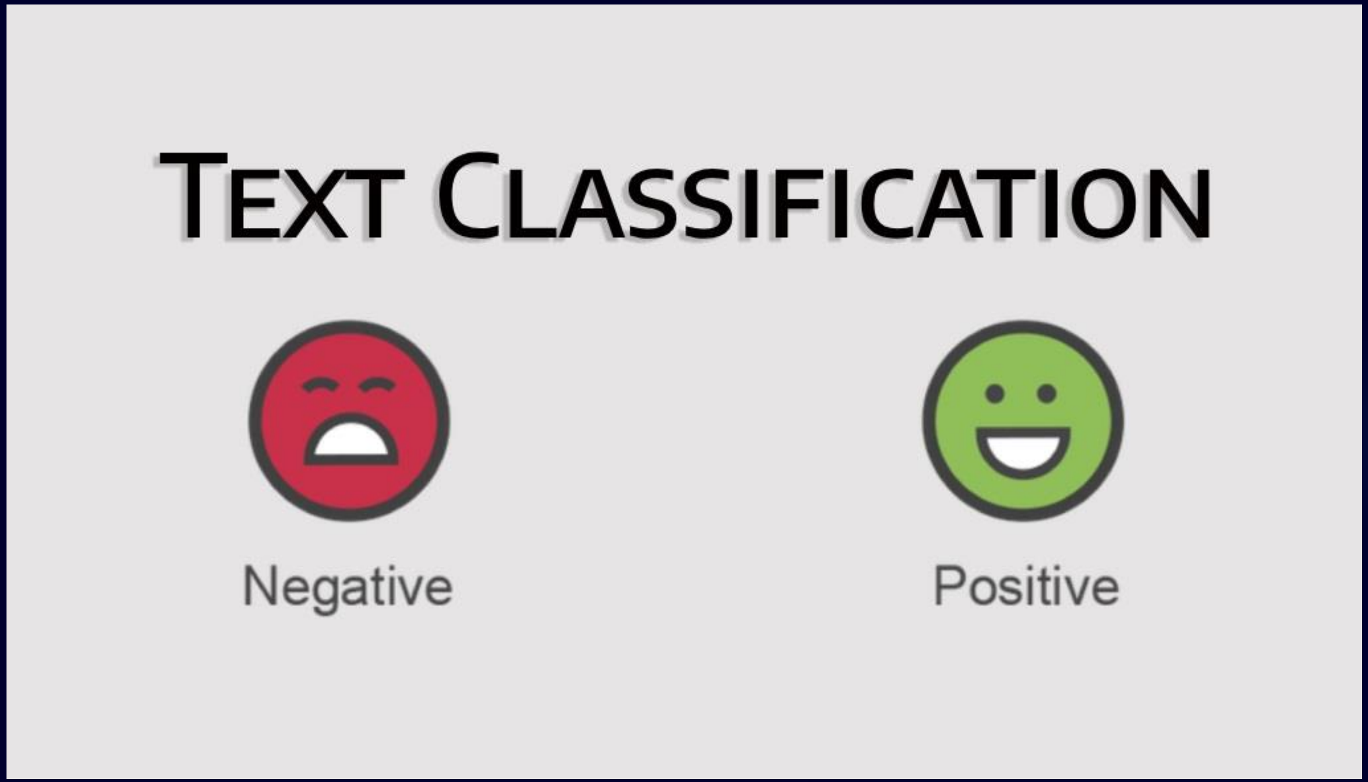
The Gradio interface is launched, enabling users to interact with the deployed model and receive sentiment predictions in real time.

Example 1

Example 2

Conclusion

The text classification project successfully demonstrates the implementation of a sentiment analysis model using the IMDb dataset. By following a structured approach—from data preprocessing and model training to evaluation and deployment—the project highlights the capabilities of deep learning in understanding human sentiments through text.



Future Work

1 Model Improvement

Exploring more advanced architectures, such as transformers (e.g., BERT or GPT), could enhance the model's performance and ability to understand complex linguistic nuances.

2 Data Augmentation

Incorporating additional datasets or expanding the current dataset to include more diverse reviews may improve the model's robustness and generalizability.

3 Hyperparameter Tuning

Experimenting with different hyperparameters (e.g., learning rates, batch sizes) could further optimize the model's performance.