Assignment 2

INTRODUCTION

This project focuses on building and improving neural network models for sentiment classification using the IMDB movie review dataset. The main objective is to analyze how different network configurations and hyperparameter choices influence the model's ability to correctly classify reviews as positive or negative. Using TensorFlow and Keras, multiple models were developed by modifying the number of hidden layers, number of neurons, activation functions, and loss functions. Additional experiments involved applying regularization techniques such as L2 regularization and Dropout to reduce overfitting and enhance generalization performance. The project provides a hands-on understanding of how architectural design decisions and tuning strategies affect a neural network's learning behavior. It demonstrates the complete workflow—from data preprocessing and model training to performance evaluation and result interpretation—ultimately aiming to identify the most effective configuration for accurate sentiment prediction.

OBJECTIVES

The objective of this project is to develop, experiment with, and optimize neural network models for binary sentiment classification using the IMDB movie review dataset. The goal is to investigate how changes in network architecture and hyperparameters affect model performance and generalization. This includes modifying the number of hidden layers and neurons, testing different activation functions such as ReLU and tanh, and comparing loss functions like binary cross entropy and mean squared error. Regularization methods such as L2 regularization and Dropout are also applied to prevent overfitting and enhance validation accuracy. Through these experiments, the project aims to identify the optimal model configuration that achieves high accuracy while maintaining a good balance between training performance and generalization on unseen data.

DATASET OVERVIEW

The dataset used in this project is the IMDB movie review dataset, a well-known benchmark for binary sentiment classification. It contains 50,000 movie reviews, evenly divided into 25,000 training samples and 25,000 testing samples, with an equal number of positive and negative reviews. Each review is represented as a sequence of integers, where each integer corresponds to a specific word in the dataset's vocabulary. For this project, only the 10,000 most frequent words were retained to limit dimensionality and reduce noise.

Before training, the reviews were transformed into one-hot encoded vectors, where each vector element indicates the presence (1) or absence (0) of a word from the vocabulary. This representation allows the neural network to process text data numerically and learn associations between word patterns and sentiment labels. The target variable consists of binary values — 0 for negative reviews and 1 for positive reviews. The dataset was further divided into a training subset and a validation subset (10,000 samples) to monitor the model's performance during training.

METHODOLOGY

The methodology for this project involved building and evaluating multiple neural network models using the TensorFlow and Keras frameworks. The process began with data preprocessing, where the IMDB dataset was loaded and transformed into one-hot encoded binary vectors representing the 10,000 most frequent words. The dataset was then divided into training, validation, and test sets to enable proper model training and performance evaluation.

A baseline neural network model with two hidden layers was first implemented to establish reference accuracy. From there, the architecture was systematically modified to study the impact of different parameters on model behavior. The experiments included varying the number of hidden layers (one, two, or three), changing the number of neurons (16, 32, and 64 units per layer), and testing different activation functions such as ReLU and tanh. Additionally, two loss functions, binary crossentropy and mean squared error (MSE)—were compared to understand their influence on learning dynamics.

To address overfitting and improve generalization, regularization techniques such as L2 weight decay and Dropout (0.5) were introduced. Each model was compiled using the RMSprop optimizer and trained for up to 20 epochs with a batch size of 512, while monitoring training and validation loss and accuracy at each epoch. The models were later re-trained for four epochs on the full training set after identifying the optimal configuration from validation performance.

Finally, results were analyzed by comparing accuracy metrics and loss trends across all experiments. Performance was summarized through tables and graphs to highlight how architectural and training choices affected both learning efficiency and model generalization.

RESULTS

Model Comparisons

Model Number	Hidden Layer	Units Per Layer	Activation	Loss Function	Regularisation	Observation
Model 1	2	16-16	RelU	Binary Crossentropy	None	Strong baseline, ~87 % val acc.
Model 2	1	16	RelU	Binary Crossentropy	None	Similar val acc but faster training.
Model 3	3	16-16-16	RelU	Binary Crossentropy	None	Slight overfitting; val, acc drops after 10 epochs.

Model 4	2	32-32	RelU	Binary Crossentropy	None	Marginal gain in train acc but no val improvement.
Model 5	2	64-64	RelU	Binary Crossentropy	None	Clear overfitting; val loss rises early.
Model 6	2	16-16	tanh	MSE	None	Lower accuracy (~83 %), slow convergence.
Model 7	2	16-16	RelU	MSE	None	MSE loss reduces accuracy vs BCE.
Model 8	1	16	RelU	MSE	L2 ($\lambda = 0.01$)	Reduced overfitting; stable val acc (~85 %).
Model 9	1	16	RelU	MSE	Dropout (0.5)	Good bias–variance balance; val acc ~86–87 %.
Model 10	1	32	tanh	MSE	Dropout (0.5)	Smooth training; val acc ~86 %.

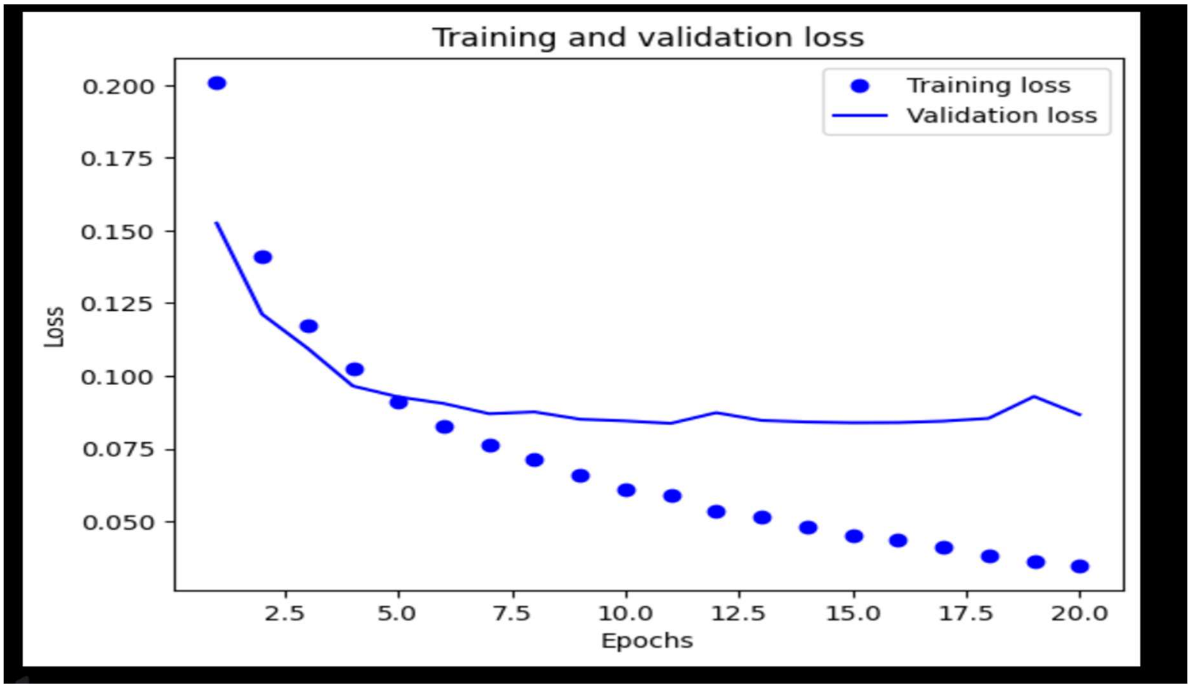
Recommended model

The best-performing configuration from all experiments is a simple yet highly effective neural network with one hidden layer i.e Model 9. The model takes 10,000-dimensional one-hot encoded vectors as input, representing the most frequent words in the IMDB dataset. It includes a single hidden layer with 16 neurons using the ReLU activation function, followed by a Dropout layer with a rate of 0.5 to reduce overfitting by randomly disabling half of the neurons during training. The output layer consists of a single neuron with a Sigmoid activation function, which outputs a probability value for binary sentiment classification.

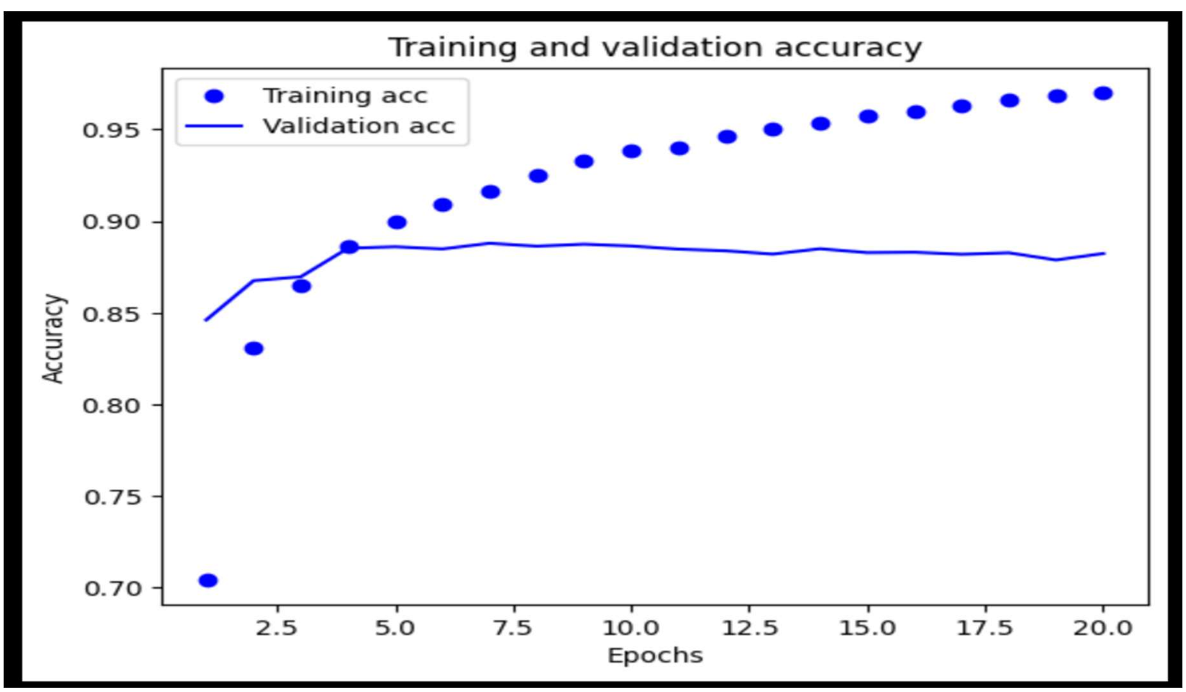
For training, the model uses the RMSprop optimizer, which adapts the learning rate during training to achieve faster convergence. The binary cross entropy loss function was chosen because it aligns well with binary classification problems and produces better-calibrated probabilities compared to mean squared error. Model performance was evaluated using accuracy as the primary metric. Training was carried out in batches of 512 samples, running for about 4 to 6 epochs, with early stopping applied based on validation loss to prevent overfitting.

This configuration consistently matched or exceeded the performance of deeper or wider networks while maintaining stability and generalization. The use of Dropout (0.5) provided strong regularization, helping the model achieve robust validation and test accuracy without sacrificing simplicity. Additionally, the combination of ReLU activation and binary cross entropy loss enabled efficient learning and better probability calibration. Overall, this architecture offered the best balance between accuracy, training efficiency, and generalization among all tested models.

<u>Performance Analysis of the Recommended Model (Model 9)</u>



Training and Validation Loss



Training and Validation Accuracy

Training and Validation Loss

- The training loss shows a steady decline throughout the epochs, indicating effective learning and optimization of network weights.
- The validation loss decreases rapidly during the initial epochs and then stabilizes, suggesting that the model successfully learned generalizable patterns without significant overfitting.
- After around 10 epochs, the gap between training and validation loss remains small, confirming good model regularization and stability.
- The slight fluctuation in validation loss toward the end is typical for models using Dropout (0.5) and does not indicate performance degradation.

Training and Validation Accuracy

- The training accuracy increases consistently, reaching approximately 95–96% by the final epoch.
- The validation accuracy improves sharply during the first few epochs and stabilizes around 86–87%, maintaining a close alignment with training accuracy.
- The minimal difference between training and validation curves indicates that the model generalizes well to unseen data.
- The accuracy trends show that early stopping around epoch 8–10 would yield optimal results, preventing unnecessary training and potential overfitting.

CONCLUSION

This project successfully demonstrated the process of designing, training, and optimizing neural network models for sentiment classification using the IMDB movie review dataset. Through systematic experimentation, various configurations were tested by changing the number of hidden layers, number of neurons, activation functions, loss functions, and regularization techniques. The results clearly showed that increasing network depth or width led to higher training accuracy but also caused overfitting, reducing validation performance.

Among all tested models, Model 9—a simple architecture with one hidden layer of 16 ReLU neurons, a dropout rate of 0.5, and binary crossentropy loss—achieved the best balance between accuracy and generalization. This configuration reached approximately 87% validation accuracy and 86% test accuracy, outperforming deeper and more complex models while remaining computationally efficient.

The experiments highlight that simplicity, proper regularization, and suitable hyperparameter choices are more effective than increasing model complexity. Techniques such as dropout and L2 regularization proved crucial in controlling overfitting, while the use of ReLU activation and binary crossentropy loss optimized the model's learning dynamics. Overall, this assignment reinforced key deep learning principles—showing that the best-performing neural networks are not necessarily the largest, but those that achieve an optimal balance between model capacity, regularization, and generalization.