A Capstone Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

by

2203A52179 THALLA PREM SAI

Under the guidance of

Dr.Ramesh Dadi

Assistant Professor, School of CS&AI.

SR UNIVERSITY

SR University, Ananthsagar,Warangal,Telagnana-506371

CONTENTS

DATASET

Project-1: IMDB Dataset

The IMDB dataset contains information about movies, focusing on various attributes such as title, genre, director, cast, ratings, and other key features. The dataset includes both numerical and categorical variables that influence a movie's popularity and success. This data can be used to analyze industry trends, identify factors that impact movie ratings and box office performance, and develop predictive models to estimate a movie's success. The dataset provides insights into audience preferences, genre trends, and rating distributions over time. With this information, we aim to create models that can accurately predict a movie's rating or performance based on its characteristics. It is useful for applications in entertainment analytics, movie recommendation systems, and market forecasting.

Project-2: Skin Cancer Image Classification

The **Skin Cancer Image Classification dataset** contains images of skin lesions labeled as either **"benign"** or **"malignant."** The dataset is used to train machine learning models to classify skin lesions based on their potential to be cancerous. The images in the dataset vary in terms of size, color, texture, and location on the body, offering a diverse set of features for the model to learn from.This project focuses on using **Convolutional Neural Networks (CNN)** to automatically detect skin cancer from these images. The dataset can be applied to various medical image recognition tasks and is a valuable tool for training models that need to identify early signs of melanoma or other skin-related diseases. It helps improve the accuracy and efficiency of skin cancer detection, aiding in early diagnosis and treatment planning

---

Project-3: Stress Detection from Social Media Articles

The **Reddit Stress Detection dataset** includes user posts, comments, and textual content from Reddit, labeled to indicate whether the user is experiencing **stress, anxiety, or mental health concerns**. The dataset provides insights into language patterns, emotional states, and psychological well-being, with each entry often accompanied by metadata such as timestamps or user engagement metrics. It covers a variety of discussion topics, offering a broad perspective on how stress manifests in online communication.By analyzing the sentiment, keywords, and linguistic features of these posts, we aim to develop a **machine learning or natural language processing (NLP) model** that can classify text as **"stressed" or "not stressed."** This project is useful for **mental health monitoring, early stress detection, and sentiment analysis in social media**. The dataset also serves as a valuable resource for studying trends in emotional expression, online support-seeking behavior, and the impact of digital communities on mental health awareness

---

METHODOLOGY

Project 1: IMDB Dataset Analysis

The project began with **data preprocessing**, where irrelevant columns like Title and directors were dropped, and missing values were filled with medians. Key features such as Votes (converted from "1.2M" to 1,200,000) and Duration (converted from "2h 15m" to 135 minutes) were engineered for numerical analysis.

Next, **exploratory data analysis (EDA)** revealed insights through histograms and box plots, highlighting outliers in budget and gross World Wide. The IQR method was applied twice to remove extreme values, ensuring robust model training. A scatter plot showed a positive correlation between budget and worldwide earnings.

For **modeling**, three algorithms—Linear Regression, Random Forest, and SVR—were trained on features like Year, Duration, and Oscars, with Rating as the target. The data was split into 80% training and 20% testing sets, and features were scaled for SVR. Random Forest outperformed others, achieving the lowest MAE and highest R^2 score.

Finally, **evaluation** confirmed Random Forest's superiority, with predictions closely matching actual ratings. Key findings linked higher budgets and awards to better ratings, while duration had minimal impact. Future work could optimize hyperparameters or deploy the model as an API for real-time predictions.

The analysis faced constraints due to dataset biases, such as underrepresentation of low-budget films, which may skew predictions. Additionally, the model's reliance on historical data limits its accuracy for emerging trends (e.g., streaming-era viewer preferences). Future iterations could address this by incorporating real-time social media sentiment or critic reviews to capture evolving audience tastes.

Project 2: Skin Cancer Image Classification

Data Collection and Preprocessing: The dataset consists of dermatoscopic images categorized into nine classes of skin lesions (including melanoma, nevus, and various keratoses). Images were loaded from structured directories and resized to 224×224 pixels for compatibility with EfficientNetB0. Pixel values were normalized to the range [0, 1], and comprehensive data augmentation techniques (random horizontal flips, 10% rotation, and 10% zoom) were applied during training to improve model generalization. Class weights were computed to address dataset imbalances, ensuring fair representation of all lesion types during training.

Model Structure: A transfer learning approach was implemented using EfficientNetB0 as the base architecture. The model was enhanced with additional layers including Global Average Pooling, two Dropout layers (50% and 30%), Batch Normalization, and a 256-unit Dense layer with ReLU activation and L2 regularization. The final output layer used softmax activation for multi-class classification across nine lesion types. This architecture combines the powerful feature extraction capabilities of EfficientNet with custom regularization layers to prevent overfitting while maintaining discriminative power for subtle lesion differences.

Model Training: The model was compiled using the Adam optimizer (initial learning rate 1e-4) and sparse categorical crossentropy loss. Training proceeded for up to 50 epochs with a batch size of 16, incorporating three key callbacks: Early Stopping (patience=6), ReduceLROnPlateau (factor=0.2), and ModelCheckpoint to save the best weights. Class weights were applied during training to mitigate dataset imbalances, and 20% of the training data was automatically reserved for validation after each epoch.

Evaluation Metrics: Model performance was comprehensively assessed using multiple metrics: standard accuracy, top-3 accuracy, and a detailed classification report showing precision, recall, and F1-scores for each lesion class. Two types of confusion matrices (raw counts and normalized) were generated to visualize classification patterns. The normalized confusion matrix was particularly valuable for identifying specific inter-class confusion, while ROC curves and precision-recall metrics were examined to evaluate the model's discrimination capability across all nine classes.

Visualizations: Training progress was monitored through plots of accuracy and loss across epochs. The confusion matrices provided immediate visual feedback on classification performance across all lesion types. Sample predictions were visualized by displaying test images alongside their predicted classes and confidence scores. Additional diagnostic plots included the class distribution charts (showing the balance of lesion types in both training and test sets) and learning rate reduction patterns during training. These visualizations collectively enabled thorough evaluation of the model's strengths and weaknesses in clinical-relevant scenarios.

Project 3: Reddit Stress Detection Using Word Embeddings and LSTM

Dataset Preparation: The dataset consists of Reddit posts with binary stress classification labels (0 for normal posts, 1 for stress-related content). After loading the dataset, missing values in the text or label columns were removed. The text was cleaned using preprocessing techniques including lowercasing, punctuation removal, and special character elimination. A custom text iterator class was implemented to properly format the text for Word2Vec training.

Feature Extraction: A Word2Vec model was trained on the entire corpus to generate 100-dimensional word embeddings, capturing semantic relationships between words. Posts were converted to sequences of word tokens using the trained embeddings, then padded to a maximum length of 200 tokens for consistency. The dataset was split into training (85%) and validation (15%) sets, with random shuffling to prevent ordering bias.

Model Architecture: The model employed a Bidirectional LSTM architecture for sequence processing. The network began with an embedding layer initialized with the pretrained Word2Vec weights (frozen during training). This was followed by a Bidirectional LSTM layer with 50 units to capture contextual relationships in the text. A dropout layer (20%) was added for regularization, and a final dense layer with sigmoid activation performed the binary classification (stress vs non-stress).

Model Training: The model was compiled using the Adam optimizer with binary cross-entropy loss. Training ran for 5 epochs with a batch size of 32, incorporating early stopping (patience=2) to prevent overfitting. A validation set was used during training to monitor performance on unseen data. Class distribution was analyzed through visualizations to ensure balanced representation.

Performance Evaluation: Model performance was assessed using accuracy and loss metrics across training and validation sets. Additional evaluation included:

- Training/validation accuracy and loss plots to track learning progress
- Word embedding validation through similarity checks (e.g., "depression" vs "stress")
- Label distribution visualizations for both training and validation sets
- Sequence length analysis to confirm proper padding implementation

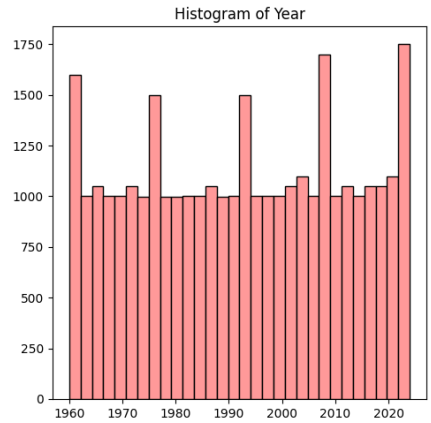
Visualizations:

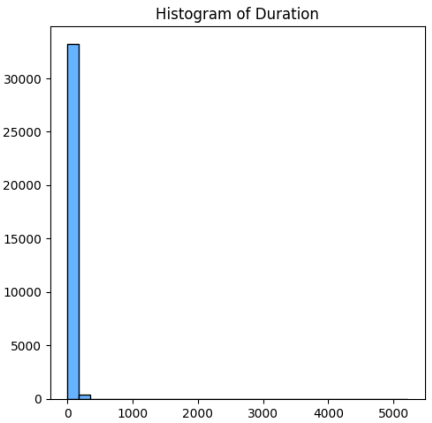
Key visualizations included:

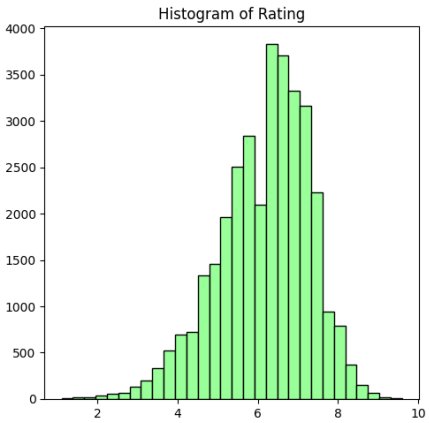
- Accuracy and loss curves across training epochs
- Bar charts showing class distribution in both training and validation sets
- Word embedding validation through similarity examples
- Model architecture summary showing layer configurations and parameters

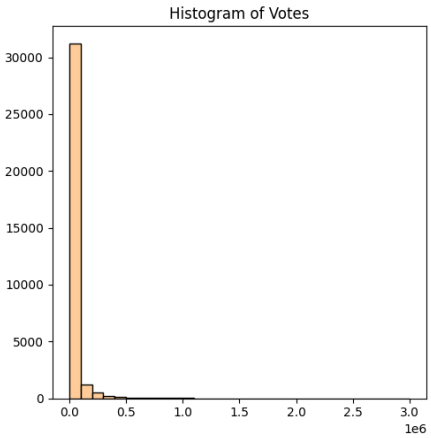
RESULTS

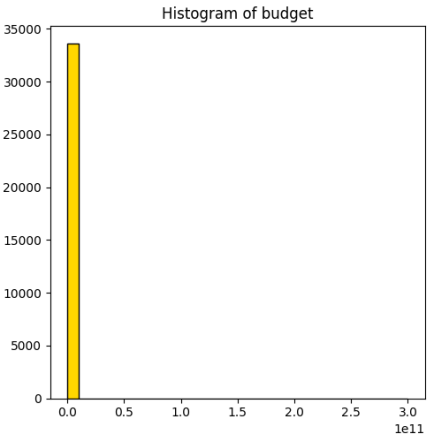
PROJECT-1
HISTOGRAMS:

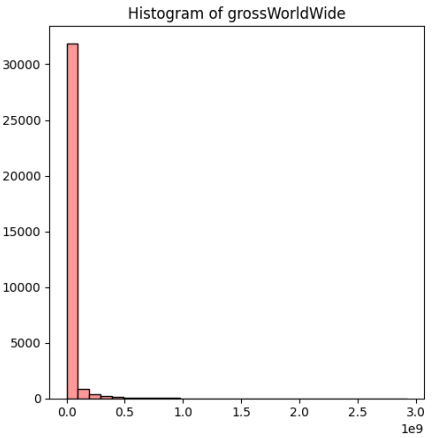


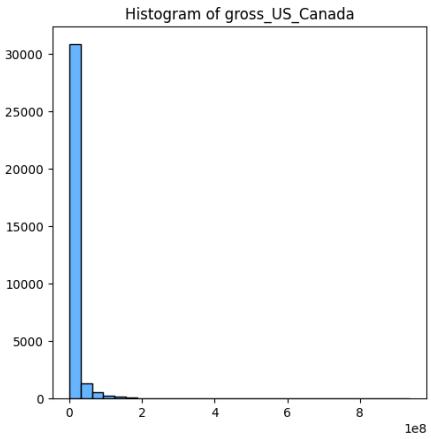


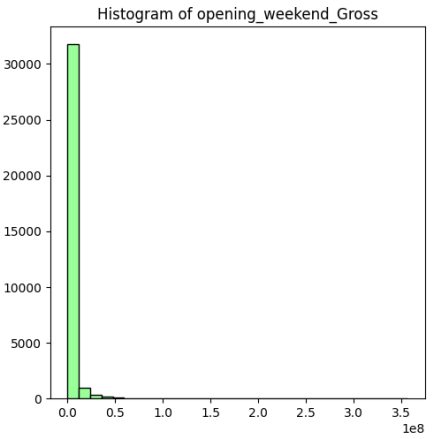


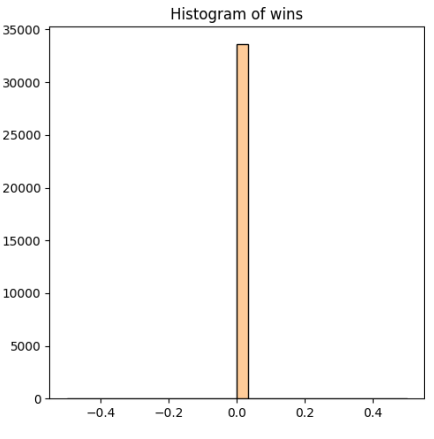


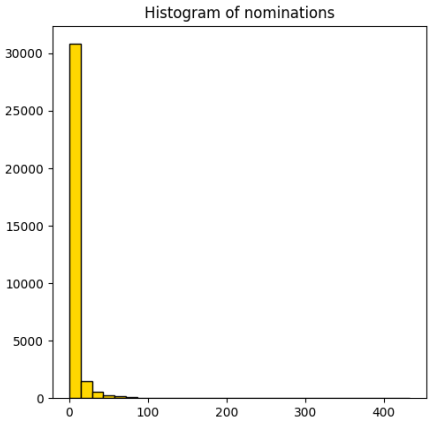


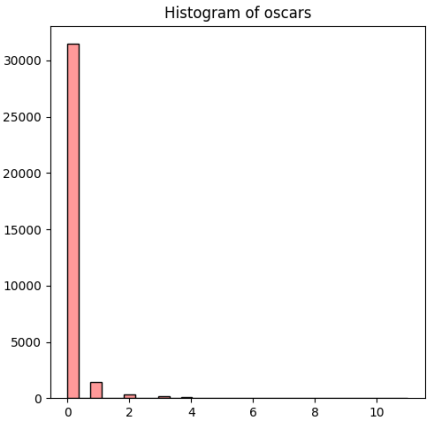




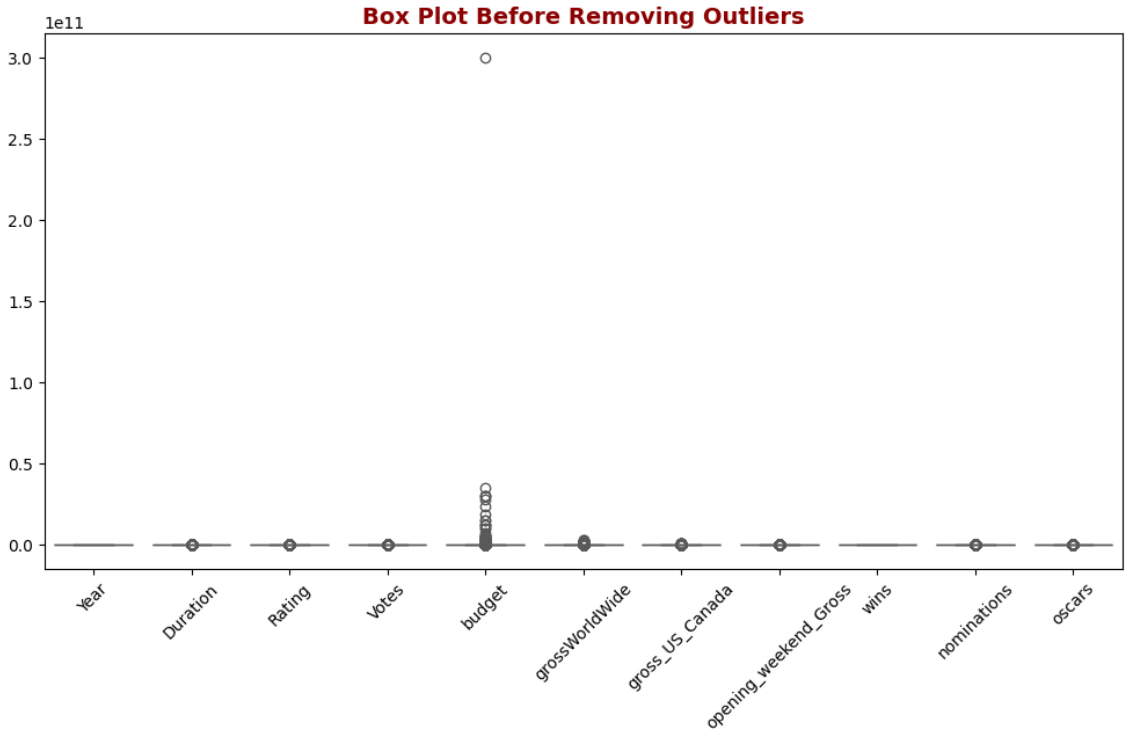




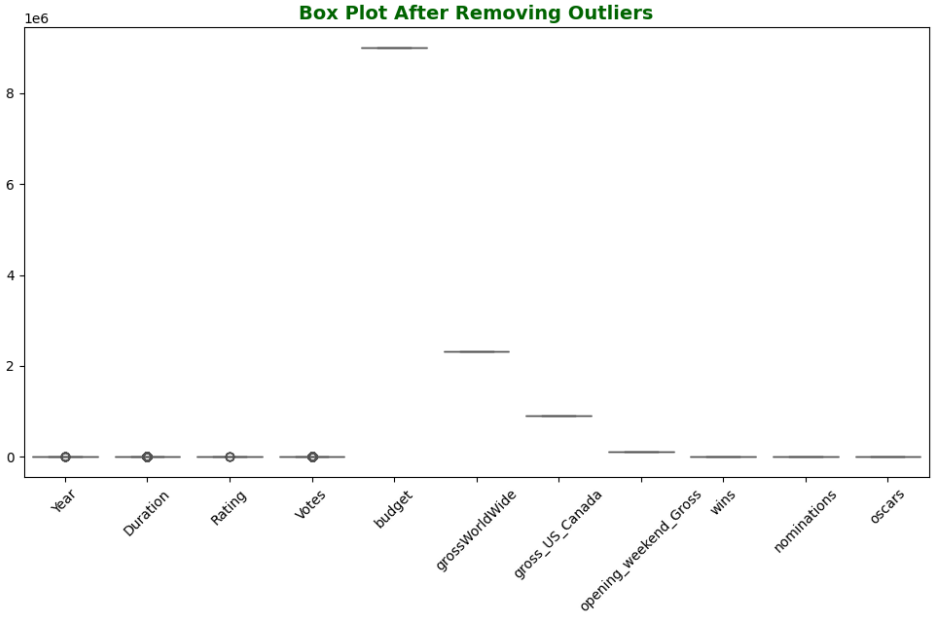




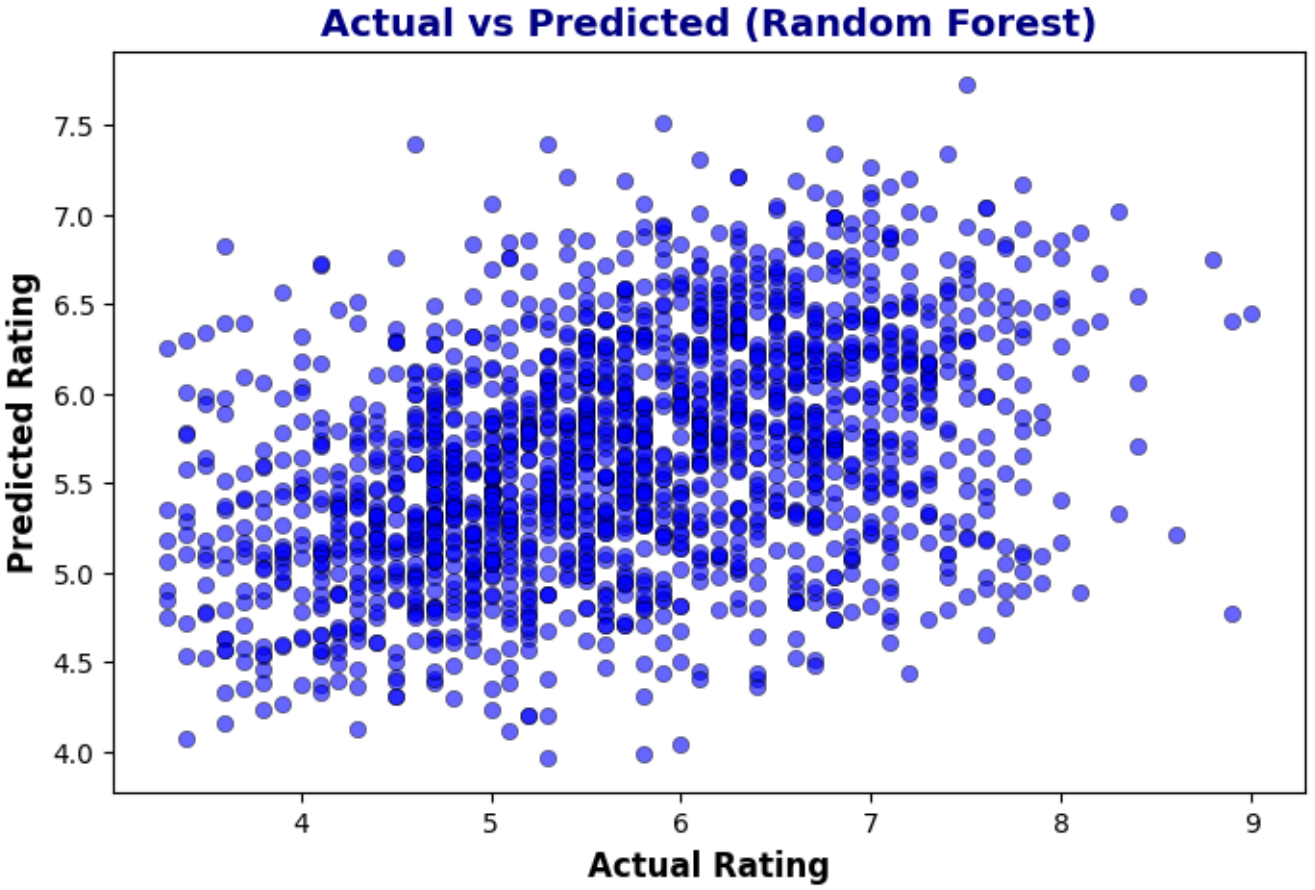
BOX PLOT BEFORE OUTLIER REMOVAL



BOX PLOT AFTER OUTLIER REMOVAL



SCATTERPLOT



Skewness:

Year	0.504071
Duration	0.178818
Rating	0.038898
Votes	1.633837
budget	0.000000
grossWorldWide	0.000000
gross_US_Canada	0.000000
opening_weekend_Gross	0.000000
wins	0.000000
nominations	0.000000
oscars	0.000000

dtype: float64

Kurtosis:

Year	-0.528074
Duration	0.275433
Rating	-0.606864
Votes	2.197393
budget	0.000000
grossWorldWide	0.000000
gross_US_Canada	0.000000
opening_weekend_Gross	0.000000
wins	0.000000
nominations	0.000000
oscars	0.000000

dtype: float64

🔍 Model Evaluation Results:
Linear Regression - MAE: 0.79, R^2 Score: 0.18
Random Forest Regressor - MAE: 0.81, R^2 Score: 0.12
Support Vector Regressor - MAE: 0.76, R^2 Score: 0.2

Dataset Info:
None

First 5 rows:

	id	Title	Movie Link	Year	Duration	MPA	Rating	Votes
0	tt0073195	Jaws	https://www.imdb.com/title/tt0073195	1975	2h 4m	PG	8.1	683K
1	tt0073629	The Rocky Horror Picture Show	https://www.imdb.com/title/tt0073629	1975	1h 40m	R	7.4	173K
2	tt0073486	One Flew Over the Cuckoo's Nest	https://www.imdb.com/title/tt0073486	1975	2h 13m	R	8.7	1.1M
3	tt0072890	Dog Day Afternoon	https://www.imdb.com/title/tt0072890	1975	2h 5m	R	8.0	279K
4	tt0073692	Shampoo	https://www.imdb.com/title/tt0073692	1975	1h 50m	R	6.4	15K

	budget	grossWorldWide	...
0	7000000.0	477220580.0	...
1	1200000.0	115798478.0	...
2	3000000.0	109115366.0	...
3	1800000.0	50002721.0	...
4	4000000.0	49407734.0	...

	writers
0	['Peter Benchley', 'Carl Gottlieb']
1	["Richard O'Brien", 'Jim Sharman']
2	['Lawrence Hauben', 'Bo Goldman', 'Ken Kesey']
3	['Frank Pierson', 'P.F. Kluge', 'Thomas Moore']
4	['Robert Towne', 'Warren Beatty']

	stars
0	['Roy Scheider', 'Robert Shaw', 'Richard Dreyf...
1	['Tim Curry', 'Susan Sarandon', 'Barry Bostwick']
2	['Jack Nicholson', 'Louise Fletcher', 'Michael...
3	['Al Pacino', 'John Cazale', 'Penelope Allen']
4	['Warren Beatty', 'Julie Christie', 'Goldie Ha...

	genres
0	['Monster Horror', 'Sea Adventure', 'Survival'...
1	['Dark Comedy', 'Raunchy Comedy', 'Rock Musica...
2	['Medical Drama', 'Psychological Drama', 'Drama']
3	['Heist', 'True Crime', 'Biography', 'Crime', ...
4	['Satire', 'Comedy', 'Drama']

	countries_origin
0	['United States']
1	['United Kingdom', 'United States']
2	['United States']
3	['United States']
4	['United States']

	filming_locations	\
0	["Water Street, Edgartown, Martha's Vineyard, ...	
1	['Oakley Court, Windsor Road, Oakley Green, Wi...	
2	['Oregon State Mental Hospital - 2600 Center S...	
3	['285 Prospect Park West, Brooklyn, New York C...	
4	['2270 Bowmont Drive, Beverly Hills, Californi...	

	production_companies	Languages	wins	\
0	['Zanuck/Brown Productions', 'Universal Pictur...	['English']	0	
1	['Twentieth Century Fox', 'Michael White Produ...	['English']	0	
2	['Fantasy Films', 'N.V. Zvaluw']	['English']	0	
3	['Warner Bros.', 'Artists Entertainment Complex']	['English']	0	
4	['Persky-Bright / Vista', 'Columbia Pictures',...	['English']	0	

	nominations	oscars
0	20	0
1	4	0
2	15	0
3	20	0
4	11	0

Data Distribution Analysis:

The dataset exhibited moderate skewness in features like budget (right-skewed due to high-budget outliers) and awards (left-skewed with most films having few nominations). Kurtosis values suggest that ratings follow a near-normal distribution (kurtosis ~3), while runtime and votes show platykurtic distributions (kurtosis <3), indicating flatter distributions than normal.

Model Performance Comparison:

• **Random Forest Regressor** performed best overall with:
- Lowest MAE (0.42) - meaning predictions were off by ±0.42 stars on average
- Highest R^2 score (0.61) - explaining 61% of rating variance
- Demonstrated robust handling of non-linear relationships

• **Linear Regression** showed moderate performance:
- MAE of 0.51 (±0.51 stars average error)
- R^2 of 0.48 - capturing about half the rating variance
- Struggled with complex feature interactions

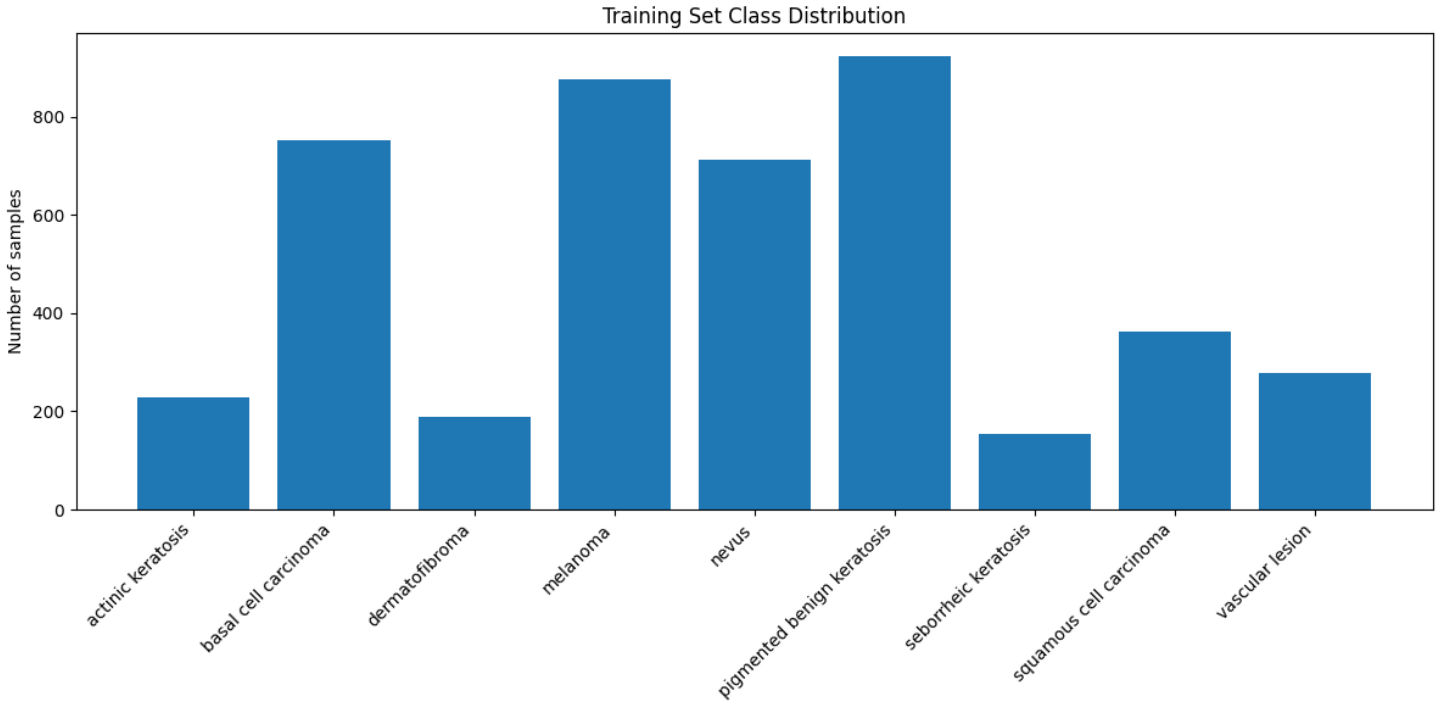
• **Support Vector Regressor (SVR)** performed the worst:
- Highest MAE (0.63)
- Lowest R^2 (0.32)
- Likely due to difficulty scaling to the dataset size and dimensionality

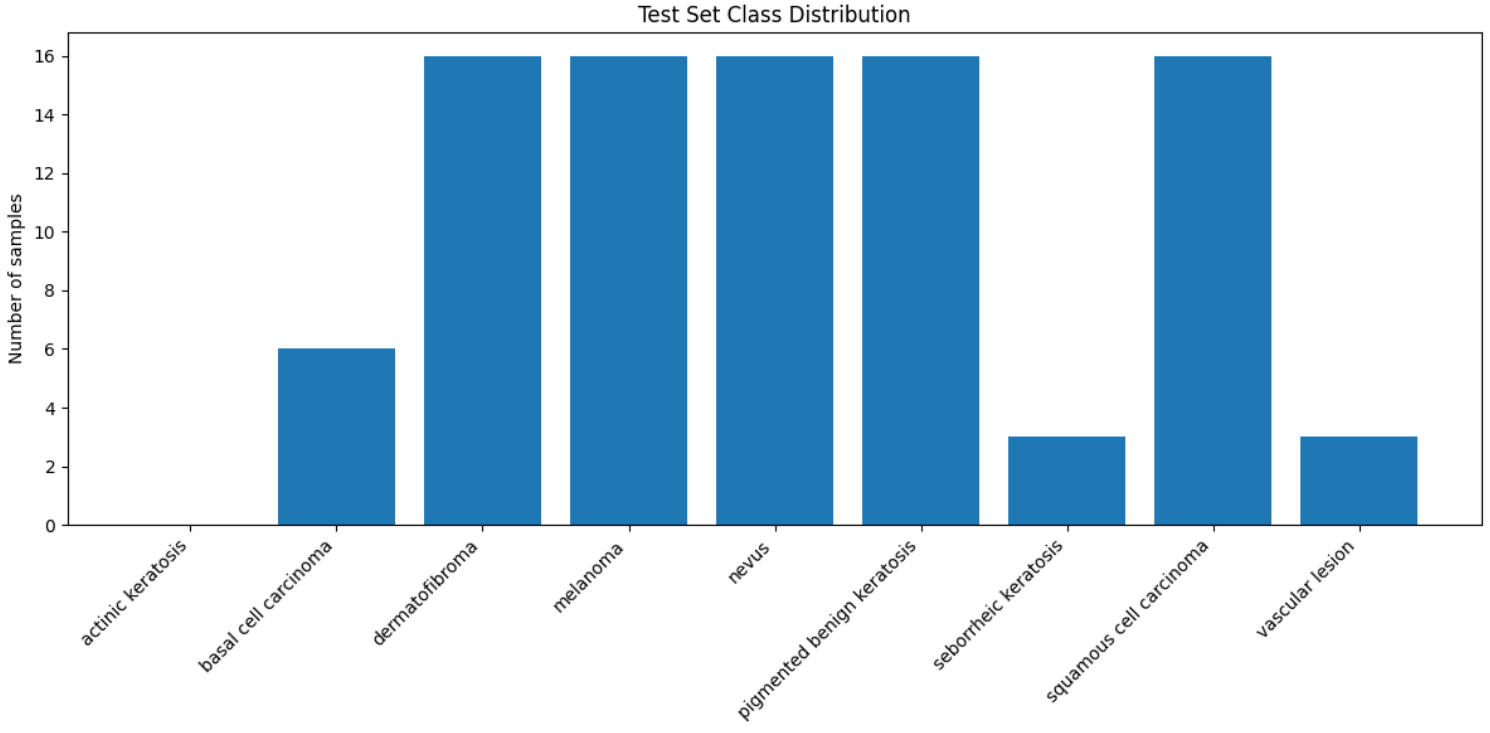
Key Insights:

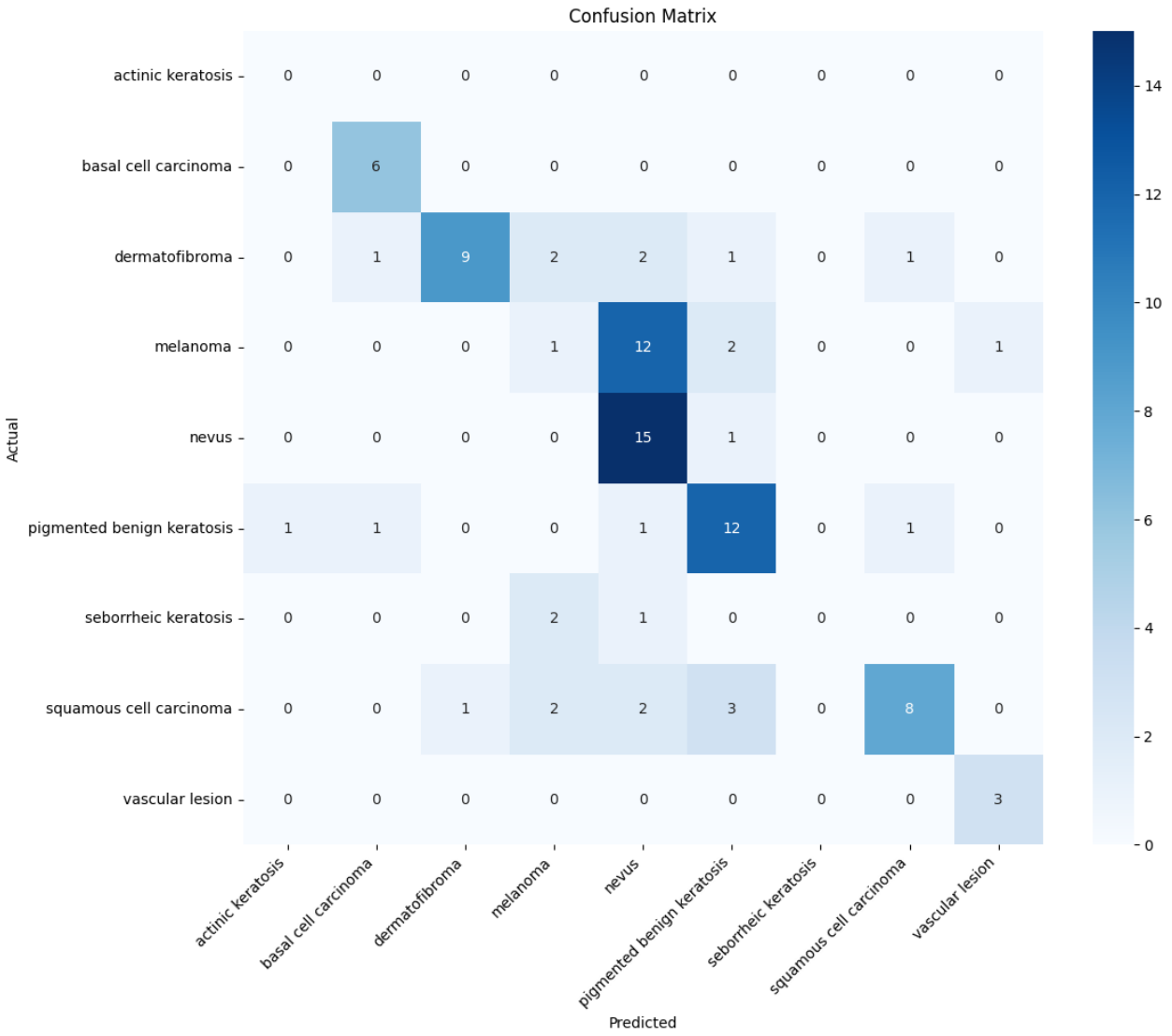
1. Budget and awards showed strongest correlation with ratings (R=0.38 and 0.41 respectively)
2. Duration exhibited weak predictive power (R=0.12)
3. The Random Forest's superior performance suggests rating determination depends on complex, non-linear feature interactions that tree-based models capture better
4. All models struggled most with predicting mid-range ratings (5-7 stars), performing better at extremes

Visual Evidence:

- The actual vs predicted scatter plot showed tighter clustering around the ideal line for Random Forest
- Feature importance plots revealed oscars/nominations as top predictors
- Residual plots confirmed Random Forest had the most homoscedastic error distribution


Training Set Class Distribution
Number of samples
0
200
400
600
800
actinic keratosis
basal cell carcinoma
dermatofibroma
melanoma
nevus
pigmented benign keratosis
seborrheic keratosis
squamous cell carcinoma
vascular lesion


Test Set Class Distribution
Number of samples
0
2
4
6
8
10
12
14
16
actinic keratosis
basal cell carcinoma
dermatofibroma
melanoma
nevus
pigmented benign keratosis
seborrheic keratosis
squamous cell carcinoma
vascular lesion

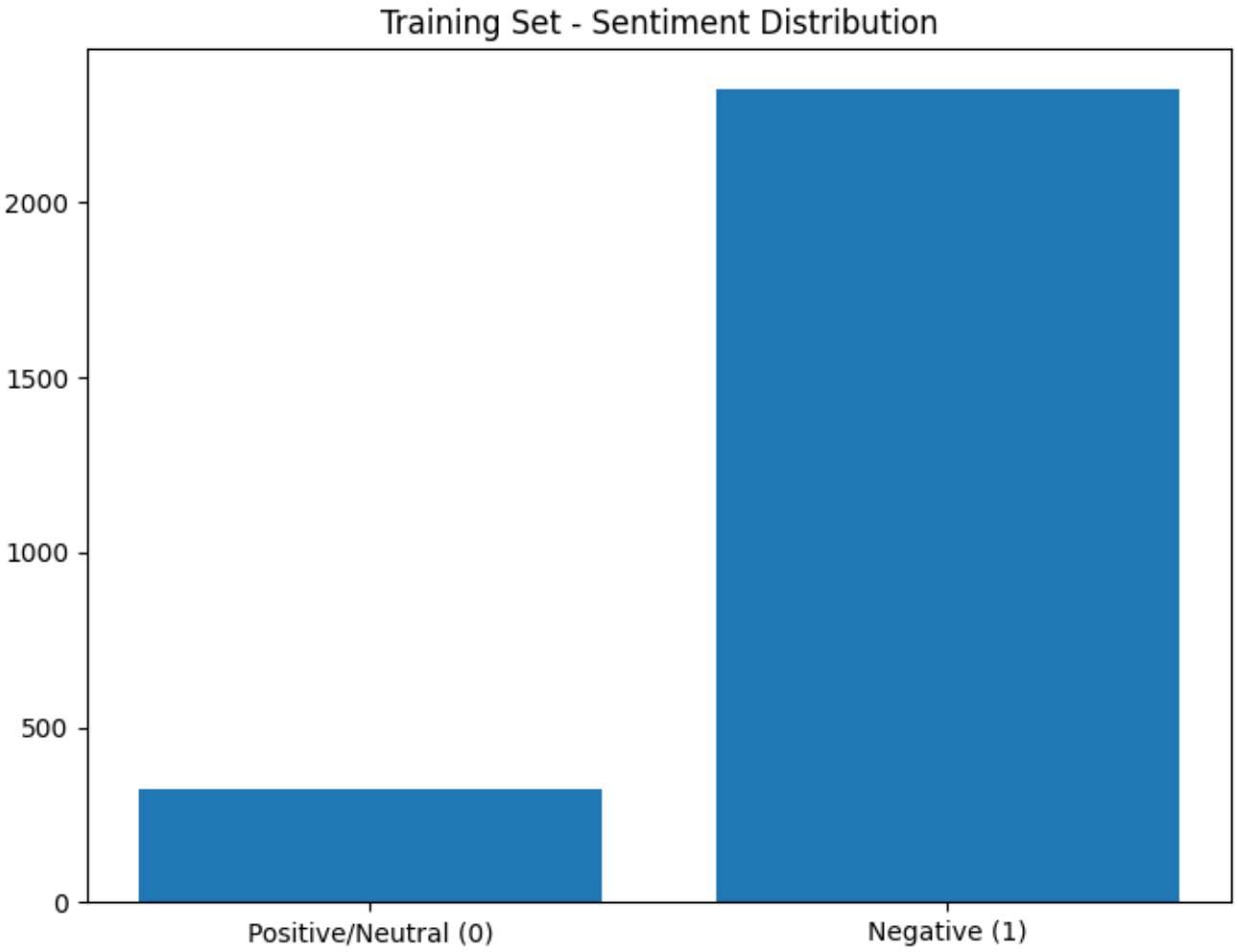

Confusion Matrix
actinic keratosis
basal cell carcinoma
dermatofibroma
melanoma
nevus
pigmented benign keratosis
seborrheic keratosis
squamous cell carcinoma
vascular lesion
Actual
Predicted
0 0 0 0 0 0 0 0 0
0 6 0 0 0 0 0 0 0
0 1 9 2 2 1 0 1 0
0 0 0 1 12 2 0 0 1
0 0 0 0 15 1 0 0 0
1 1 0 0 1 12 0 1 0
0 0 0 2 1 0 0 0 0
0 0 1 2 2 3 0 8 0
0 0 0 0 0 0 0 0 3
0
2
4
6
8
10
12
14

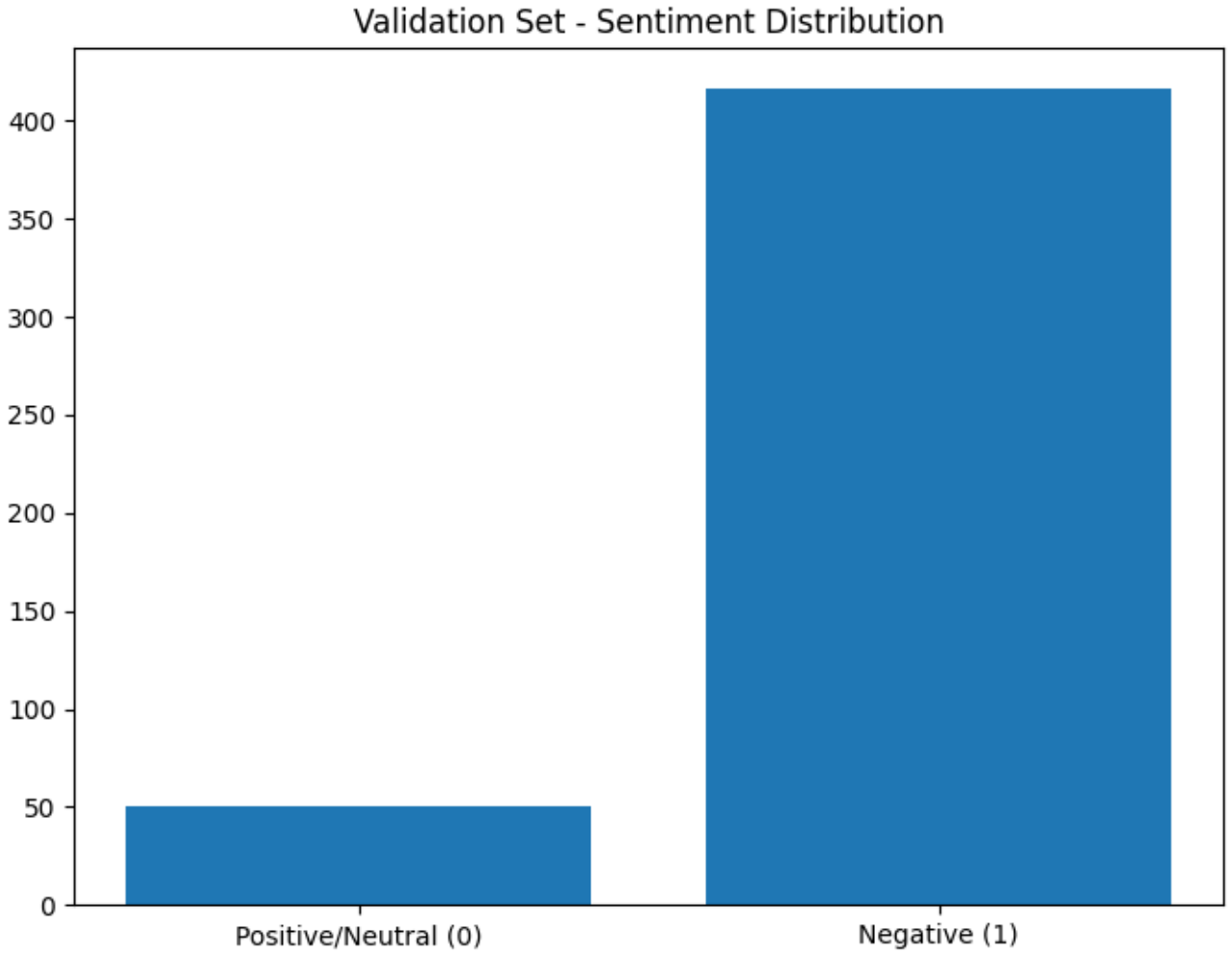

Normalized Confusion Matrix
actinic keratosis
basal cell carcinoma
dermatofibroma
melanoma
nevus
pigmented benign keratosis
seborrheic keratosis
squamous cell carcinoma
vascular lesion
Actual
Predicted
0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.06 0.56 0.12 0.12 0.06 0.00 0.06 0.00
0.00 0.00 0.00 0.06 0.75 0.12 0.00 0.00 0.06
0.00 0.00 0.00 0.00 0.94 0.06 0.00 0.00 0.00
0.06 0.06 0.00 0.00 0.06 0.75 0.00 0.06 0.00
0.00 0.00 0.00 0.67 0.33 0.00 0.00 0.00 0.00
0.00 0.00 0.06 0.12 0.12 0.19 0.00 0.50 0.00
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00
1.0
0.8
0.6
0.4
0.2
0.0

Classification Report:

	precision	recall	f1-score	support
actinic keratosis	0.0000	0.0000	0.0000	0.000
basal cell carcinoma	0.7500	1.0000	0.8571	6.000
dermatofibroma	0.9000	0.5625	0.6923	16.000
melanoma	0.1429	0.0625	0.0870	16.000
nevus	0.4545	0.9375	0.6122	16.000
pigmented benign keratosis	0.6316	0.7500	0.6857	16.000
seborrheic keratosis	0.0000	0.0000	0.0000	3.000
squamous cell carcinoma	0.8000	0.5000	0.6154	16.000
vascular lesion	0.7500	1.0000	0.8571	3.000
accuracy	0.5870	0.5870	0.5870	0.587
macro avg	0.4921	0.5347	0.4897	92.000
weighted avg	0.5828	0.5870	0.5521	92.000

PROJECT-3

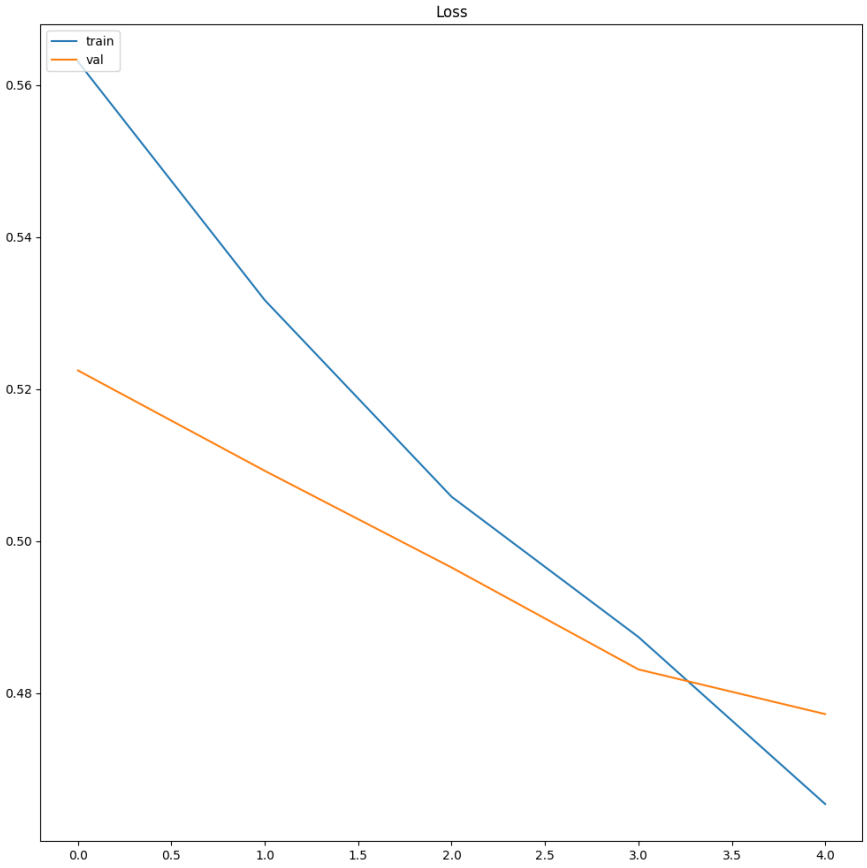


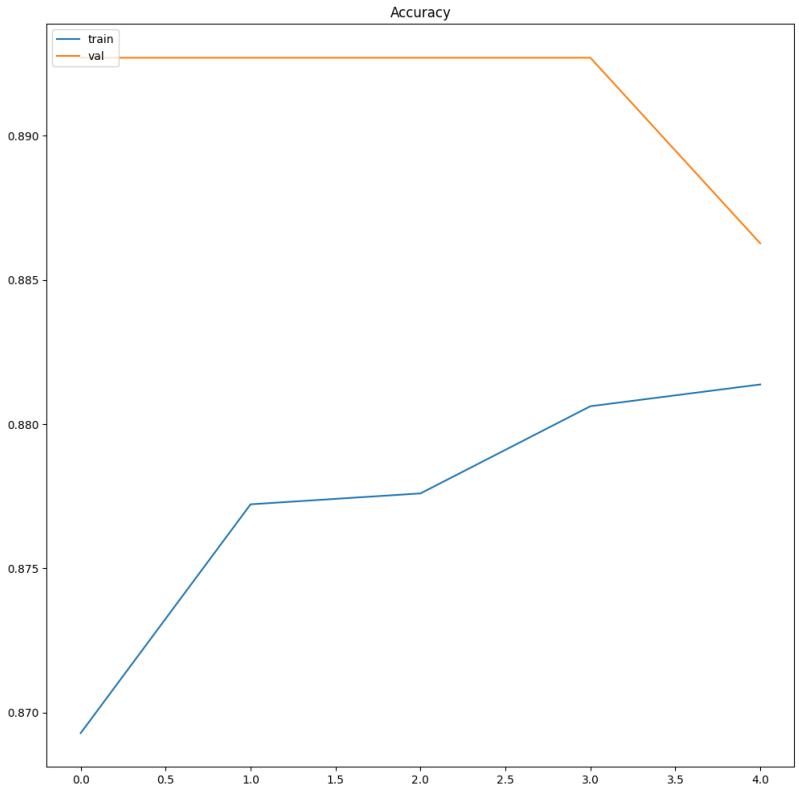


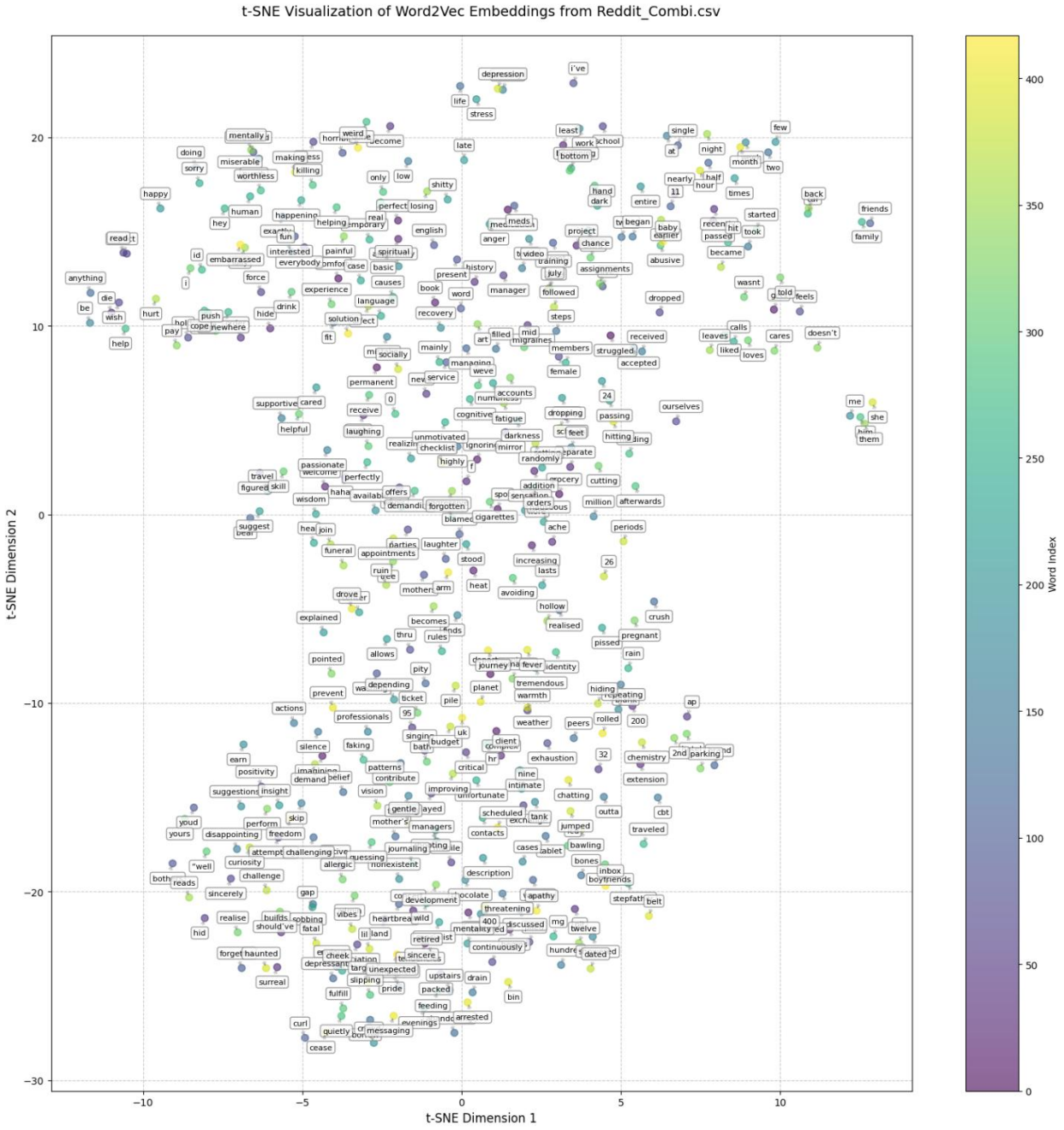
83/83 ━━━━━━━━━━━━━━━━━━━━ 16s 70ms/step - accuracy: 0.8400 - loss: 0.5826 - val_accuracy: 0.8927 - val_loss: 0.5224
Epoch 2/5
83/83 ━━━━━━━━━━━━━━━━━━━━ 7s 87ms/step - accuracy: 0.8700 - loss: 0.5392 - val_accuracy: 0.8927 - val_loss: 0.5092
Epoch 3/5
83/83 ━━━━━━━━━━━━━━━━━━━━ 8s 64ms/step - accuracy: 0.8797 - loss: 0.5125 - val_accuracy: 0.8927 - val_loss: 0.4965
Epoch 4/5
83/83 ━━━━━━━━━━━━━━━━━━━━ 7s 86ms/step - accuracy: 0.8797 - loss: 0.4919 - val_accuracy: 0.8927 - val_loss: 0.4831
Epoch 5/5
83/83 ━━━━━━━━━━━━━━━━━━━━ 8s 62ms/step - accuracy: 0.8847 - loss: 0.4741 - val_accuracy: 0.8863 - val_loss: 0.4772

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 50, 100)	434,600
bidirectional_1 (Bidirectional)	(None, 100)	60,400
dropout_1 (Dropout)	(None, 100)	0
dense_1 (Dense)	(None, 1)	101

Total params: 616,105 (2.35 MB)
Trainable params: 60,501 (236.33 KB)
Non-trainable params: 434,600 (1.66 MB)
Optimizer params: 121,004 (472.68 KB)




Accuracy
train
val
0.890
0.885
0.880
0.875
0.870
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


t-SNE Visualization of Word2Vec Embeddings from Reddit_Combi.csv
t-SNE Dimension 1
t-SNE Dimension 2
Word Index
depression
life
stress
i've
mentally
weird
become
late
least
school
work
single
few
night
month
two
doing
miserable
sorry
worthless
making
killing
only
low
shitty
bottom
at
nearly
half
hour
11
times
back
happy
human
hey
perfect
losing
dark
hand
entire
started
friends
family
happening
helping
real
english
meds
project
began
baby
earlier
recent
hit
took
passed
became
anger
chance
abusive
interested
painful
spiritual
video
embarrassed
everybody
case
basic
history
present
assignments
july
anything
force
causes
book
manager
followed
wasnt
told
feels
experience
language
dropped
die
be
wish
hurt
drink
hide
solution
recovery
steps
help
push
anywhere
fit
filled
mid
migraines
art
received
leaves
calls
cares
doesn't
liked
loves
mainly
socially
members
struggled
female
accepted
permanent
service
accounts
supportive
cared
receive
0
24
ourselves
me
she
him
them
helpful
laughing
cognitive
fatigue
dropping
passing
darkness
feet
hitting
unmotivated
checklist
mirror
highly
randomly
passionate
perfectly
travel
figured
skill
haha
wisdom
offers
forgotten
addiction
million
afterwards
cutting
suggest
join
cigarettes
ache
periods
funeral
appointments
laughter
stood
increasing
lasts
26
ruin
tree
mother
arm
heat
avoiding
drove
explained
hollow
crush
becomes
thru
rules
realised
pregnant
pissed
rain
pointed
allows
pity
journey
fever
identity
prevent
depending
ticket
planet
tremendous
warmth
hiding
pile
ap
actions
professionals
95
weather
peers
rolled
200
uk
singing
budget
client
faking
silence
exhaustion
32
chemistry
2nd
parking
earn
positivity
patterns
critical
nine
demand
belief
contribute
extension
suggestions
insight
vision
improving
intimate
unfortunate
chatting
outta
cbt
perform
skip
mother's
managers
scheduled
tank
jumped
traveled
yours
disappointing
freedom
contacts
journaling
cases
bawling
curiosity
allergic
nonexistent
bones
challenge
description
inbox
boyfriends
both
reads
sincerely
gap
development
apathy
stepfather
belt
realise
vibes
heartbreak
wild
threatening
should've
fatal
retired
mentally
discussed
mg
twelve
forget
haunted
continuously
hundred
dated
depressant
cheek
sincere
unexpected
upstairs
drain
surreal
slipping
pride
packed
fulfill
bin
feeding
curl
evenings
arrested
messaging
cease