Sentiment Analysis on IMDb Movie Reviews

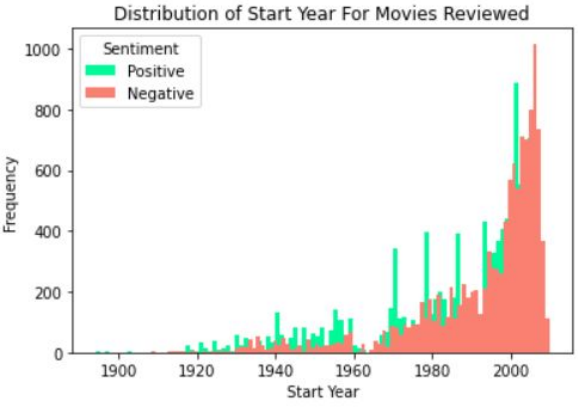
Tim McCormack

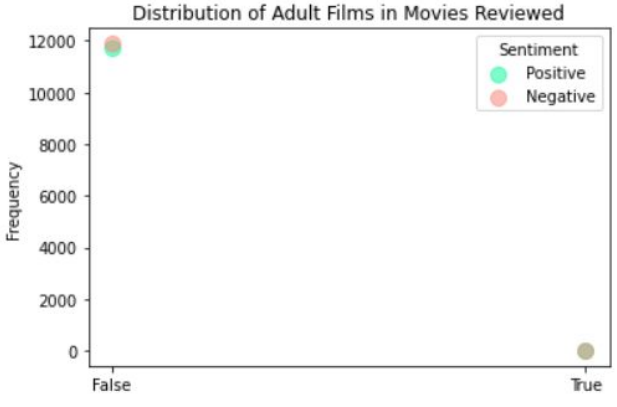
Motivation

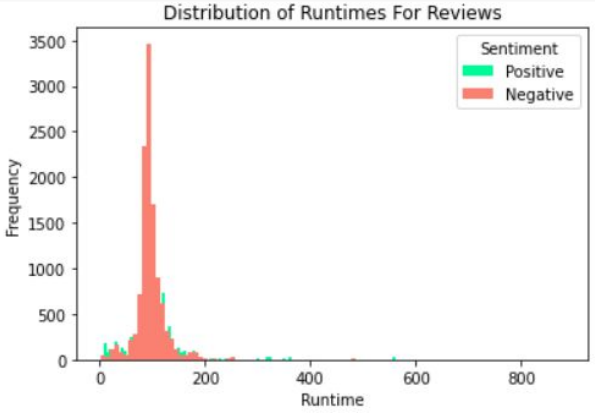
Scenario: A startup company wants to design a subscription based website for movie buffs. The site allows users to host watch parties for their favorite movies. The company would like to work on their recommender system, and they need a way to extract a numerical rating from a written review of a movie. You are a Data Scientist and are writing a proposal to work on this task.

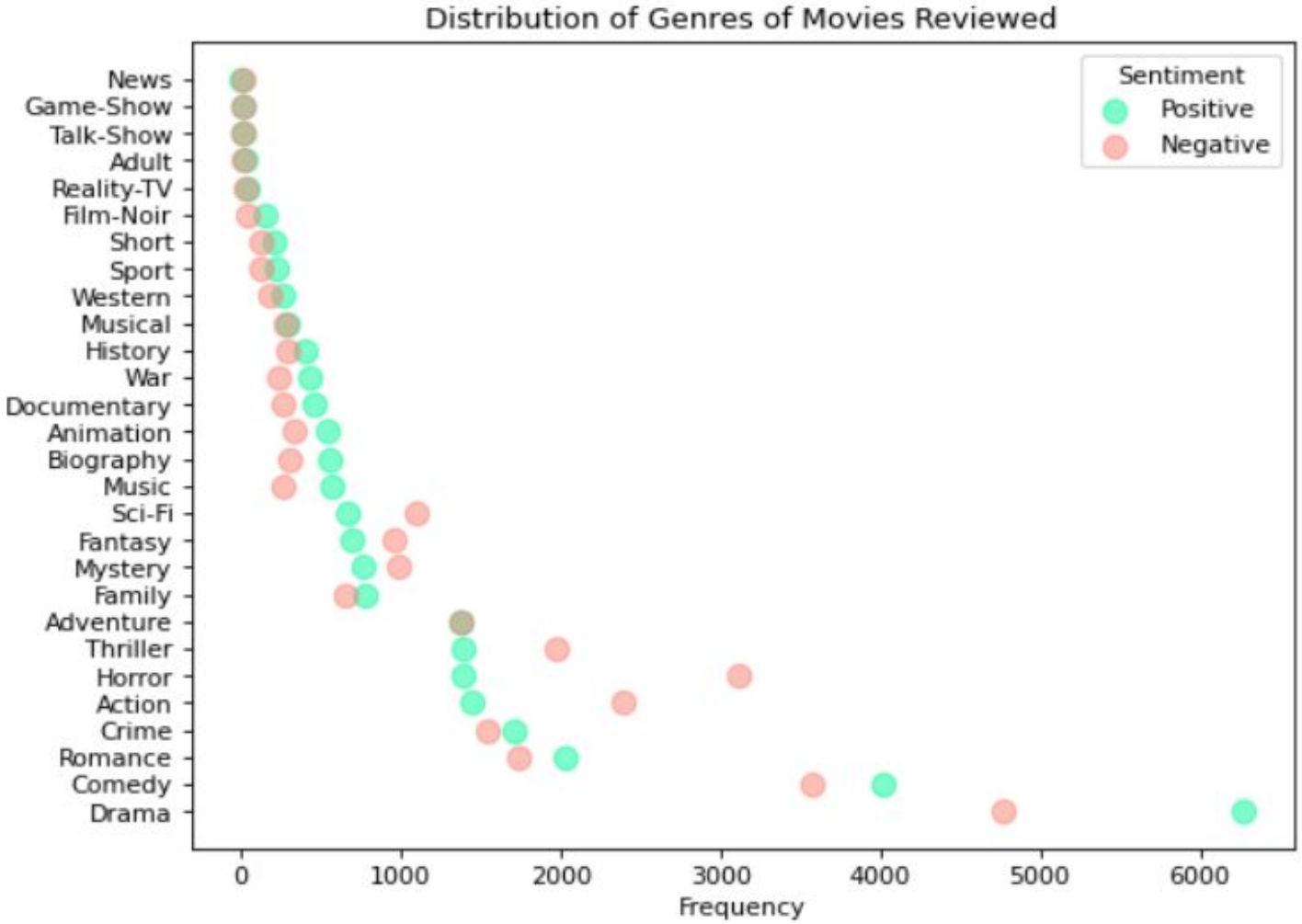
Problem Definition: How Accurately Can You Predict The Numerical Rating From Textual Reviews?

	ReviewID	titleid	titletype	primarytitle	originaltitle	isadult	startyear	endyear	runtimeminutes	genres	Review	Score
0	0	tt0064354	movie	Futz	Futz	False	1969	NaN	92	["'Comedy'"]	Story of a man who has unnatural feelings for ...	3







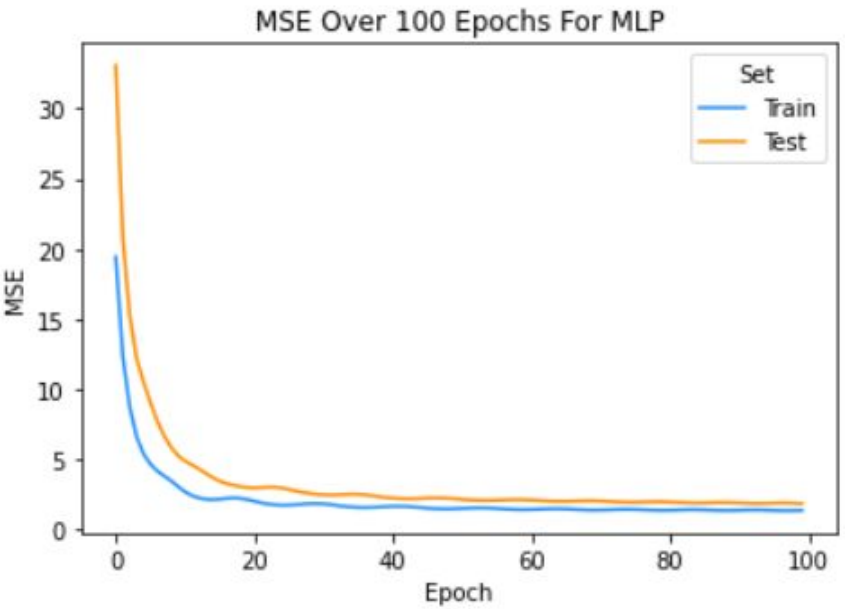

Distribution of Genres of Movies Reviewed
Sentiment
Positive
Negative
News
Game-Show
Talk-Show
Adult
Reality-TV
Film-Noir
Short
Sport
Western
Musical
History
War
Documentary
Animation
Biography
Music
Sci-Fi
Fantasy
Mystery
Family
Adventure
Thriller
Horror
Action
Crime
Romance
Comedy
Drama
0
1000
2000
3000
4000
5000
6000
Frequency

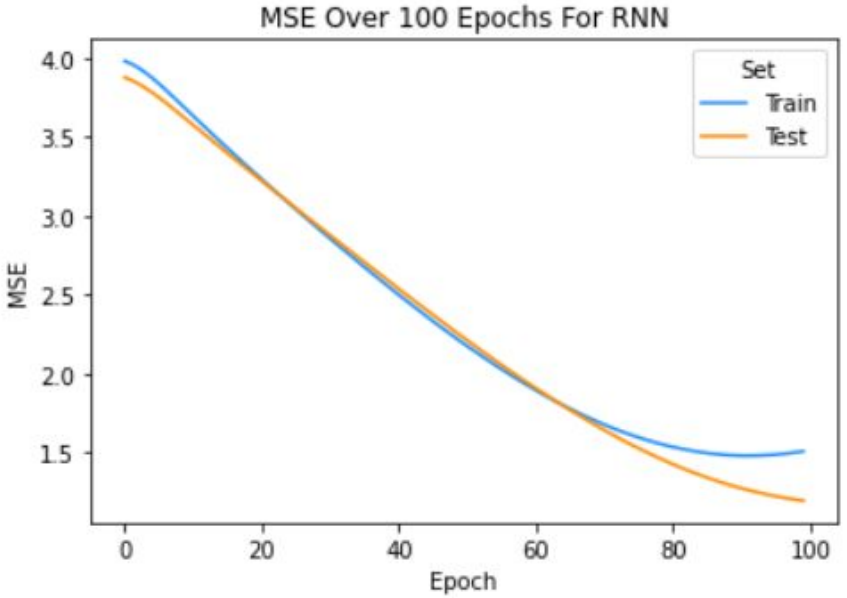
Methodology

1. Multilayer Perceptron
2. RNN (Stacked & Bi-directional)
3. LSTM (1 & 2 Layers)
4. GRU

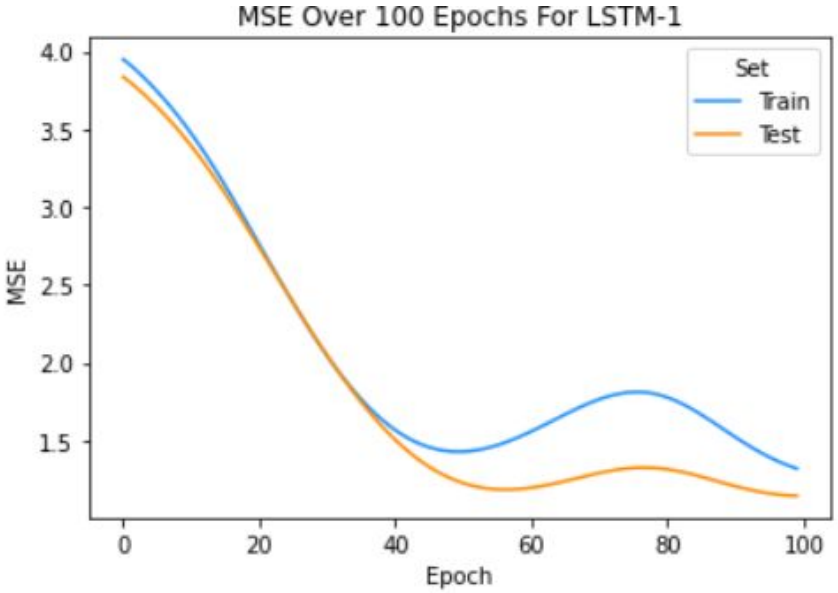
Architecture	Test MSE Loss	Train MSE Loss
MLP	1.89	1.70
RNN	1.15	1.45
LSTM-1	1.18	1.46
LSTM-2	**0.59**	**1.20**
GRU	0.93	1.17

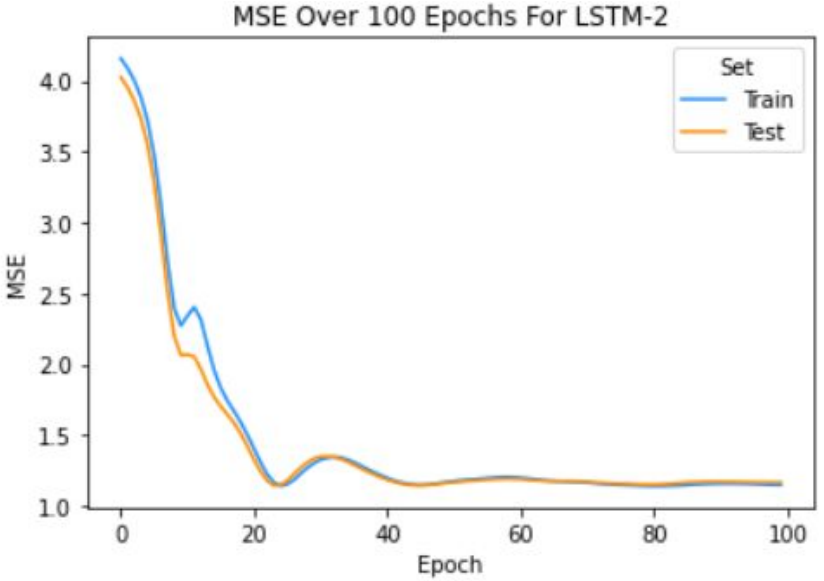
Evaluation - MLP & RNN



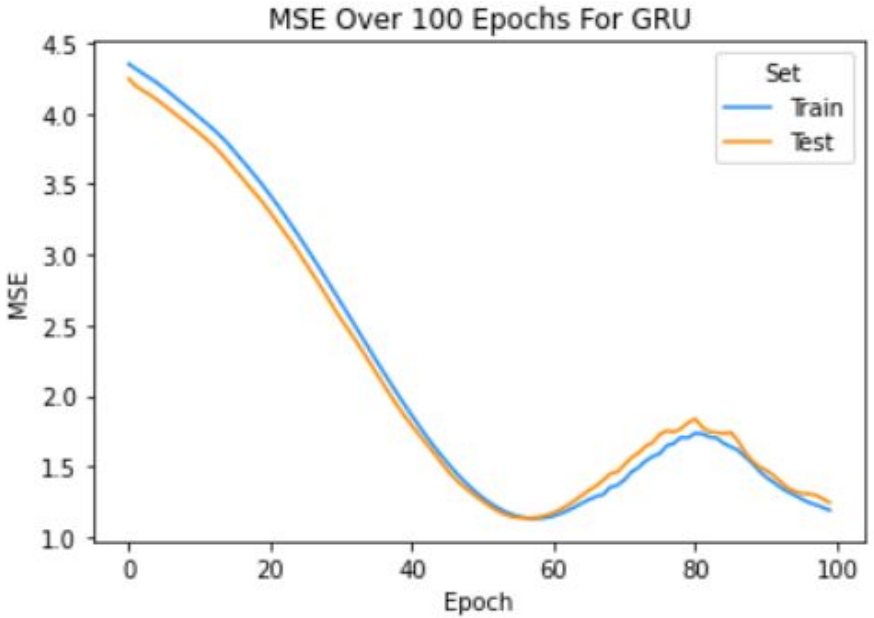


Evaluation - LSTM





Evaluation - GRU



Key Takeaway

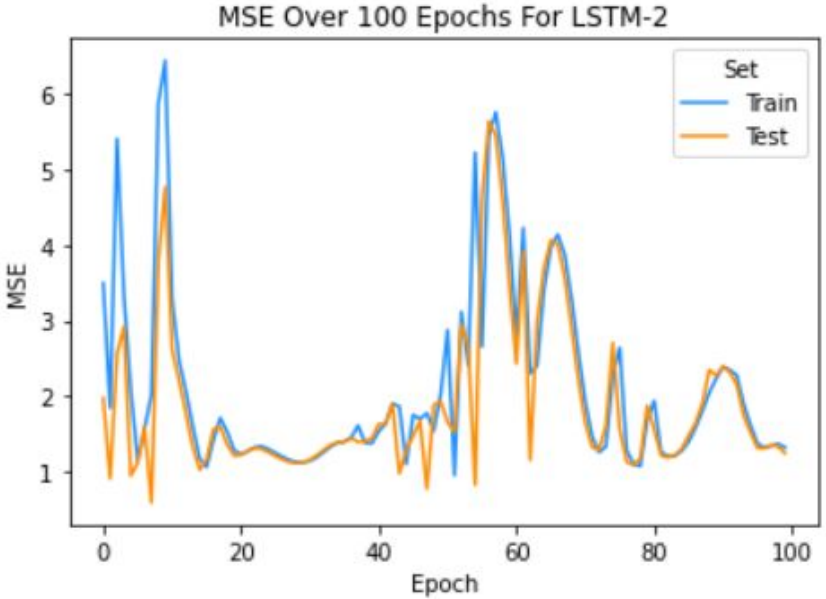
What is unique about my solution?

1. My solution solves a realistic problem given the scarcity of labelled data.
1. I have not found this dataset used in a regression problem.
2. My solution is more accessible than other solutions because my models are created in PyTorch from scratch.

What was surprising?

1. A simple MLP can still be efficacious when extracting a numerical rating from text with a MSE as low as 1.8

Roadblocks


MSE Over 100 Epochs For LSTM-2
Set
Train
Test
MSE
Epoch
1
2
3
4
5
6
0
20
40
60
80
100