# Sentiment Analysis on Movie Reviews

Sabin K. Pradhan

# Intro

- Rotten Tomato Movie Review Dataset
- Amazon's Mechanical Turk to create fine-grained labels
- Sentiment-analysis model benchmark
- Label phrases on a scale of five values: 0) negative, 1) somewhat negative, 2) neutral, 3) somewhat positive, 4) positive

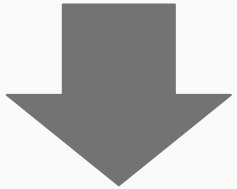
# Data Description

- Dataset: tab-separated files with phrases
- Imported data into pandas dataframe
- Cleaned data using the sklearn.feature_extraction.text module
- Divided train.csv to 80% train and 20% test

# Cleaning Data

Tokenize the Data set

| PhraseID | SentenceID | Phrase | Sentiment |
|---|---|---|---|
| 1 | 1 | Hello World | 2 |



| PhraseID | SentenceID | Hello | World | Sentiment |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 2 |

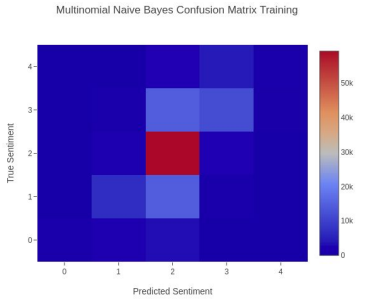
# Models

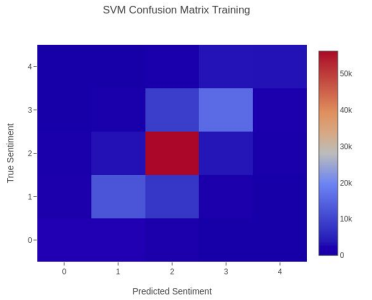
1. Multinomial Naive Bayes
2. Support Vector Machine
3. Neural Networks
4. Decision Tree
5. Random Forest

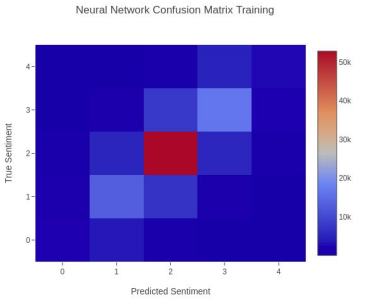
# Model Metrics

| Metrics\ Model | Multinomial Naive Bayes | SVM (Linear Kernel) | Neural Network | Decision Trees | Random Forest |
|---|---|---|---|---|---|
| Training Accuracy | 63.17% | 72.27% | 69.22% | 95.24% | 93.56% |
| Testing Accuracy | 58.34% | 64.16% | 64.67% | 58.46% | 62.42% |

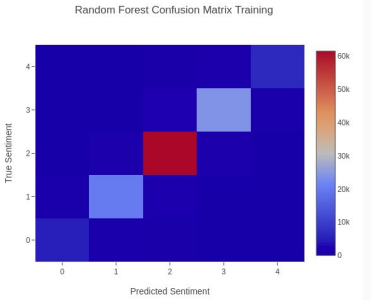
# Training Confusion Matrix



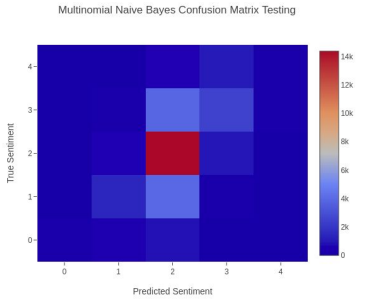



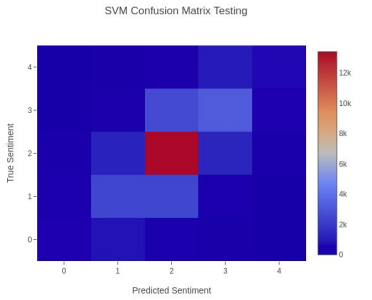



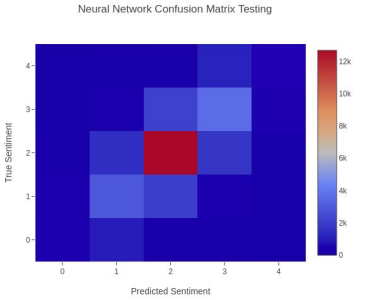


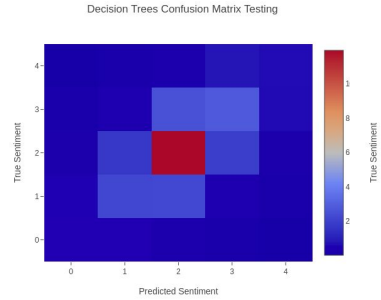


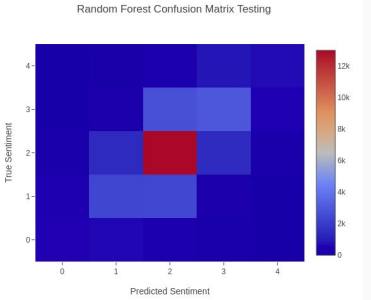

| Multinomial Naive Bayes | SVM Linear Kernel | Neural Network | Decision Trees | Random Forest |
|---|---|---|---|---|




# Testing Confusion Matrix

# Conclusion

Neural Networks performed best with the test data.

“Artificial Intelligence is the new electricity”

- Andrew Ng

# Thanks!

sabinpradhan@outlook.com
github.com/spradh