SENTIMENT ANALYSIS OF MOVIE REVIEWS

GA DSI-SG-26 Capstone

By: Matthew Lio

Agenda

Introduction	• Background • Problem Statement
Data Analysis	• Cleaning • Exploratory Data Analysis
Modelling	• Machine Learning and Lexicon • Deep Learning and RNN
Conclusion	• Model Insights • Limitations & Future Work

Introduction

- Background
- Problem Statement

“Any business is obliged to understand their clients, their needs, opinions, and their satisfaction with the product.”

FEEDBACK IS IMPORTANT

Star ratings

Reviews

Feedback

Users

Positive Feedback

- Customers' satisfaction
- Product promotion

- Increase customer pool
- Increase revenue
- Business growth

Negative Feedback

- Shortcomings
- Suggestions for improvement

- Improvement of products

Opinion Mining & Collection of Data

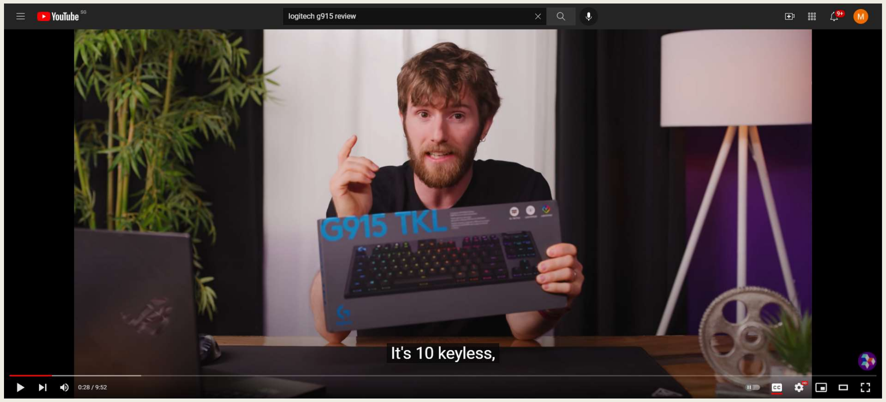
Active Collection of Data

- Incentivising customers to review
- Sponsoring products

Content Creators

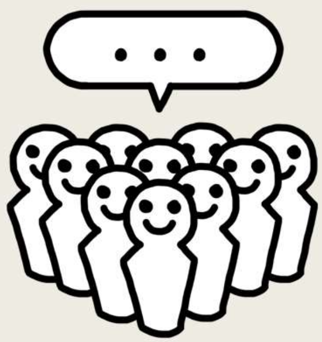
Influencers

YouTubers



The Problem

Thousands of unrated reviews and comments!



Silxnce 1 year ago
I've had the big one before and I loved it but I needed a ten keyless and now they have it! It is an amazing keyboard and feels better than most mechanical keyboard even though they are low profile
1 REPLY

Albert W. 1 year ago
I just got this keyboard over a week ago. Truely an amazing keyboard so far. It is premium looking and incredibly thin. I got the clicky version, it is not loud at all. The click sound provides a great feedback and is pretty satisfying. And it doesn't have issue like a lot of other RGB keyboards – ...
13 REPLY

UNTAPPED DATA

Lackoffaith 1 year ago
The lack of USB C is enough for me to wait for the next iteration
204 REPLY
View 17 replies

Red Robbo's Workshop 1 year ago (edited)
One massive issue with this and the larger version - none of the secondary key legends are illuminated!
I can't believe such an oversight made it into production.
A real shame as the build is superb as are the switches.
So it's ok if you just use it for gaming but if you want to also use it for more general work in subdued lighting, forget it.
1 REPLY

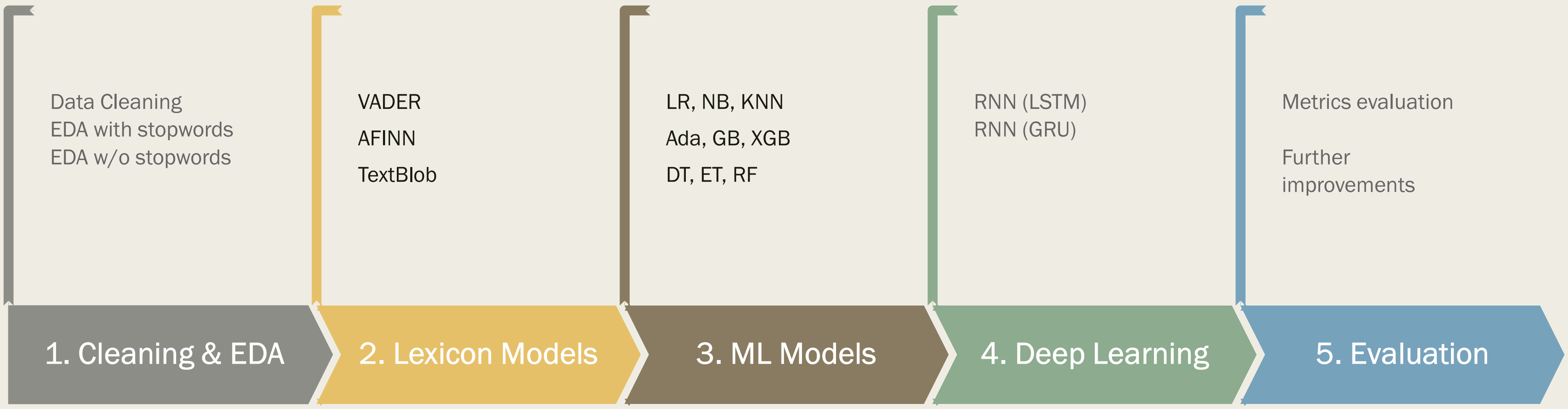
SOLUTION

To create a model that classifies the polarities of sentiments effectively in texts using sentiment analysis

NLP: SENTIMENT ANALYSIS

- A natural language processing (NLP) technique
- The task of classifying polarity of a given text, whether expressed opinion in sentences are positive or negative
- Model would be able to sieve through thousands of unrated reviews/comments and effectively classify them to sentiment polarities

Methodology & Workflow



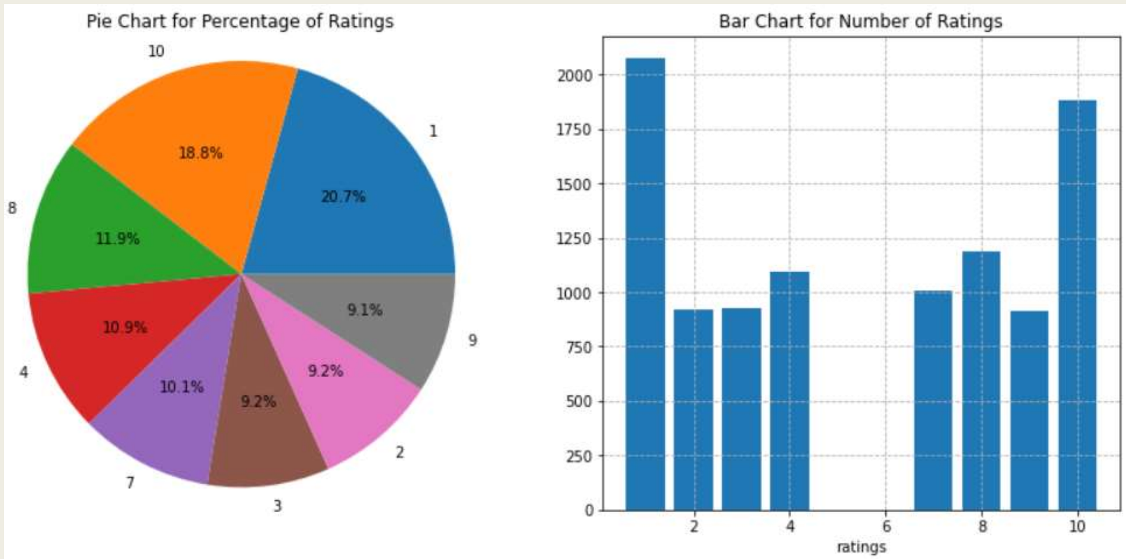
Data Analysis

Dataset: IMDB Movie Reviews

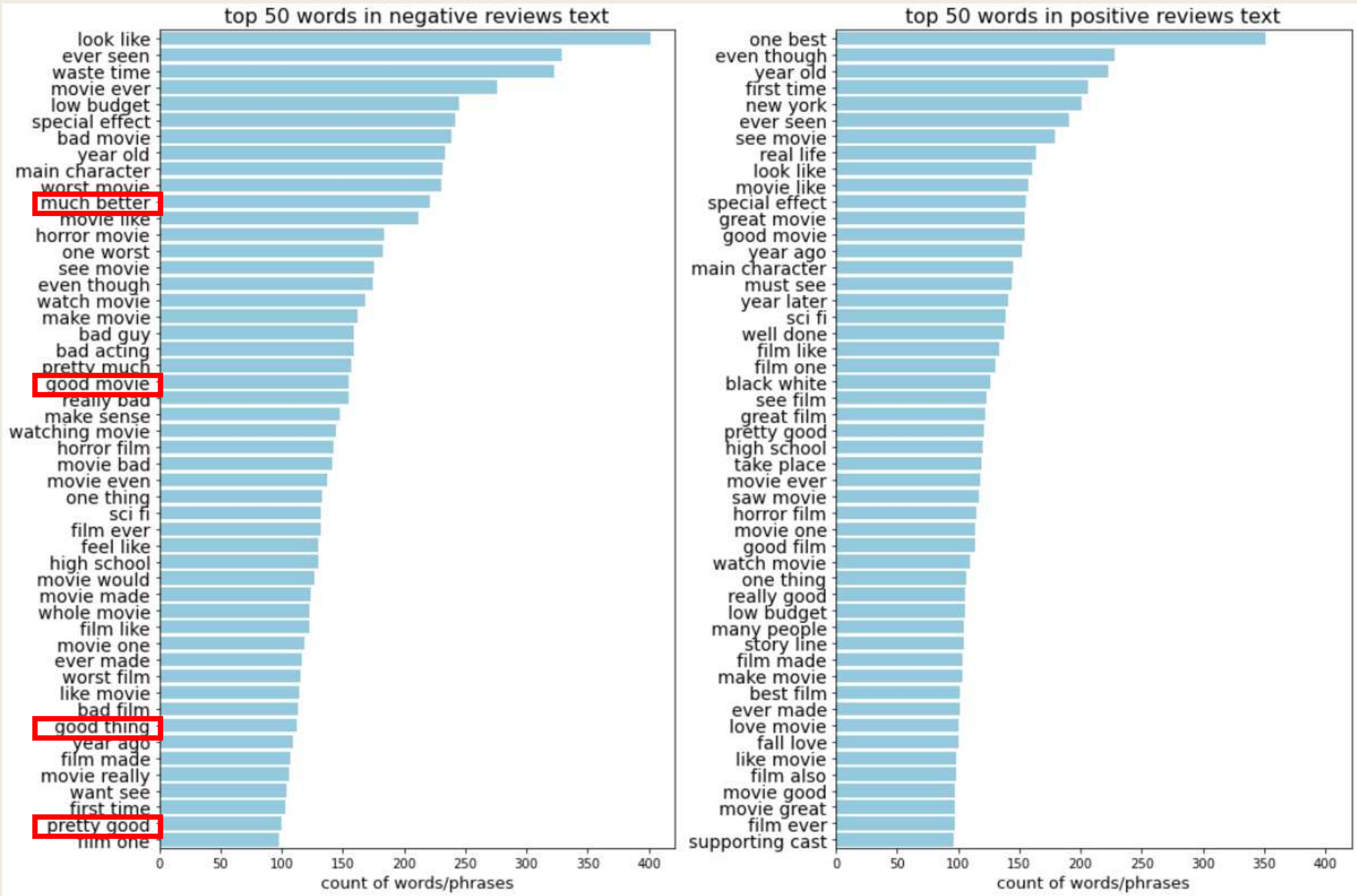
- Data Cleaning
- EDA (With removal of stopwords)
- EDA (Without removal of stopwords)

Data Cleaning

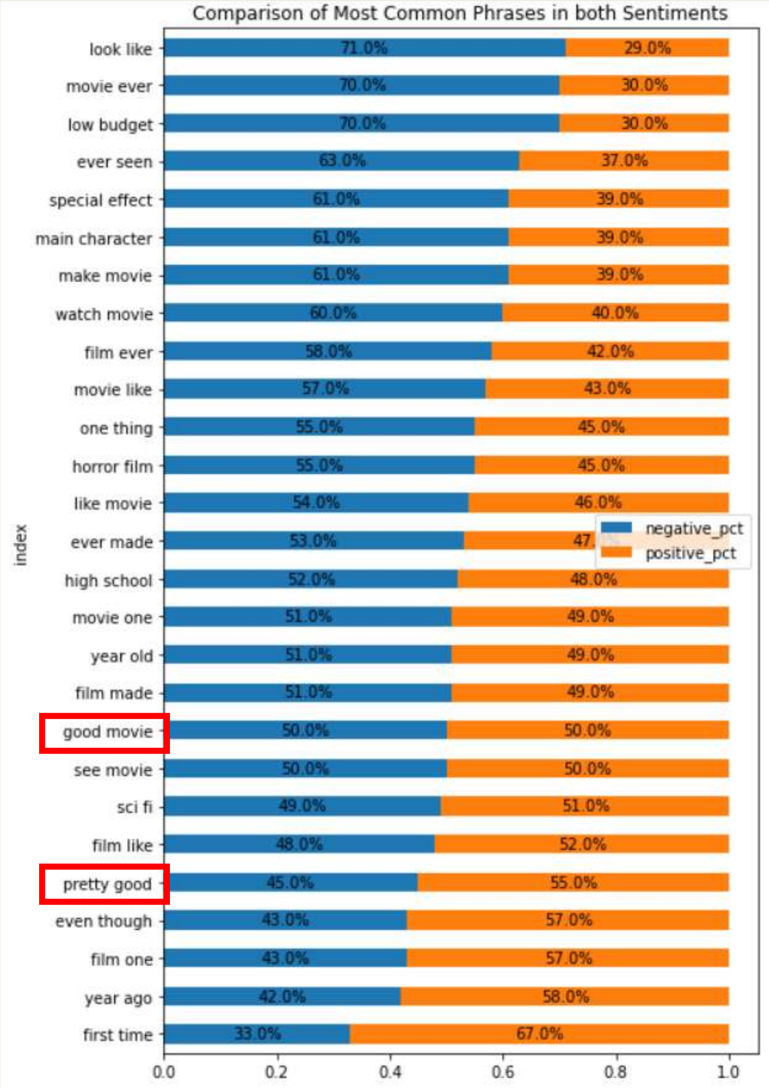
- Removing links
- Removing
- Removing special characters
- Removing duplicate reviews
- Lowercase all letters
- Tokenizing
- Lemmatizing



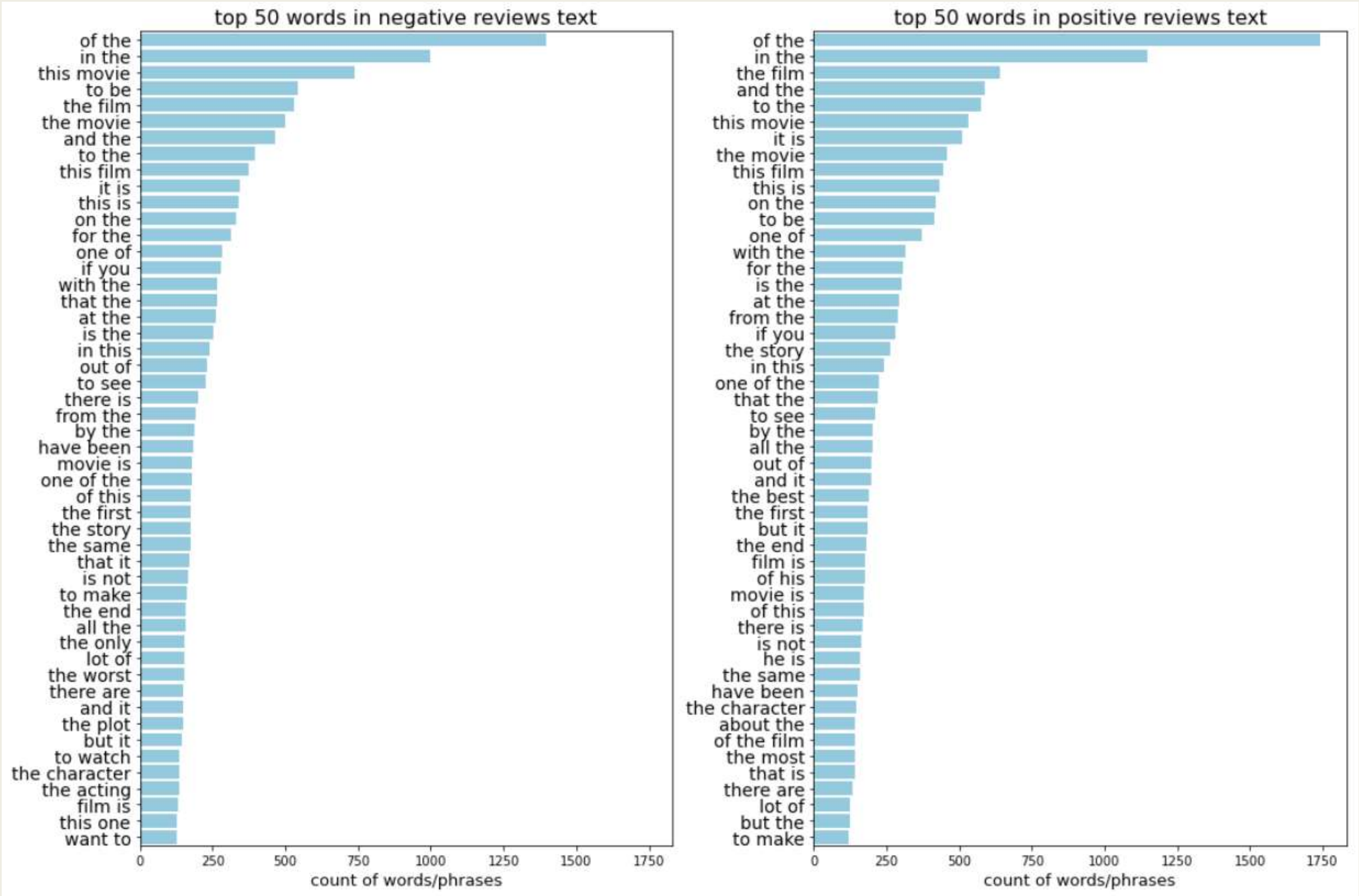
EDA (With removal of stopwords)



Most common phrases in both polarities



EDA (Without removal of stopwords)



Evaluation Metrics

- ROC AUC
- Accuracy
- Specificity

- ROC AUC
  - Better than accuracy.
  - ROC AUC calculated based on predicted scores
- Accuracy
  - Easily interpretable
  - Helpful to explain to non-technical stakeholders
- Specificity
  - Correct predictions of Negatives
  - Negatives may be more important than positive reviews
  - Need to be robust if dataset is imbalanced, with lesser negatives

Lexicon-based Models

- VADER Lexicon
- AFINN Lexicon
- TextBlob Lexicon

- Lexicon means the vocabulary of a person, language or branch of knowledge
- Every word in the dictionary contains a corresponding sentiment score to it
- Combining function makes the final sentimental prediction regarding the total text component

ISSUES

- Meaning of the whole corpus might be different than each individual words used, based on phrases or sentences that imply sarcasm or in a context of comparison

Lexicon Models: Incorrect Predictions

zero day lead you to think even re think why two boy young men would do what they did commit mutual suicide via slaughtering their classmates. it capture what must be beyond a bizarre mode of being for two human who have decided to withdraw from common civility in order to define their own mutual world via coupled destruction. it is not a perfect movie but given what money time the filmmaker and actor had it is a remarkable product. in term of explaining the motif and action of the two young suicide murderer it is better than 'elephant' in term of being a film that get under our 'rationalistic' skin it is a far far better film than almost anything you are likely to see. flawed but honest with a terrible honesty.

Out[8]:

{'neg': 0.157, 'neu': 0.702, 'pos': 0.141, 'compound': -0.3816}

Positive, predicted Negative

there are a lot of highly talented filmmaker actor in germany now. none of them are associated with this movie . why in the world do producer actually invest money in something like this this you could have made good film with the budget of this garbage it's not entertaining to have seven grown men running around a dwarf pretending to be funny. what is funny though is that the film's producer who happens to be the oldest guy of the bunch is playing the youngest dwarf. the film is filled with moment that scream for caption saying you're supposed to laugh now . it's hard to believe that this crap's supposed to be a comedy. many people actually stood up and left the cinema minute into the movie. i should have done the same instead of wasting my time... pain

Out[9]:

{'neg': 0.079, 'neu': 0.768, 'pos': 0.153, 'compound': 0.8907}

Negative, predicted Positive

word can't describe how bad this movie is. i can't explain it by writing only. you have too see it for yourself to get at grip of how horrible a movie really can be. not that i recommend you to do that. there are so many clich s mistake and all other negative thing you can imagine here that will just make you cry. to start with the technical first there are a lot of mistake regarding the airplane. i won't list them here but just mention the coloring of the plane. they didn't even manage to show an airliner in the color of a fictional airline but instead used a painted in the original boeing livery. very bad. the plot is stupid and been done many time before only much better. there are so many ridiculous moment here that i lost count of it really early. also i on the bad guys' side all the time in the movie because the good guy were so stupid. executive decision should without a doubt be you're choice over this one even the turbulence movie are better. in fact every other movie in the world is better than this one.

Out[10]:

{'neg': 0.122, 'neu': 0.744, 'pos': 0.134, 'compound': 0.7007}

Negative, predicted Positive. WHY

Lexicon Models: Evaluation

VADER Lexicon

```
print(classification_report(train['sentiment'], train['v_sentiment']))
confusion_matrix(train['sentiment'], train['v_sentiment'])
```

```
              precision    recall  f1-score   support

           0       0.78      0.54      0.64     12432
           1       0.65      0.85      0.74     12472

    accuracy                           0.69     24904
   macro avg       0.72      0.69      0.69     24904
weighted avg       0.72      0.69      0.69     24904


array([[ 6670,  5762],
       [ 1843, 10629]], dtype=int64)
```

AFINN Lexicon

```
print(classification_report(train['sentiment'], train['afinn_sentiment']))
confusion_matrix(train['sentiment'], train['afinn_sentiment'])
```

```
              precision    recall  f1-score   support

           0       0.79      0.58      0.67     12432
           1       0.67      0.85      0.75     12472

    accuracy                           0.71     24904
   macro avg       0.73      0.71      0.71     24904
weighted avg       0.73      0.71      0.71     24904


array([[ 7198,  5234],
       [ 1916, 10556]], dtype=int64)
```

TextBlob Lexicon

```
print(classification_report(train['sentiment'], train['tb_sentiment']))
confusion_matrix(train['sentiment'], train['tb_sentiment'])
```

```
              precision    recall  f1-score   support

           0       0.89      0.43      0.58     12432
           1       0.62      0.95      0.75     12472

    accuracy                           0.69     24904
   macro avg       0.76      0.69      0.66     24904
weighted avg       0.76      0.69      0.66     24904


array([[ 5289,  7143],
       [  668, 11804]], dtype=int64)
```

Machine Learning Models

Models

- Logistic Regression
- Multinomial Naïve Bayes
- K-Nearest Neighbors
- AdaBoost Classifier
- Gradient Boost Classifier
- XGBoost Classifier
- Decision Tree Classifier
- Extra Trees Classifier
- Random Forest Classifier

Vectorizers

- Count Vectorizer
- TF-IDF Vectorizer

ML Models: Evaluation

BEST!

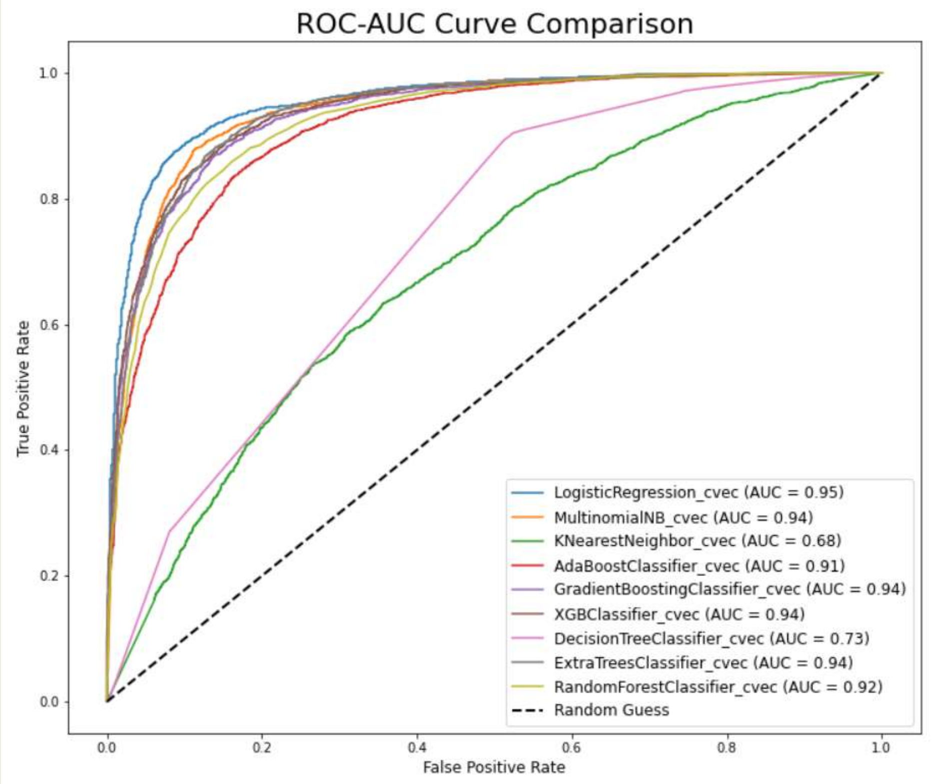
Logistic Regression with TF-IDF Vectorizer

```
Fitting 5 folds for each of 32 candidates, totalling 160 fits
============= Best model parameters ===============

{'lr__C': 10,
 'lr__class_weight': 'balanced',
 'lr__penalty': 'l2',
 'lr__solver': 'newton-cg',
 'tvec__max_df': 0.95,
 'tvec__max_features': None,
 'tvec__min_df': 4,
 'tvec__ngram_range': (1, 2),
 'tvec__stop_words': None}

================== METRICS ======================

{'model': 'lr',
 'train_auc': 1.0,
 'test_auc': 0.9657,
 'accuracy': 0.9054,
 'specificity': 0.8964}

True Negatives: 2786
False Positives: 322
False Negatives: 267
True Positives: 2851

array([[2786,  322],
       [ 267, 2851]], dtype=int64)
```

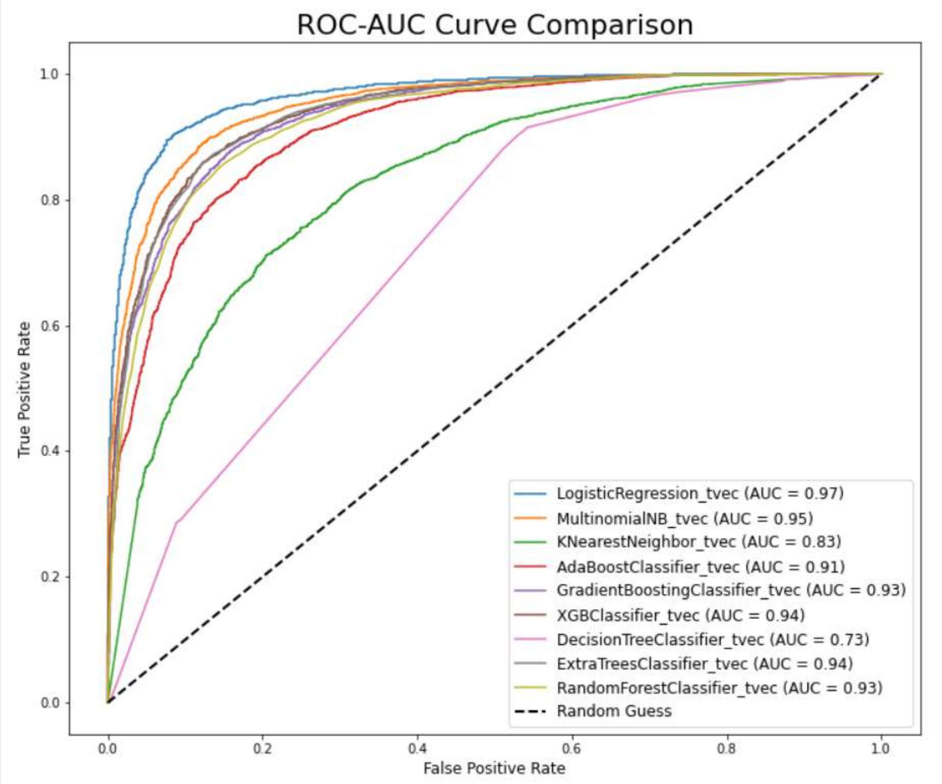
	model	train_auc	test_auc	accuracy	specificity	vectorizer
1	lr	1.0000	0.9657	0.9054	0.8964	tvec
0	lr	1.0000	0.9537	0.8908	0.8835	cvec
3	nb	0.9933	0.9505	0.8807	0.8887	tvec
2	nb	0.9936	0.9408	0.8824	0.8867	cvec
11	xgb	0.9988	0.9397	0.8648	0.8481	tvec
10	xgb	0.9972	0.9394	0.8678	0.8481	cvec
14	et	1.0000	0.9392	0.8704	0.8604	cvec
15	et	1.0000	0.9373	0.8673	0.8745	tvec
8	gb	0.9722	0.9356	0.8628	0.8362	cvec
9	gb	0.9784	0.9338	0.8577	0.8308	tvec
17	rf	1.0000	0.9273	0.8532	0.8571	tvec
16	rf	1.0000	0.9246	0.8487	0.8388	cvec
6	ada	0.9192	0.9129	0.8350	0.8137	cvec
7	ada	0.9260	0.9121	0.8304	0.8076	tvec
5	knn	1.0000	0.8286	0.7512	0.6866	tvec
12	dt	0.7451	0.7315	0.6907	0.4875	cvec
13	dt	0.7472	0.7297	0.6865	0.4701	tvec
4	knn	1.0000	0.6850	0.6303	0.4755	cvec

ML Models: ROC AUC Curves

Count Vectorizer

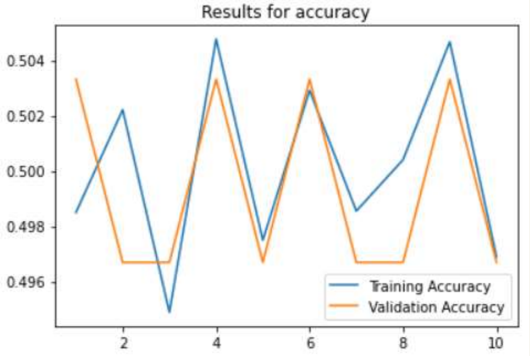


TF-IDF Vectorizer



Deep Learning Models: RNN

```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_4 (Embedding)     (None, 2447, 100)         8169500

 gru_4 (GRU)                 (None, 64)                31872

 dense_4 (Dense)             (None, 1)                 65

=================================================================
Total params: 8,201,437
Trainable params: 8,201,437
Non-trainable params: 0
_________________________________________________________________
```




```
Epoch 9/10
312/312 - 3467s - loss: 0.6932 - accuracy: 0.5047 - specificity: 0.4944 - precision_1: 0.5048 - average_metric: 0.5010 - recal
l: 0.5076 - val_loss: 0.6931 - val_accuracy: 0.5033 - val_specificity: 0.0000e+00 - val_precision_1: 0.5033 - val_average_metri
c: 0.5000 - val_recall: 1.0000 - 3467s/epoch - 11s/step
Epoch 10/10
312/312 - 3463s - loss: 0.6933 - accuracy: 0.4969 - specificity: 0.4026 - precision_1: 0.4975 - average_metric: 0.4974 - recal
l: 0.5922 - val_loss: 0.6932 - val_accuracy: 0.4967 - val_specificity: 1.0000 - val_precision_1: 0.0000e+00 - val_average_metri
c: 0.5000 - val_recall: 0.0000e+00 - 3463s/epoch - 11s/step
```

Conclusion & Further Improvements

- Best model: Logistic Regression with TF-IDF Vectorizer
  - Test ROC AUC: 0.9657
  - Accuracy: 0.9054
  - Specificity: 0.8964
- Use on other balanced datasets
- Effective on classifying sentiment polarities
- Reviews and comments from untapped data can be utilized for improvement of products, or product promotion/advertising

Further Improvements

- Tune RNN
- Other deep learning models
  - Word2vec embeddings
  - Pre-trained: BERT
- Ordinal Regression


[Words in negative reviews]
even though
bad movie
worst movie
movie would
ever seen
whole movie
main character
watch movie
sci fi
first time
make sense
bad film
really bad
good movie
waste time
like movie
movie one
feel like
one worst
special effect
movie bad
film ever
want see
bad acting
horror film
film made
movie made
movie even
horror movie
ever made
look like
pretty good
one thing
film one
movie really
movie like
much better
bad guy
high school
movie ever
film like
good thing
year ago
watching movie
low budget
year old
pretty much
make movie
worst film
see movie
[Words in positive reviews]
main character
one thing
must see
fall love
see film
movie great
one best
great film
high school
film also
take place
like movie
ever seen
see movie
look like
best film
year later
film made
film like
even though
special effect
film ever
supporting cast
movie like
real life
black white
good movie
pretty good
year old
done
make movie
horror film
many people
story line
new york
well
really good
great movie
year ago
sci fi
low budget
film one
ever made
movie good
first time
good film
watch movie
saw movie
love movie
movie ever
movie one

THANK YOU