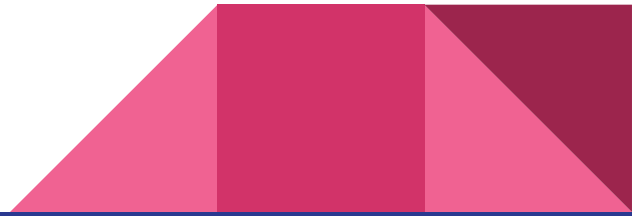
Predicting Amazon Book Reviews Ratings Using Customers Review

Jin Park

Objectives

- Data Cleaning / Exploratory Data Analysis.
- Predict Ratings using the sentiment analysis.

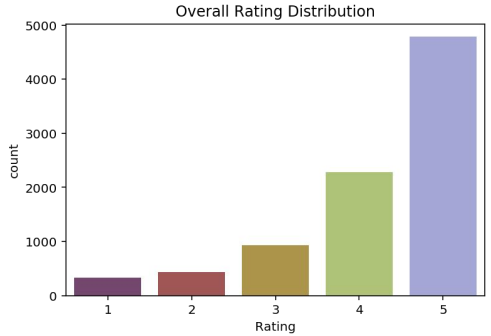


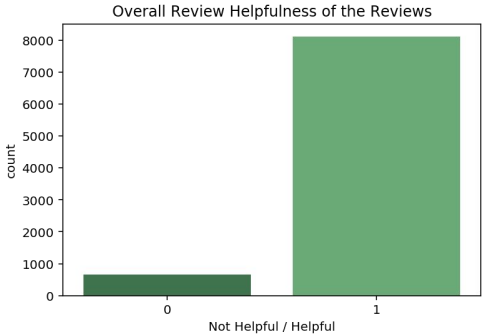
Data

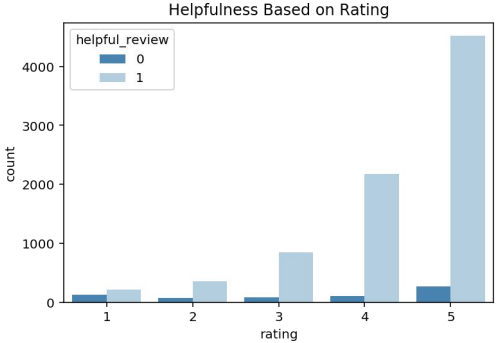
- The dataset for this project was collected from the University of California San Diego (UCSD) website, http://jmcauley.ucsd.edu/data/amazon.
- Dataset contains Amazon customer book reviews and metadata from May 1996 - July 2014.
- Web scraping Amazon.com to gather book categories by using Selenium.

Exploratory Data Analysis (EDA)

- Distributions of ratings, helpful counts, and helpful counts by ratings

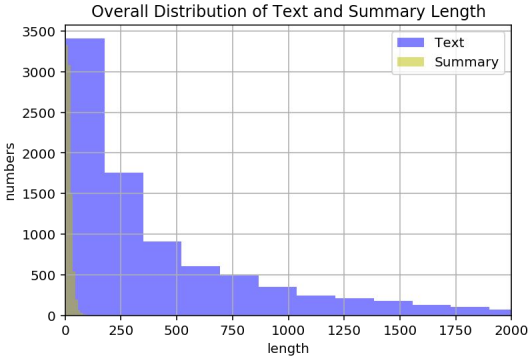


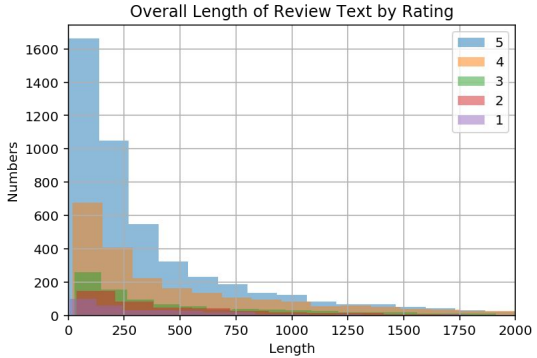


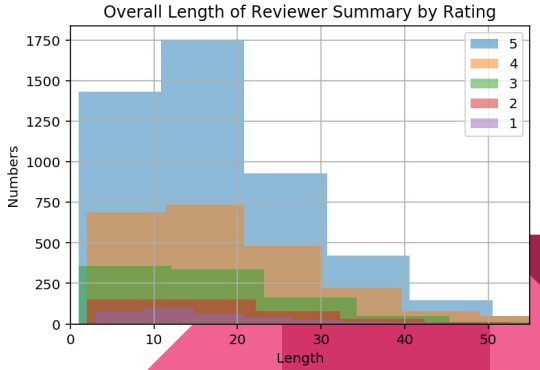


Reviews Length

- Length of review text and summary by ratings.







Top Least Occuring Words

	review_text	count
0	00	1
1	mitrokhin2	1
2	mitrokhins	1
3	mitt	1
4	mittenburg	1
5	mittens	1
6	mitts	1
7	mitzi	1
8	mixedup	1
9	mixeswhile	1

	review_text	count
0	00 farina	1
1	peopleat fascinating	1
2	peopleaside overused	1
3	peopleas story	1
4	peopleas actual	1
5	people34 mentality	1
6	people34 martin	1
7	people34 dumb	1
8	people34 could	1
9	people zombies	1

	review_text	count
0	00 farina flour	1
1	point view well	1
2	point view world	1
3	point view worthwhile	1
4	point view5 question	1
5	point viewand please	1
6	point viewcommentators political	1
7	point view various	1
8	point viewi would	1
9	point views concerning	1

	review_summary	count
0	lara	1
1	nazism	1
2	nazi	1
3	nazarea	1
4	navarro	1
5	naughty	1
6	nativity	1
7	native	1
8	neal	1
9	nate	1

	review_summary	count
0	10 best	1
1	pace star	1
2	paced adventure	1
3	paced brutal	1
4	paced historical	1
5	paced romance	1
6	paced suspenseful	1
7	paced thrill	1
8	paced unique	1
9	pacedkept attention	1

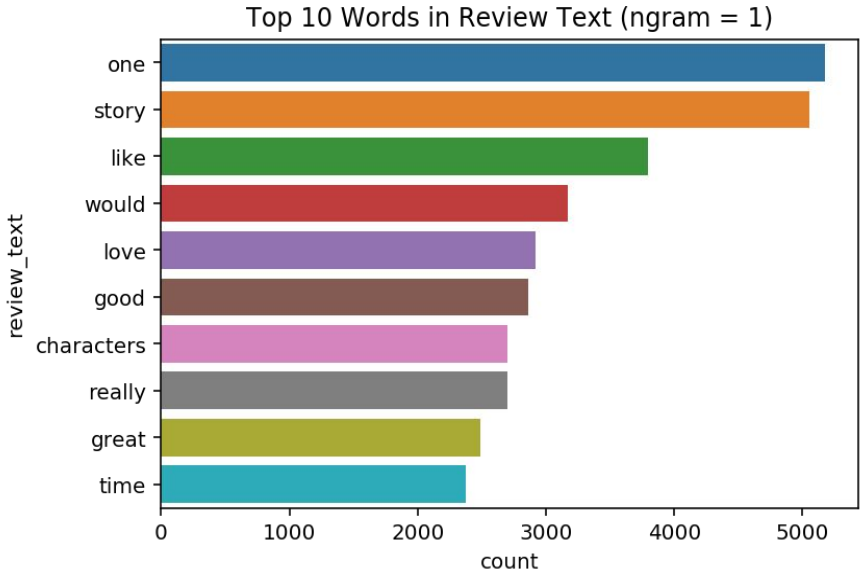
	review_summary	count
0	10 best novels	1
1	pageturner sad sad	1
2	pages turning rapid	1
3	pages keep comming	1
4	page turner memoir	1
5	page bronies rule	1
6	page amazonspecifichtml codes	1
7	paen 80s nerd	1
8	padding short story	1
9	packed well ordered	1

Most Frequent Words in Reviews Word Clouds

Overall Most Frequent Words in Review Text

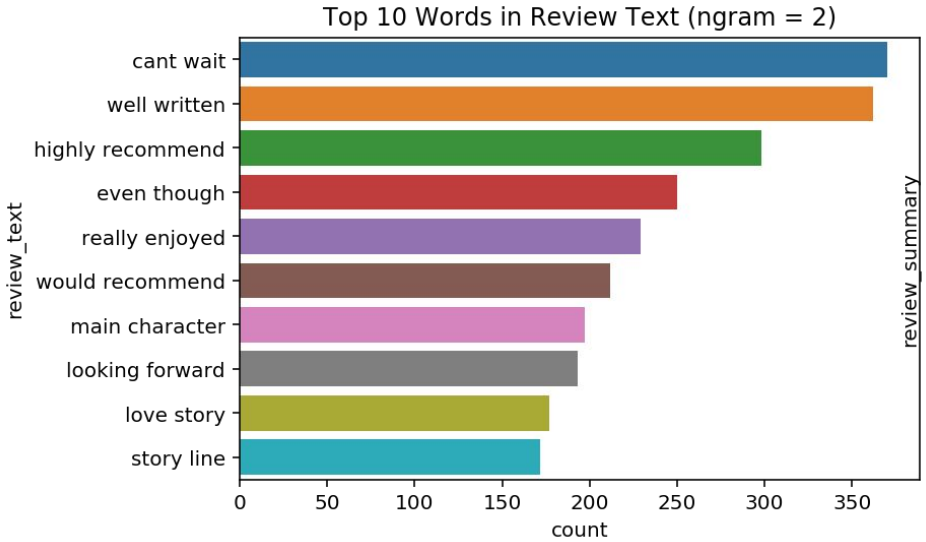
Overall Most Frequent Words in Review Summary

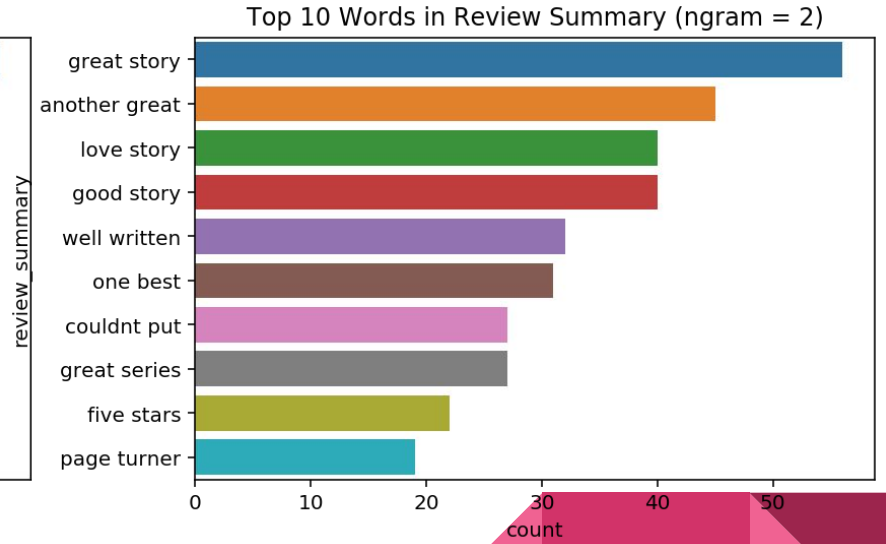
Unigram


Top 10 Words in Review Text (ngram = 1)
review_text
one
story
like
would
love
good
characters
really
great
time
0
1000
2000
3000
4000
5000
count

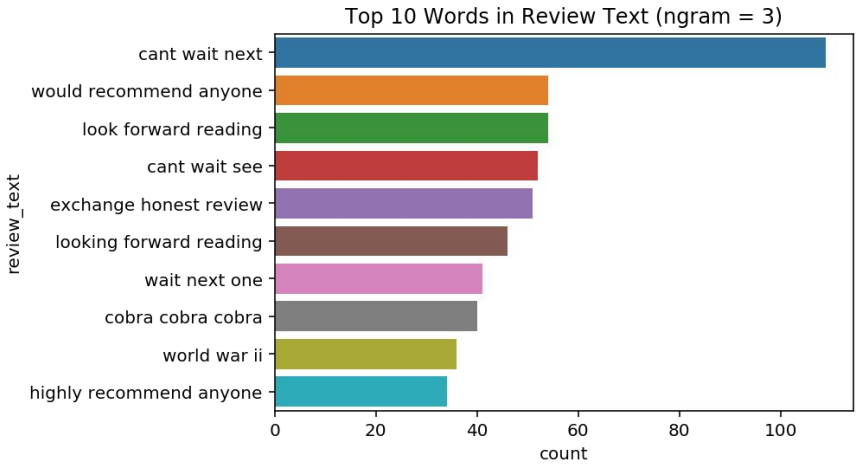

Top 10 Words in Review Summary (ngram = 1)
review_summary
great
good
story
love
series
one
best
fun
loved
excellent
0
100
200
300
400
500
600
700
800
count

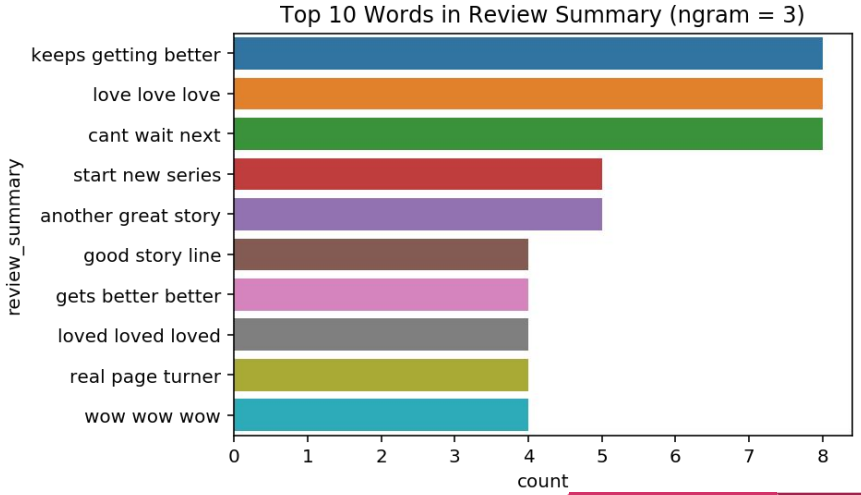
Bigrams


Top 10 Words in Review Text (ngram = 2)
cant wait
well written
highly recommend
even though
really enjoyed
would recommend
main character
looking forward
love story
story line
review_text
0
50
100
150
200
250
300
350
count

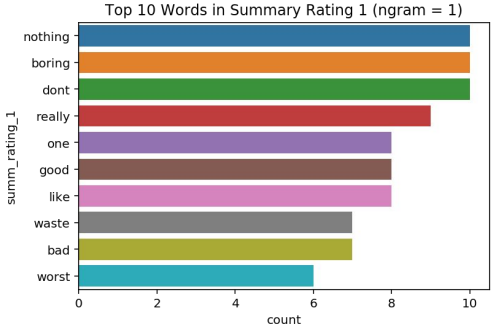

Top 10 Words in Review Summary (ngram = 2)
great story
another great
love story
good story
well written
one best
couldnt put
great series
five stars
page turner
review_summary
0
10
20
30
40
50
count

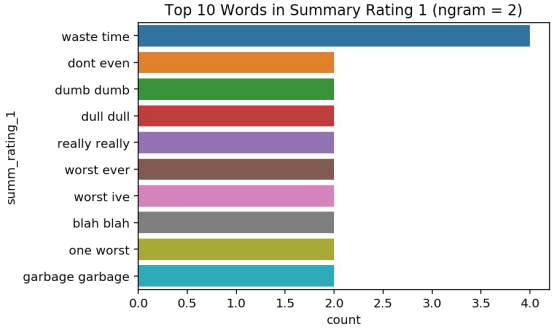
Trigrams

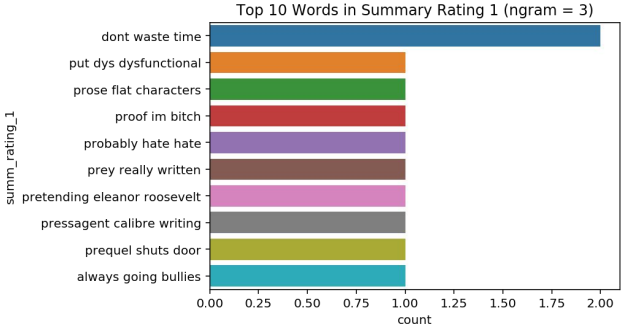

Top 10 Words in Review Text (ngram = 3)
cant wait next
would recommend anyone
look forward reading
cant wait see
exchange honest review
looking forward reading
wait next one
cobra cobra cobra
world war ii
highly recommend anyone
review_text
0
20
40
60
80
100
count


Top 10 Words in Review Summary (ngram = 3)
keeps getting better
love love love
cant wait next
start new series
another great story
good story line
gets better better
loved loved loved
real page turner
wow wow wow
review_summary
0
1
2
3
4
5
6
7
8
count

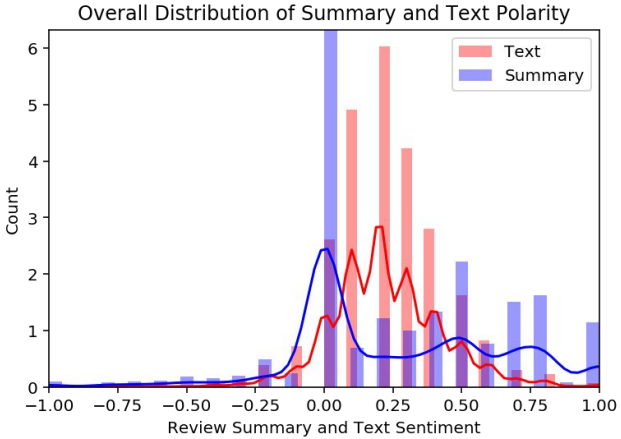
Lowest Rating Words

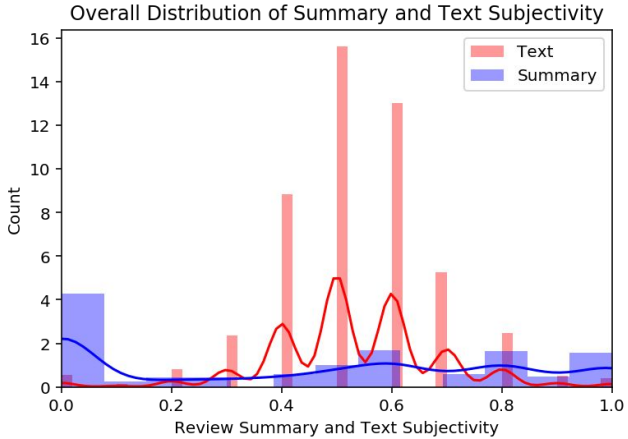




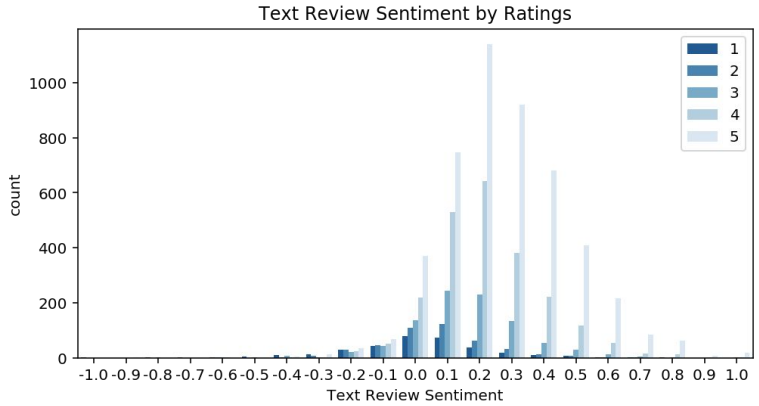


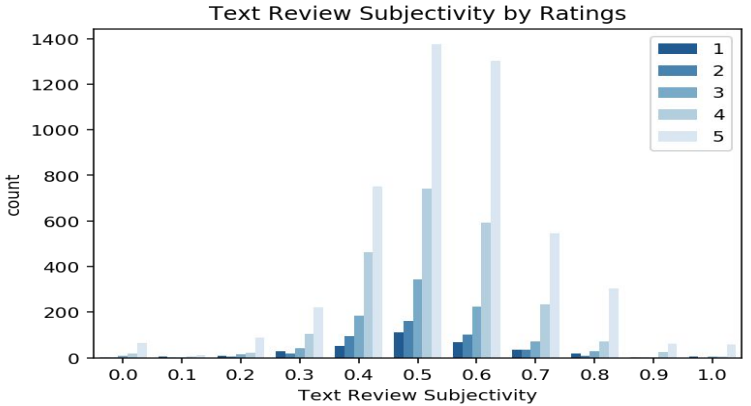
Sentiment Analysis

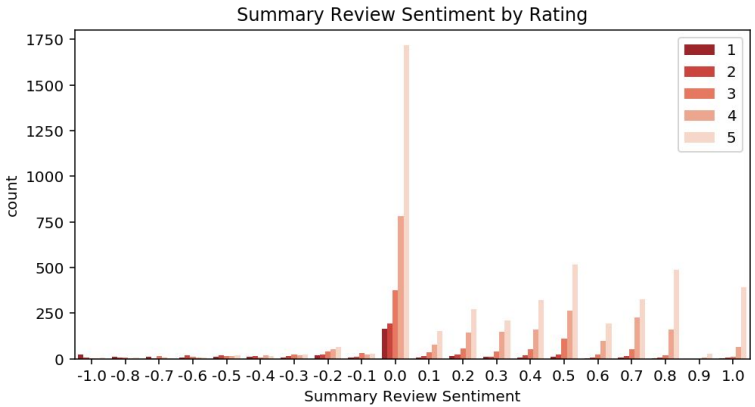


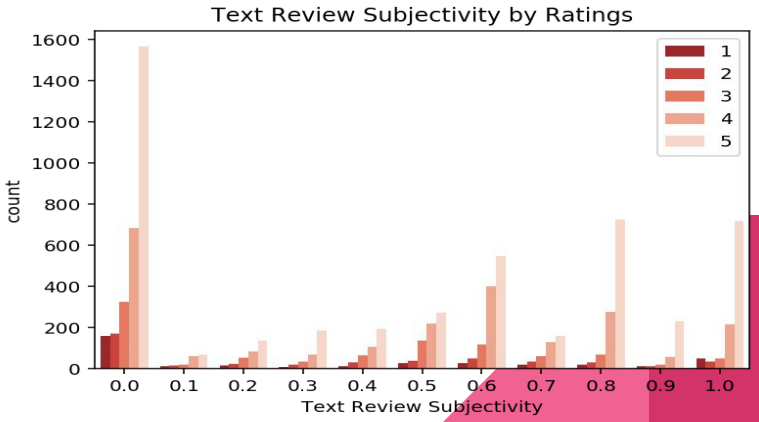


Sentiment Analysis by Ratings

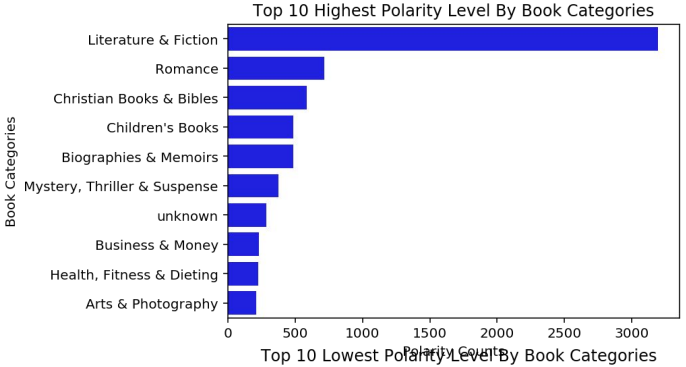


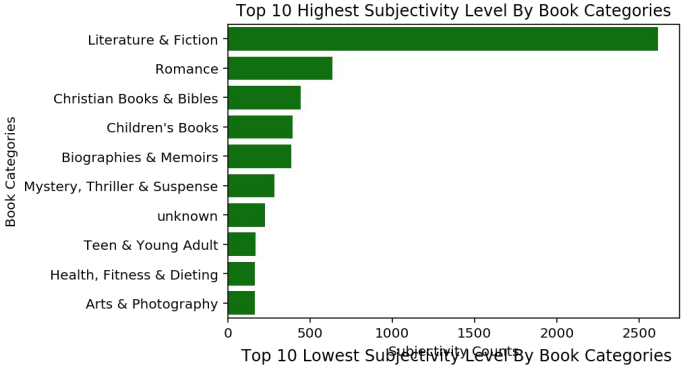


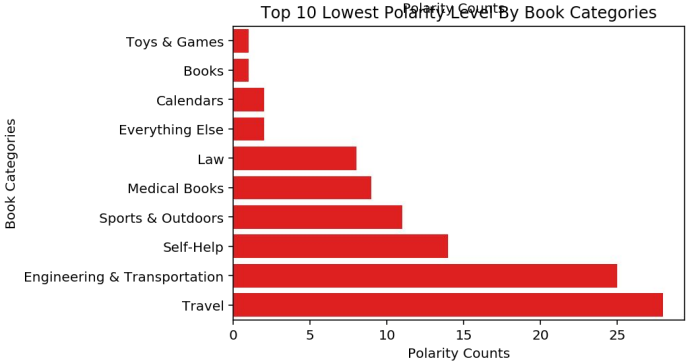


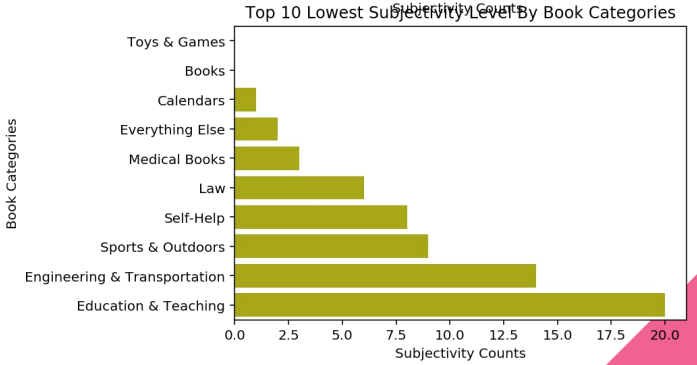


Polarity and Subjectivity Level by Book Categories

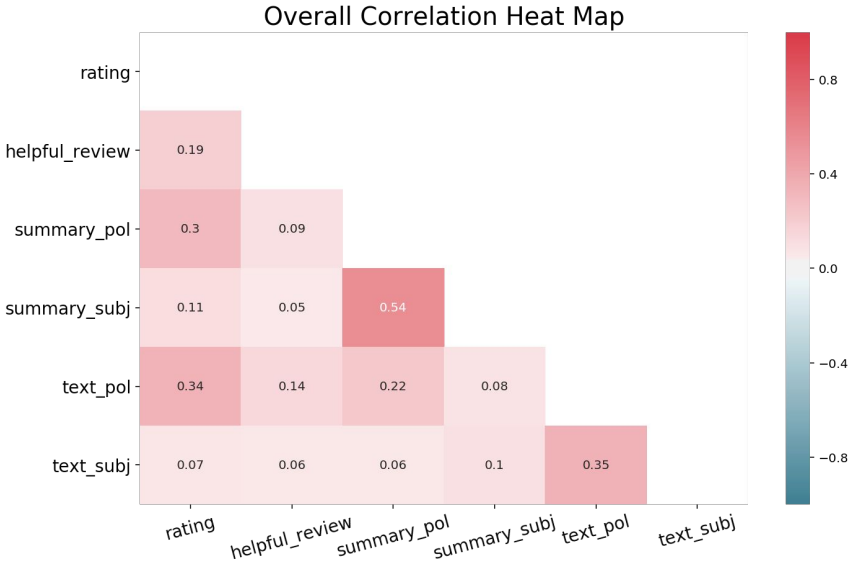






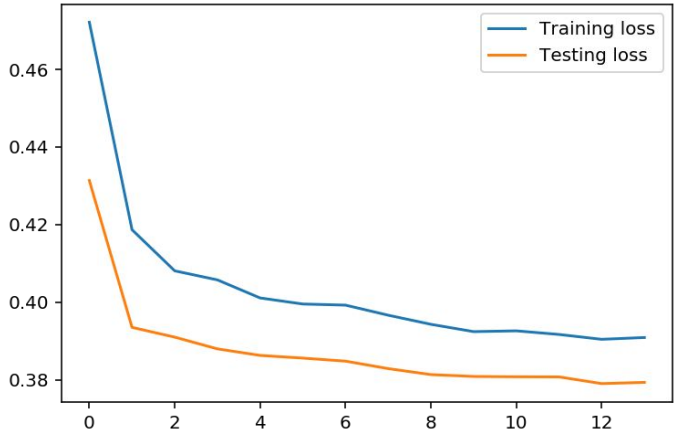


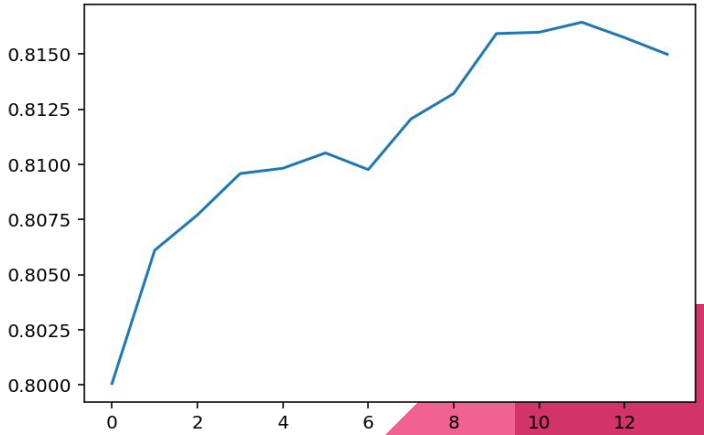
Correlation Heatmap



Modeling

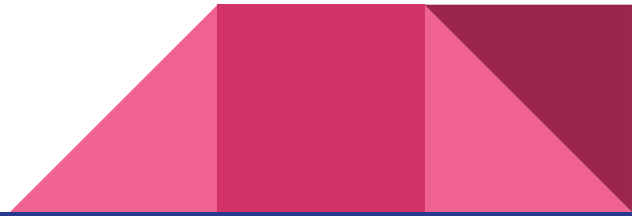
- Baseline: 55%
- RandomForest Classifier, Logistic Regression, Neural Network using Keras.
- Neural Network: 82% Accuracy on test set.





Conclusion

- The product ratings can be predicted just by using customers sentiment analysis.
- Text data itself could not outperform the neural network that uses sentiment analysis as features.



Any Questions?

This is the end of my presentation and thank you for listening.