

Amazon Book Reviews:

Classification Modeling & Sentiment Analysis

By: Liisa Isaacson
March 17, 2023
Nod Coding Bootcamp 2023-1
Project 6

Dataset used from Kaggle.com
(<https://www.kaggle.com/datasets/mohamedbakhmet/amazon-books-reviews>)



Agenda

01

Introduction

02

Building my
Model

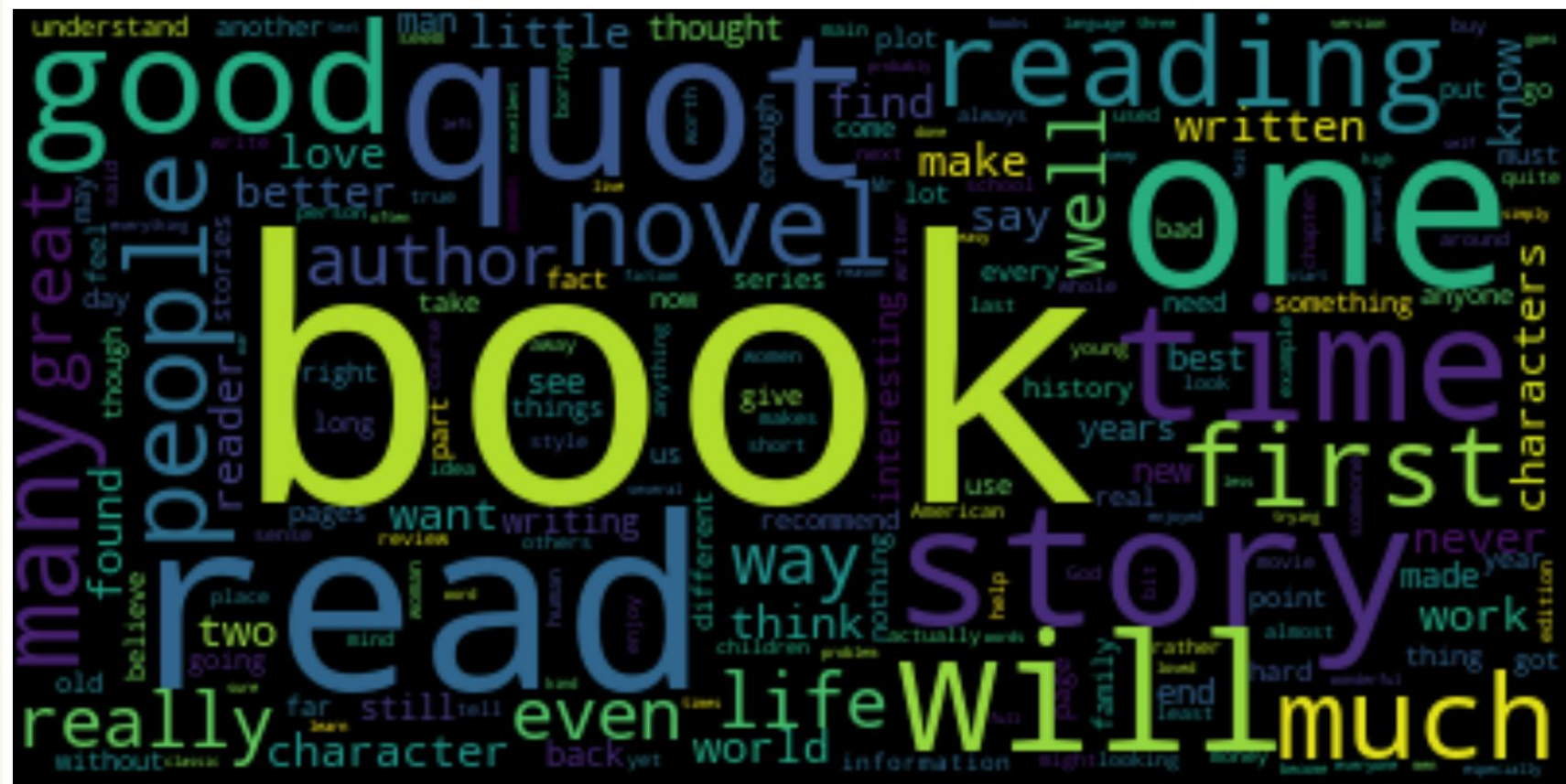
03

Sentiment
Analysis with
other models

04

Conclusion







4/5

"This is a good place," he said.

"There's a lot of liquor," I agreed."

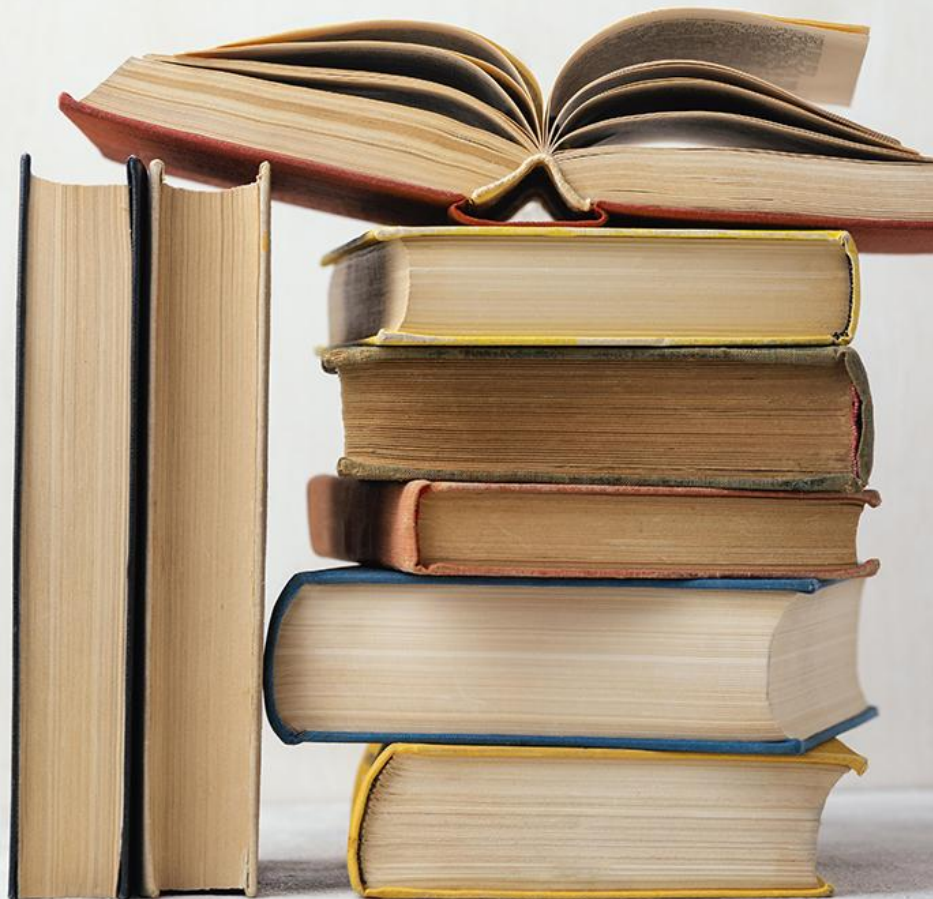
— Ernest Hemingway
The Sun Also Rises

"The Sun Also Rises (TSAR) chronicles the drinking and traveling of a handful of privileged young Americans in Europe. Picture MTV's Real World in the 1920's with lots of wine."

— Amazon Reviewer

Machine Learning

Building a
Classification
model using
NLP



“

My Goal:

Create a machine learning model that can **accurately** classify negative and positive reviews based on the review text.



The data

Rating

Review Text

Building an NLP Classification Model



Prepare the data

Create a sample dataframe,
take care of missing values,
duplicates, unnecessary columns,
etc.

1's and 2's became
Negative (0)

1 2 3 4 5

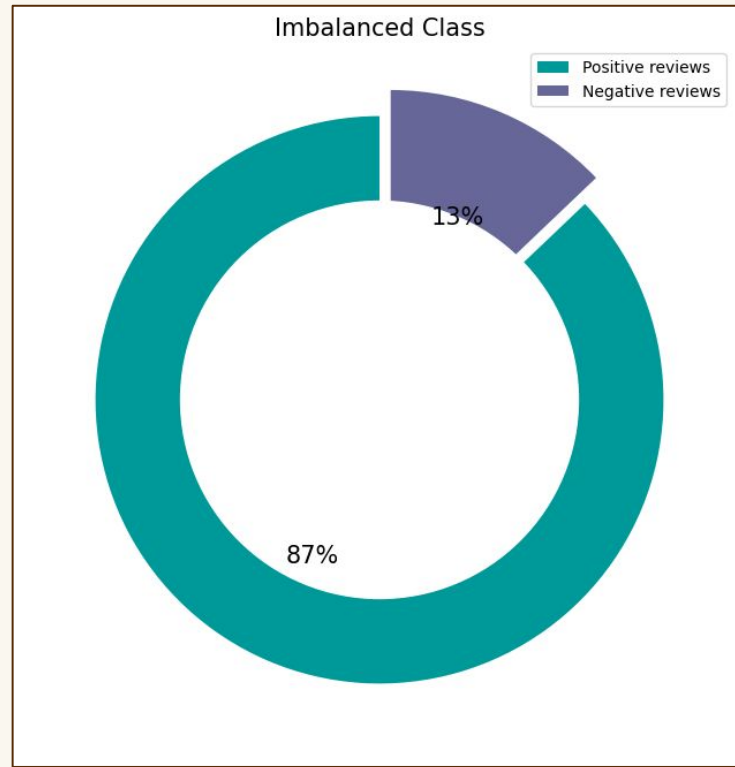
4's and 5's became
Positive (1)

Building an NLP Classification Model



Decide on the Baseline

Because of the imbalanced class, I chose the **Zero Rate Classifier** as my baseline: my target is for my model to achieve better than **87%** accuracy.



Building an NLP Classification Model


Preprocessing



- ❑ Balanced sample
- ❑ Split sentences into words
- ❑ Remove unneeded characters
- ❑ Change all to lowercase
- ❑ Remove “stopwords” - can even modify list
- ❑ Lemmatize text - change all variants to a single “root” of the word
- ❑ Count Vectorizer - creating matrix of words, with adjustments for strings
- ❑ TF-IDF assigns values based on frequency

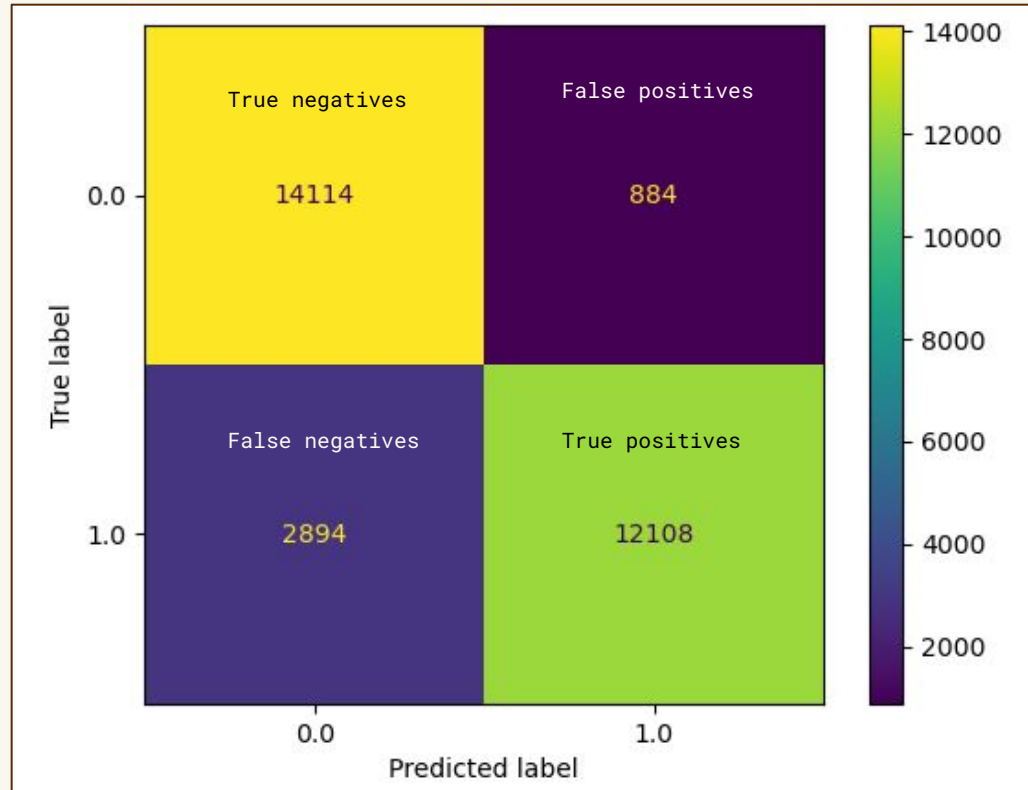
review/summary	review/text	tokenized_text	cleaned	review_cleaned_lemmatized
Sprout your business!	I read Sprout! in a couple of hours. Once I st...	[read, sprout, I, couple, hours, ., started, c...	read sprout ! couple hours . started could n't...	read sprout couple hour start could nt put ...
Finally, a balanced history text	Some of the reviews posted here are just bizar...	[reviews, posted, bizarre, -, read, ?, yes, ,....	reviews posted bizarre - read ? yes , writes m...	review post bizarre read yes writes minorit...
This bio has		[bio, merit	bio merit	

Model Selection

Model	Training Score (Accuracy)	Training Score after testing Hyperparameters
LogisticRegression	0.94147	0.89441 (C score)
NaiveBayes - Multinomial Bayes	0.97721	0.87251 (Alpha score)
XGBoost	0.89147	

Interpreting the Test Results

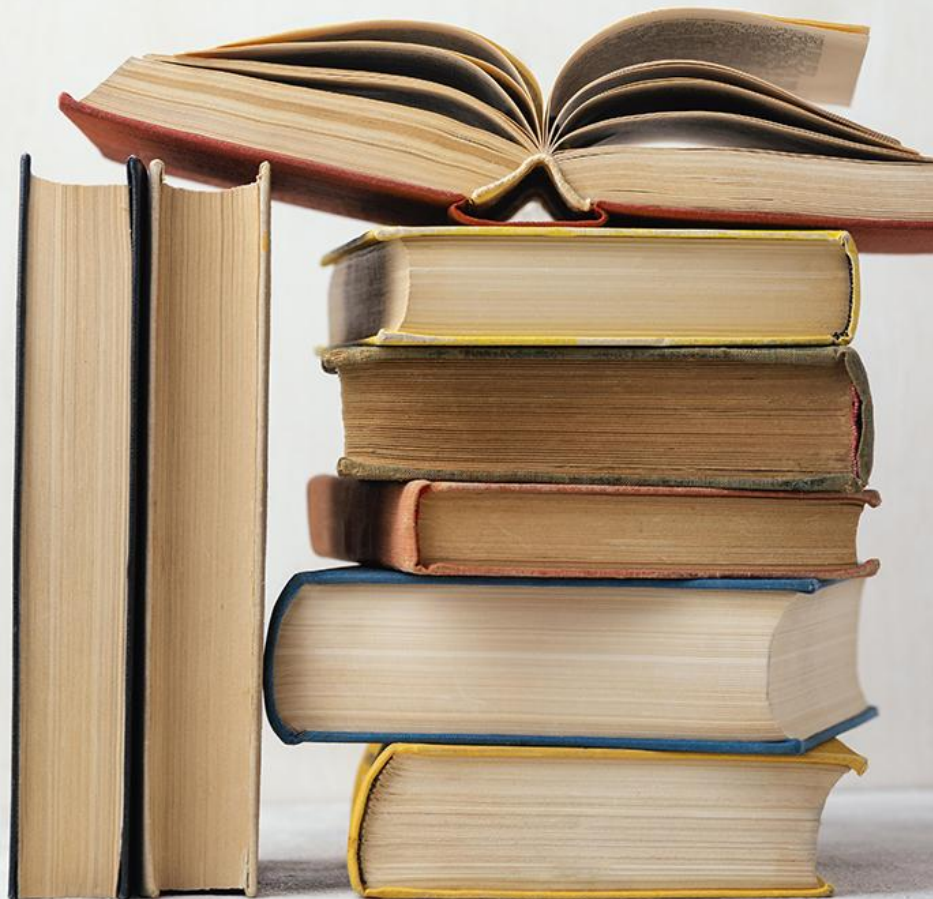
Accuracy	87.4%
Precision	93.2%
Recall	80.7%
Specificity	94.1%



Sentiment Analysis

Comparing the
VADER
&
Roberta
Models

“





“

Valence Aware Dictionary for sEntiment Reasoning

VADER is a model in the Natural Language Toolkit(NLTK) that provides sentiment scores on a given word input.

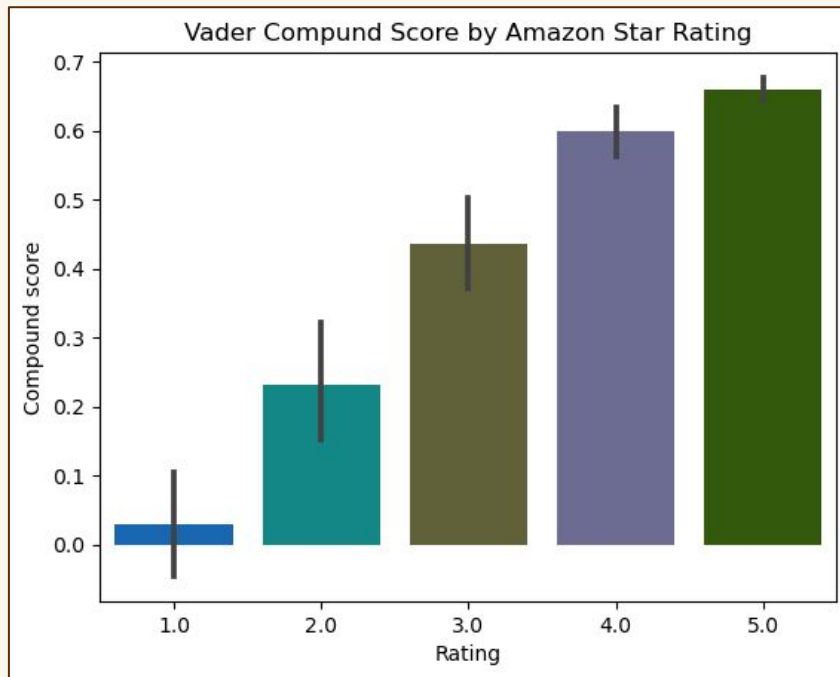
VADER Results on sample set of reviews

VADER produces four scores:

1. **Negative**
2. **Neutral**
3. **Positive**
4. **Compound**

The **Compound** score is the sum of positive, negative & neutral scores which is then normalized between -1 (most extreme negative) and +1 (most extreme positive).

The more Compound score closer to +1, the higher the positivity of the text.





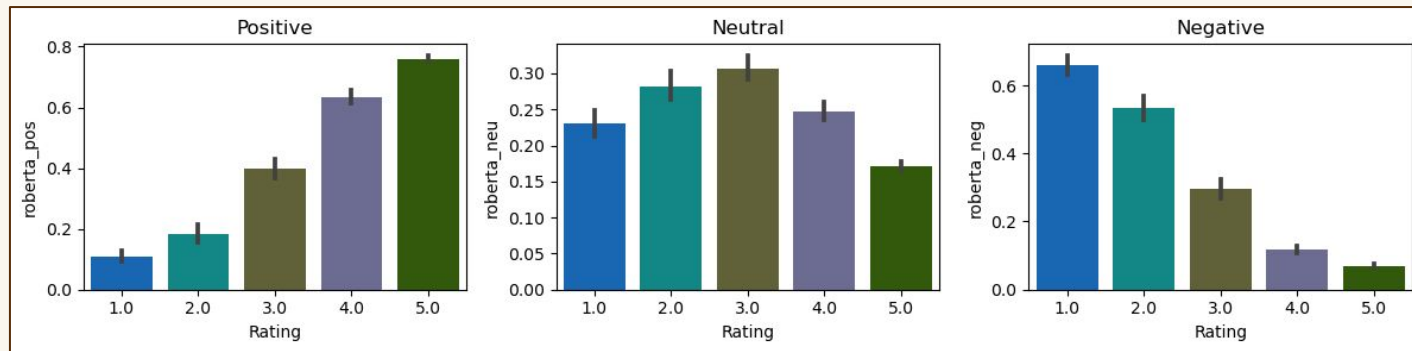
Robustly Optimized BERT Approach

ROBERTA is a variant of the BERT model. The difference is that ROBERTA was trained on a much larger dataset and is able to learn more robust and generalizable representations of words.

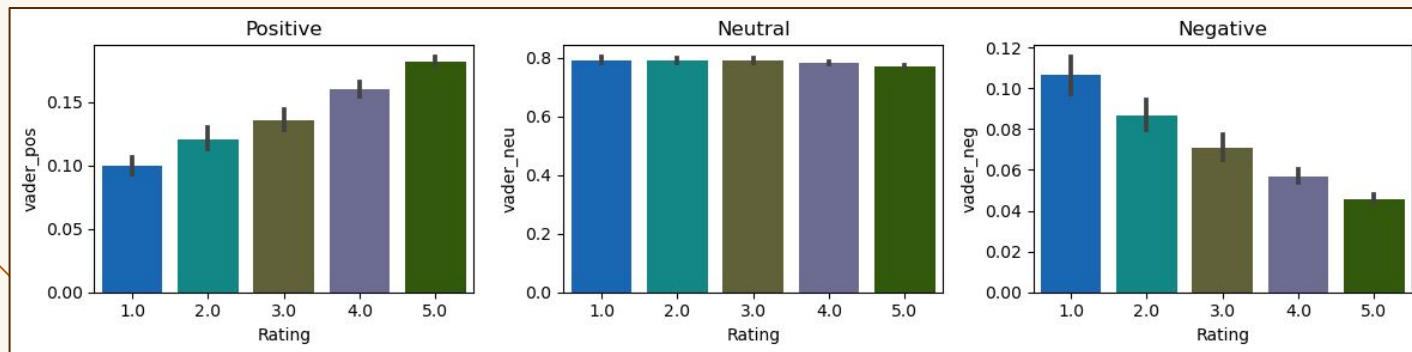
ROBERTA scoring



ROBERTA produces three scores: a negative, a neutral, and a positive. Each represent percent probability of the text falling into said category



A comparison of the models shows that ROBERTA is more nuanced in assigning sentiment scores.



This is especially apparent in the Neutral category.

In conclusion

Challenges



Large data and deep learning techniques take a lot of computing power



Reviewers are tricky.
It's easy to see why ROBERTA gave this 1-star rating a positive score of 0.991874:

On boy! This book is fantastic! There are so many good advices for me and you!'You can easily with the red score! This is one of the best books in the world!Hups... sorry.. I was just kidding :)"

Learnings



Sentiment analysis is fun and I think we're scratching the surface of what we can learn from it

"The Sun Also Rises (TSAR) chronicles the drinking and traveling of a handful of privileged young Americans in Europe. Picture MTV's Real World in the 1920's with lots of wine."

VADER compound score	-0.7536 Negative
ROBERTA highest score	0.4715 Neutral
Amazon Rating	4 Positive

"This is a good place," he said.

"There's a lot of liquor," I agreed."

VADER compound score	0.6124 Positive
ROBERTA highest score	0.8121 Positive
My Rating	5/5 Positive



Thanks!

CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)

Please keep this slide for attribution