AMAZON REVIEW POLARITY CLASSIFICATION

SPRINGBOARD DATA SCIENCE CAPSTONE PROJECT #3

PREPARED BY MIKE (XIANGNAN) SHI

MENTOR: BLAKE ARENSDORF

OUTLINE

- Project overview
- Problem statement
- Exploratory data analysis
- Modeling
- Summary
- Future research

PROJECT OVERVIEW

- Goal: Develop a model to classify Amazon reviews as positive or negative.
- Tool: Python, Jupyter Notebook
- Workflow:
  - Define the problem
  - Collect, clean and explore the data
  - Create and train models based on training set
  - Validate the models using test set and evaluate the performance

PROBLEM STATEMENT

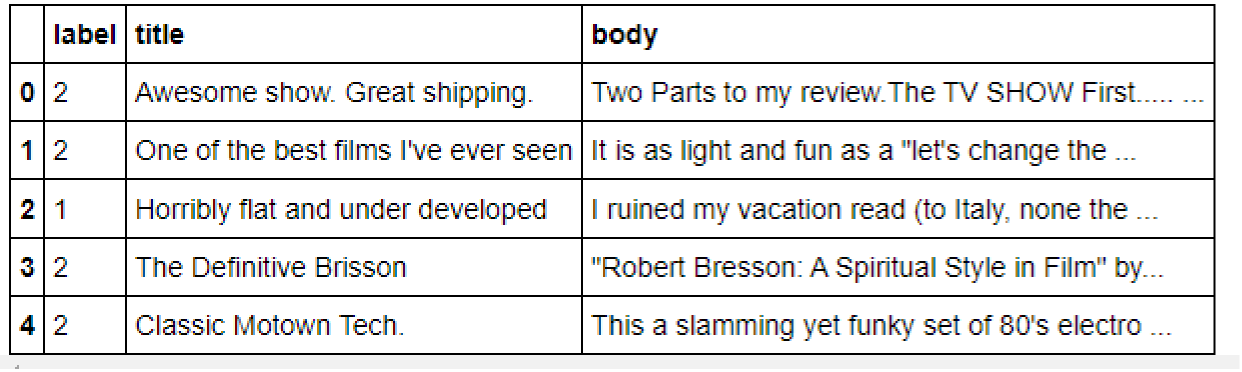
PROBLEM STATEMENT

- Context:
  - Given a large volume of customer reviews, how many of them are positive or negative?
- Stakeholders:
  - Business owners, product managers, marketing.
- Scope of solution
  - Build different models that can take review text data as input and predict sentiment
  - Compare the performance of different models
- Data:
  - Source: https://www.kaggle.com/kritanjalijain/amazon-reviews

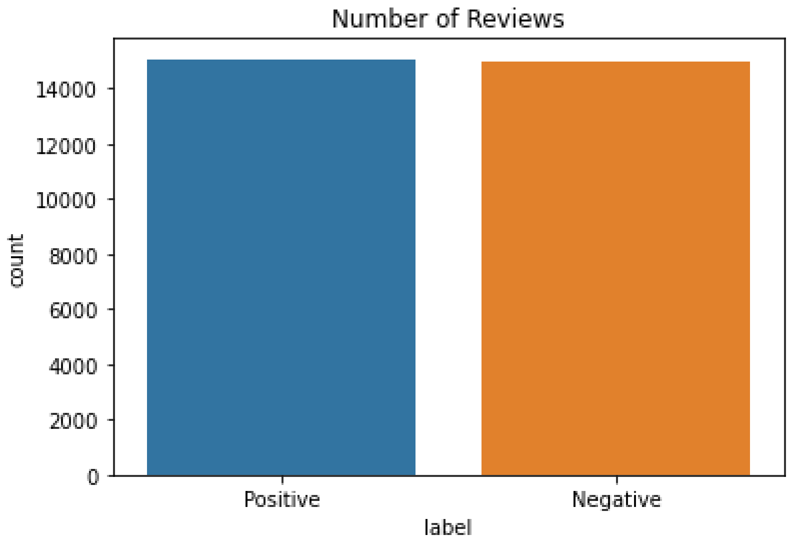
EXPLORATORY DATA ANALYSIS (EDA)

DATA OVERVIEW

- Data source: https://www.kaggle.com/kritanjalijain/amazon-reviews
- Original data set has 2 million customer reviews on Amazon. A subset of 30,000 reviews were used for this study.
- Target: label (1-negative, 2-postive, 50/50 split)
- Feature: review title, review body

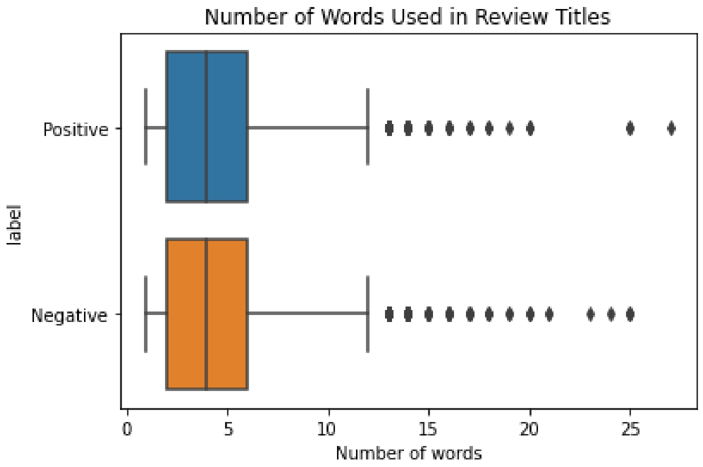


	label	title	body
0	2	Awesome show. Great shipping.	Two Parts to my review.The TV SHOW First..... ...
1	2	One of the best films I've ever seen	It is as light and fun as a "let's change the ...
2	1	Horribly flat and under developed	I ruined my vacation read (to Italy, none the ...
3	2	The Definitive Brisson	"Robert Bresson: A Spiritual Style in Film" by...
4	2	Classic Motown Tech.	This a slamming yet funky set of 80's electro ...

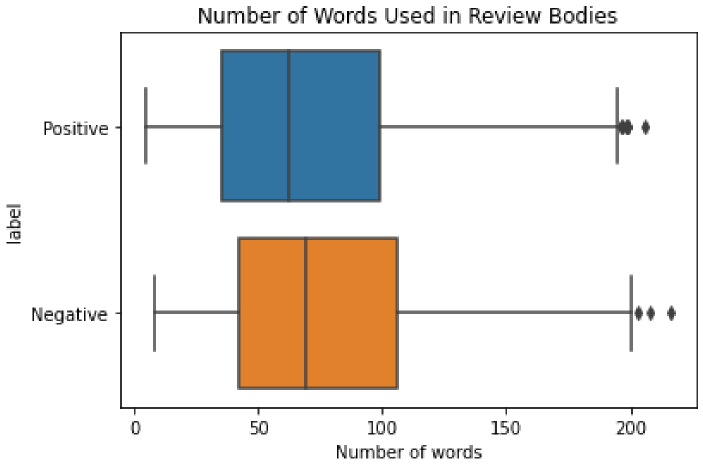


REVIEW LENGTH

- Review title
  - Most review titles are 2-6 words long.
  - There isn’t obvious difference between positive and negative reviews in terms of review length.

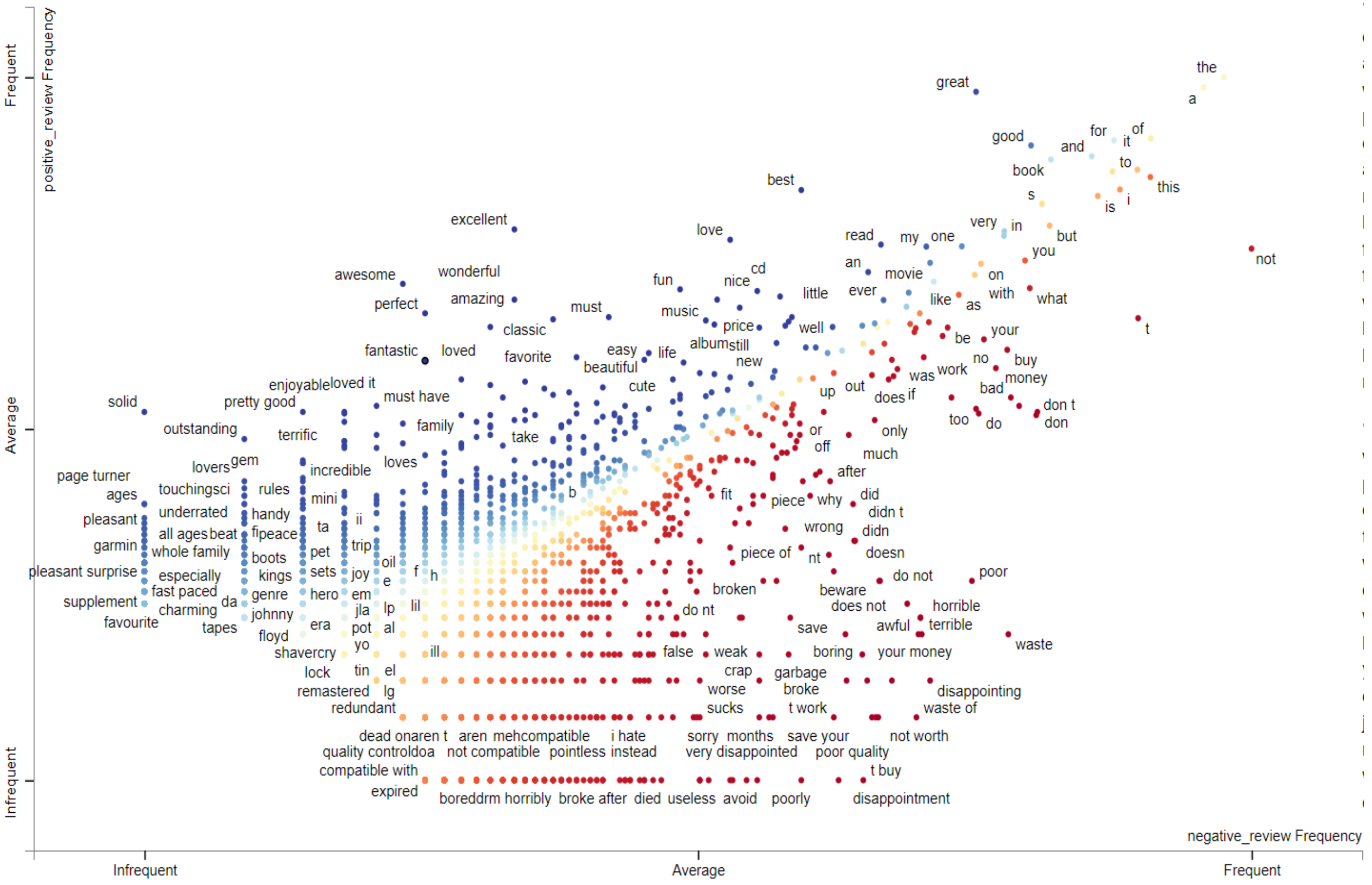

Number of Words Used in Review Titles

- Review body
  - Most review bodies are 30-110 words long.
  - An interesting fact is that negative reviews tend to be a little longer than positive reviews.


Number of Words Used in Review Bodies

WORD FREQUENCY IN NEGATIVE VS POSITIVE REVIEWS

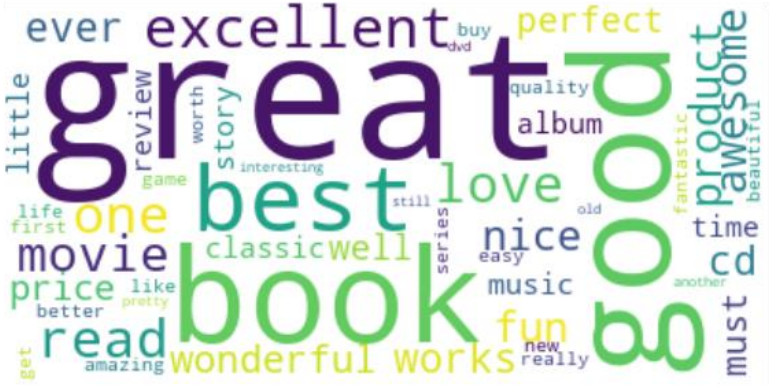
- On the bottom right, words appear more frequently in negative reviews, such as ‘disappointment’, ‘poor’, etc.
- On the upper left, words appear more frequently in positive reviews, such as ‘solid’, ‘awesome’, etc
- The scatter points are denser when they are close to bottom left, which means there’re more infrequent words than frequent words.



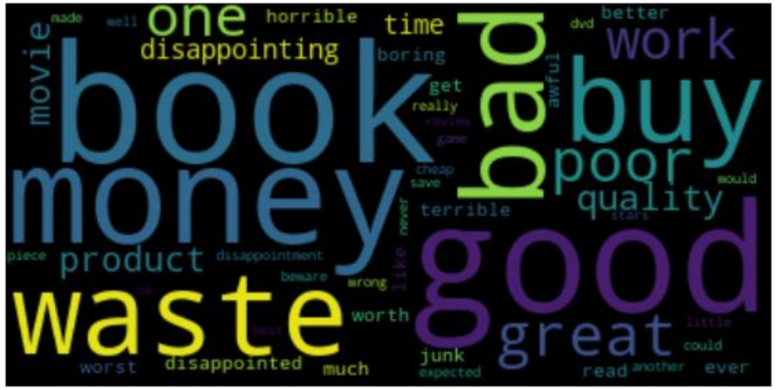
WORD CLOUD

- Two word cloud maps are generated to show the 50 most frequent words in positive and negative review titles.
- Unlike the scatter text plot before, the word cloud maps don’t show the relative frequency of a word in positive reviews comparing to negative reviews.
- Some words just appear a lot in both types of reviews, such as book, good, buy, read, etc.

Word Cloud of Positive Reviews

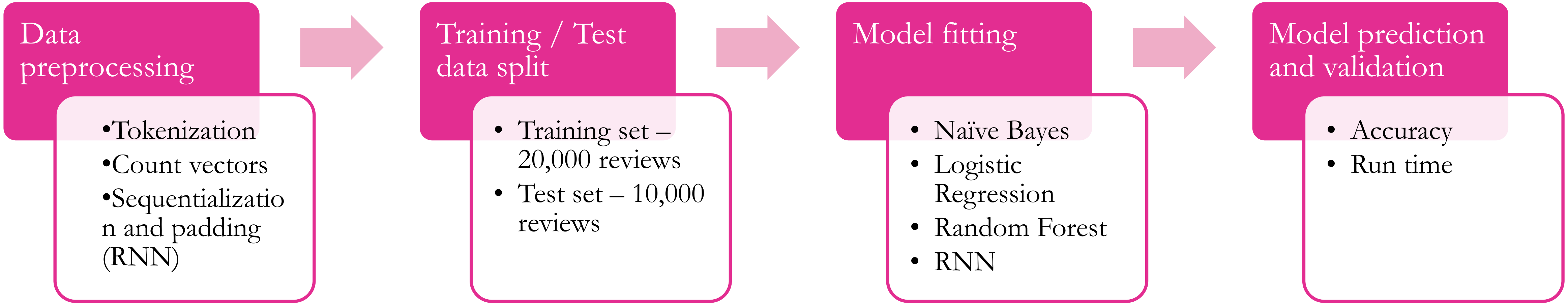


Word Cloud of Negative Reviews



MODELING

MODELING



Data preprocessing

- Tokenization
- Count vectors
- Sequentialization and padding (RNN)

Training / Test data split

- Training set – 20,000 reviews
- Test set – 10,000 reviews

Model fitting

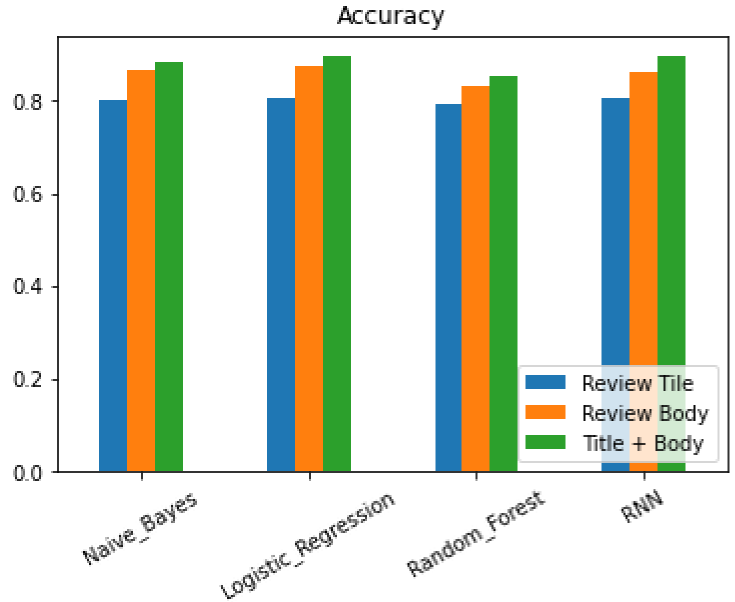
- Naïve Bayes
- Logistic Regression
- Random Forest
- RNN

Model prediction and validation

- Accuracy
- Run time

PERFORMANCE SUMMARY

- The accuracy ranges from 0.79 to 0.90.
- Using review body for inputs has better result than using review title alone.
- The performance of all models seem to be very close without much hyperparameter tuning done.
- Model run time can vary greatly depending on the method picked and parameter setting.



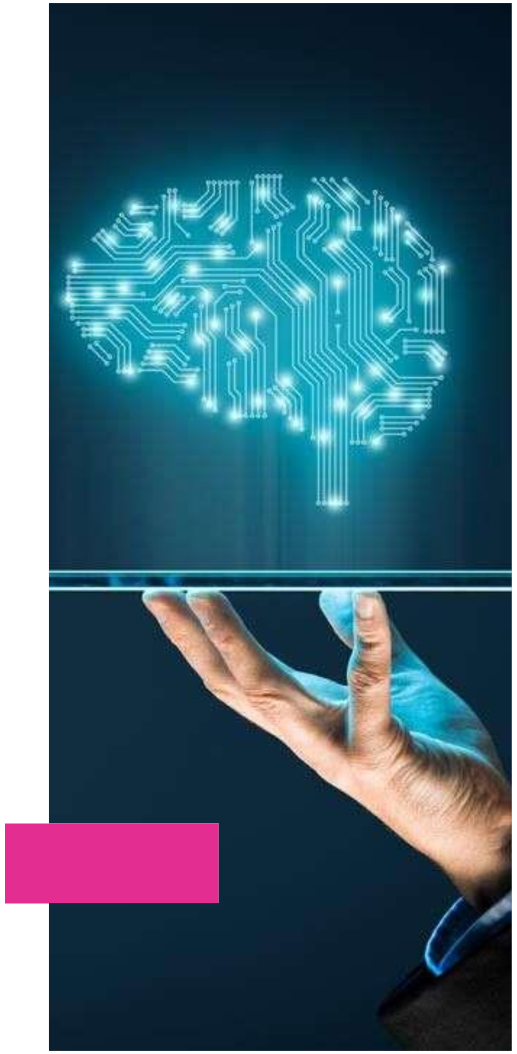
	Model_Accuracy (Title)	Model_Accuracy (Body)	Model_Accuracy (Title&Body)	Model_Run_Time (secs)
Naive_Bayes	0.79988	0.866187	0.883888	0.212183
Logistic_Regression	0.805681	0.874487	0.89539	12.955616
Random_Forest	0.794279	0.832683	0.855586	78.578343
RNN	0.805381	0.862886	0.89529	Depending on epochs

FUTURE RESEARCH

Additional data preprocessing is definitely worth exploring, such as misspelling correction, stemming and lemmatizing, etc.

Hyperparameter tuning isn’t implemented in this study. As a result, the performance achieved is likely not close to optimal. It’s strongly recommended to incorporate tuning as a future exploration topic.

It’s recommended to extend the training set to the original data set that has 2 million reviews, which is a lot more than what this study uses. It’s expected that the accuracy can increase to some extent with more training data.



CONTACT

LinkedIn
https://www.linkedin.com/in/mike-shi-pe-cem-14029a26/

GitHub
https://github.com/stonewatertx

Email
stonewatertx@email.com