MINING USERS' OPINIONS FROM AMAZON REVIEWS

BY XINYU ZHANG

GOALS

Through mining amazon review text to understand:

When reviewers like or dislikes a certain product, what do they like/dislike about?

DATA COLLECTION

METHODS AND RESULTS

- Data cleaning and integration
- Exploratory data analysis
- Sentiment analysis to identify the positive opinions and negative opinions
- Extracting the relevant product features and representative sentences

METHODS AND RESULTS

- Data cleaning and integration
- Exploratory data analysis
- Sentiment analysis to identify the positive opinions and negative opinions
- Extracting the relevant product features and representative sentences

DATA CLEANING AND INTEGRATION

There are in total 263032 products in Health and personal care department

Product metadata contain:

- product ID.
- brief description of the product
- other products that users who bought this product also bought.
- the rank of the sale amount of product.
- categories that a product can belong to.
- price
- brands

Review data contains:

- the reviewers' ID
- reviewers' user name
- the number of people who find the review helpful or not helpful
- the review text
- the rating (1 – 5)
- summary of the review
- time the review was uploaded
- product ID

Two datasets are linked by product ID

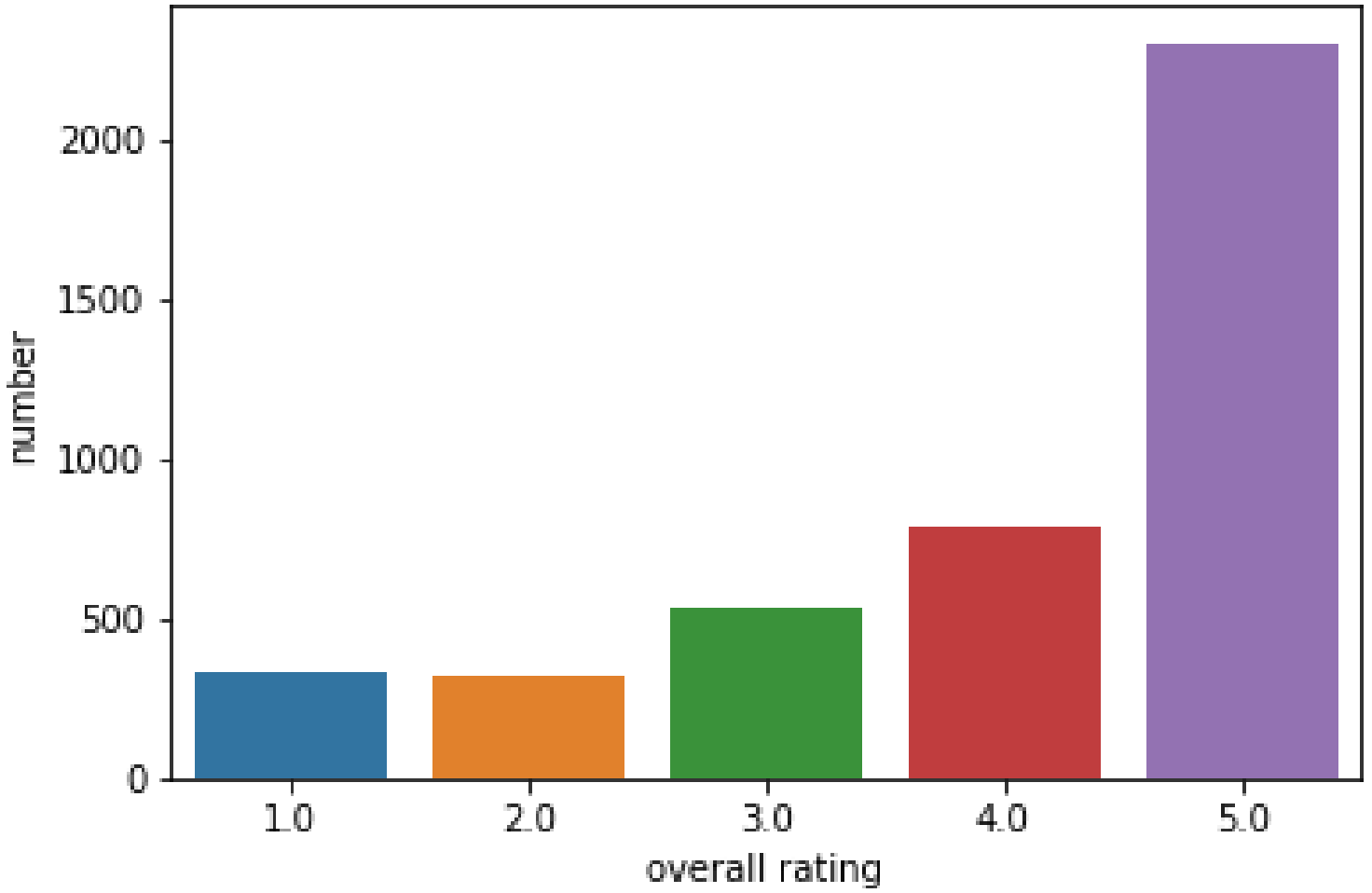
DATA CLEANING AND INTEGRATION

- Inner join product metadata with review data by column of product ID
- Select products under category “Sleep & Snoring”
- Select subset of columns for sentiment analysis
  - Review ID
  - product ID
  - Review text
  - Rating
  - Summary

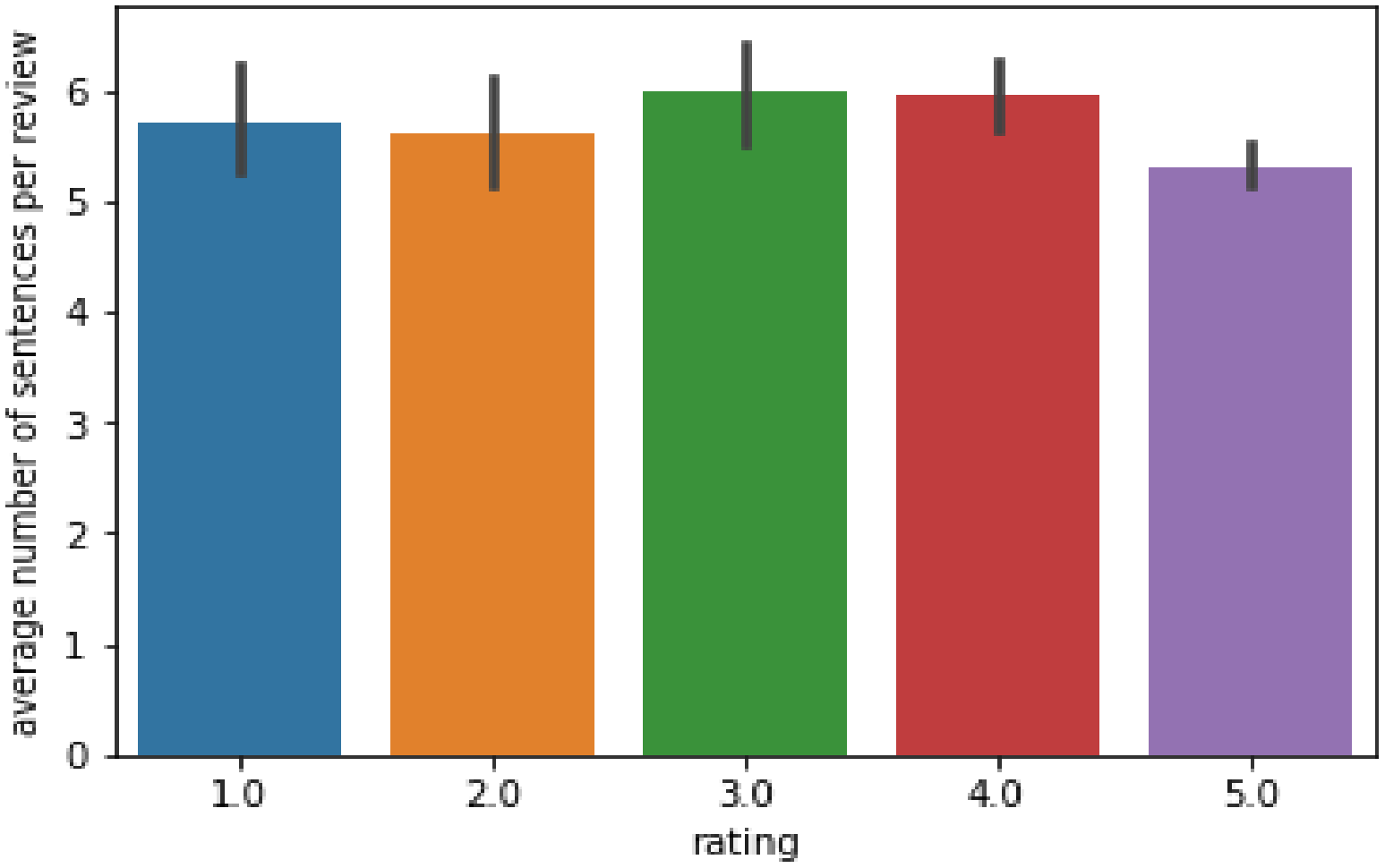
METHODS AND RESULTS

- Data cleaning and integration
- Exploratory data analysis
- Sentiment analysis to identify the positive opinions and negative opinions
- Extracting the relevant product features and representative sentences

Number of reviews by rating



Average number of sentences per review by rating



An example of review text and summary:

- *review text:*

"Ok... so I got this because a friend recommended melatonin to help me sleep at night. The first night I tried it I got a little relaxed, fuzzy sleeping feeling for maybe 5 minutes (I was already a little tired) and then BAM I felt hyper. Every other time I tried it since then? Nada. I might as well be drinking water. As for taste... I love me some straight up spirits, so the alcoholy part didn't bother me. It reminded me a bit of NyQuil meets a bit of, I don't know..some sort of flavored vodka. I'm certainly drank worse things."

- *summary sentence:*

"Can't decide if I should dump the rest or just keep using it."

Review text are less focused and cannot be labeled using rating

The summary sentence more straightforwardly reflects the user's opinion than the review text

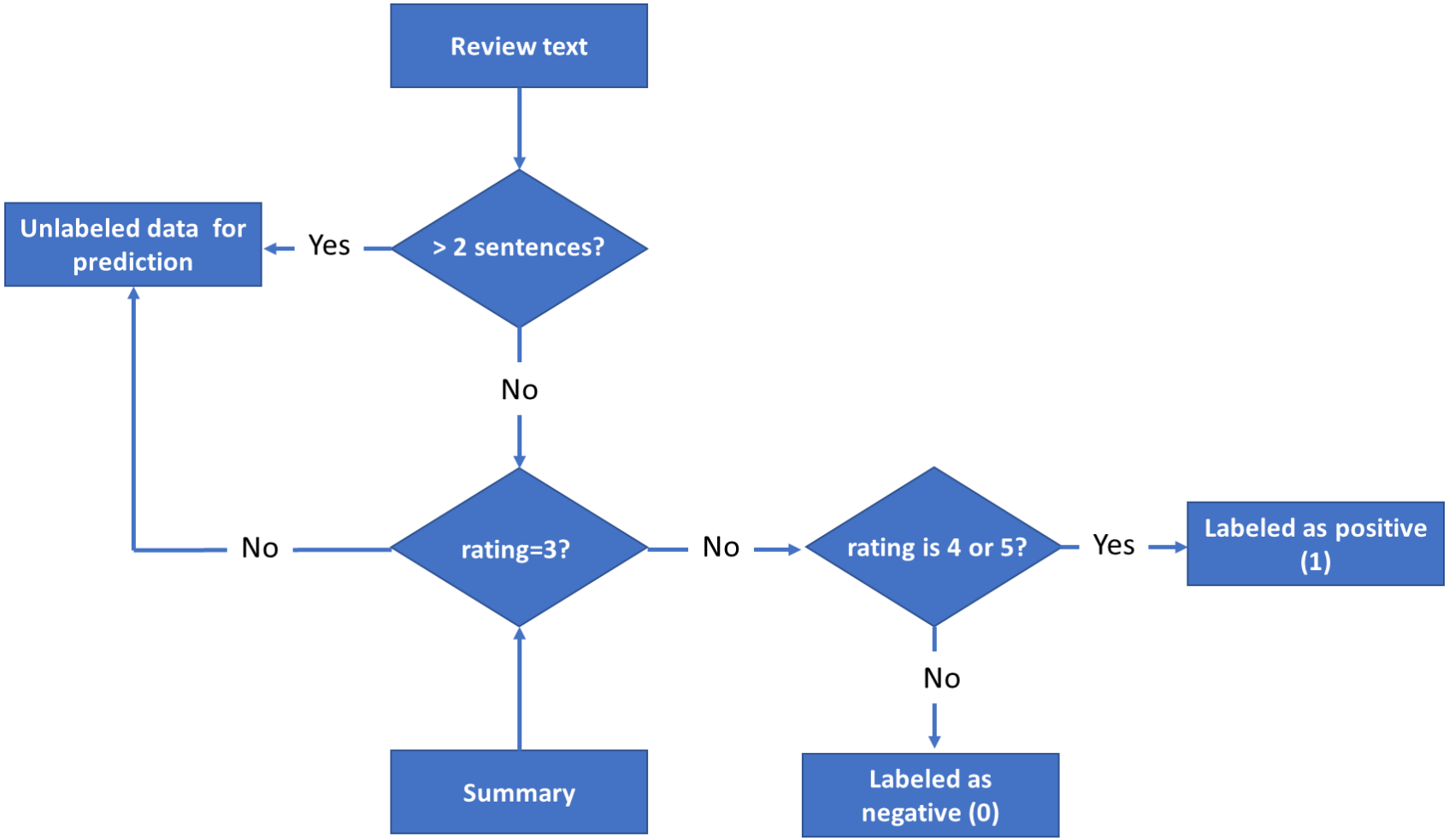
METHODS AND RESULTS

- Data cleaning and integration
- Exploratory data analysis
- Sentiment analysis to identify the positive opinions and negative opinions
- Extracting the relevant product features and representative sentences

SENTIMENT ANALYSIS

- **Selecting data for training, validation and testing**
- **Data preprocessing**
- **Test-train split**
- **Classification pipeline**
- **Evaluate tuned classifiers**
- **Custom cutoff probabilities for classification**

Selecting data for training, validation and testing



Data Preprocessing

- All letters changed to lower cases;
- Separating "not" as a single word. For example: don't -> do not, won't -> will not.
- Spelling correction. For example, amazzzzing -> amazing.
- Removing punctuations
- Lemmatization

Test-train split

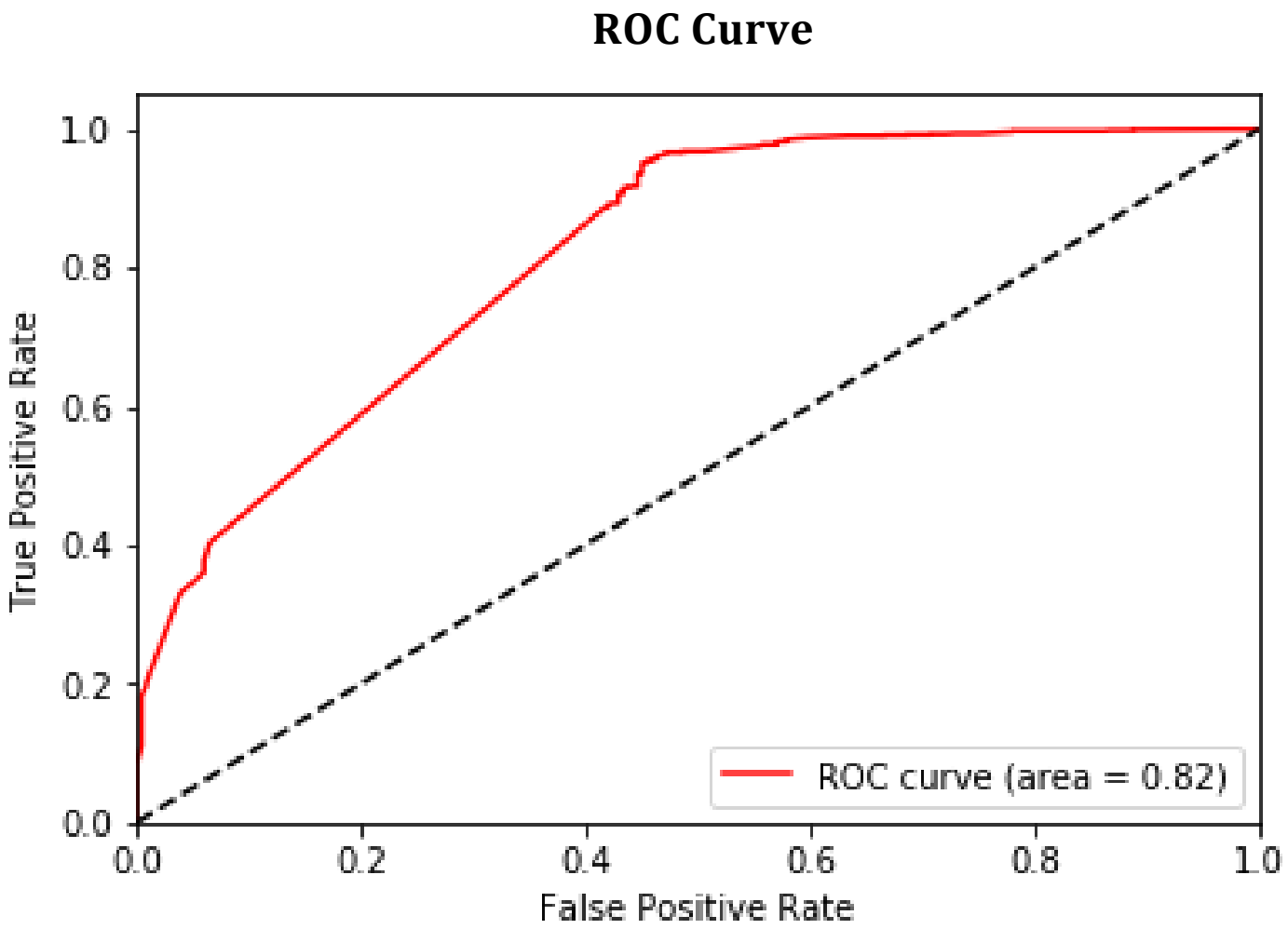
- 70% for train and cross validation and 30% data for testing

Classification pipeline

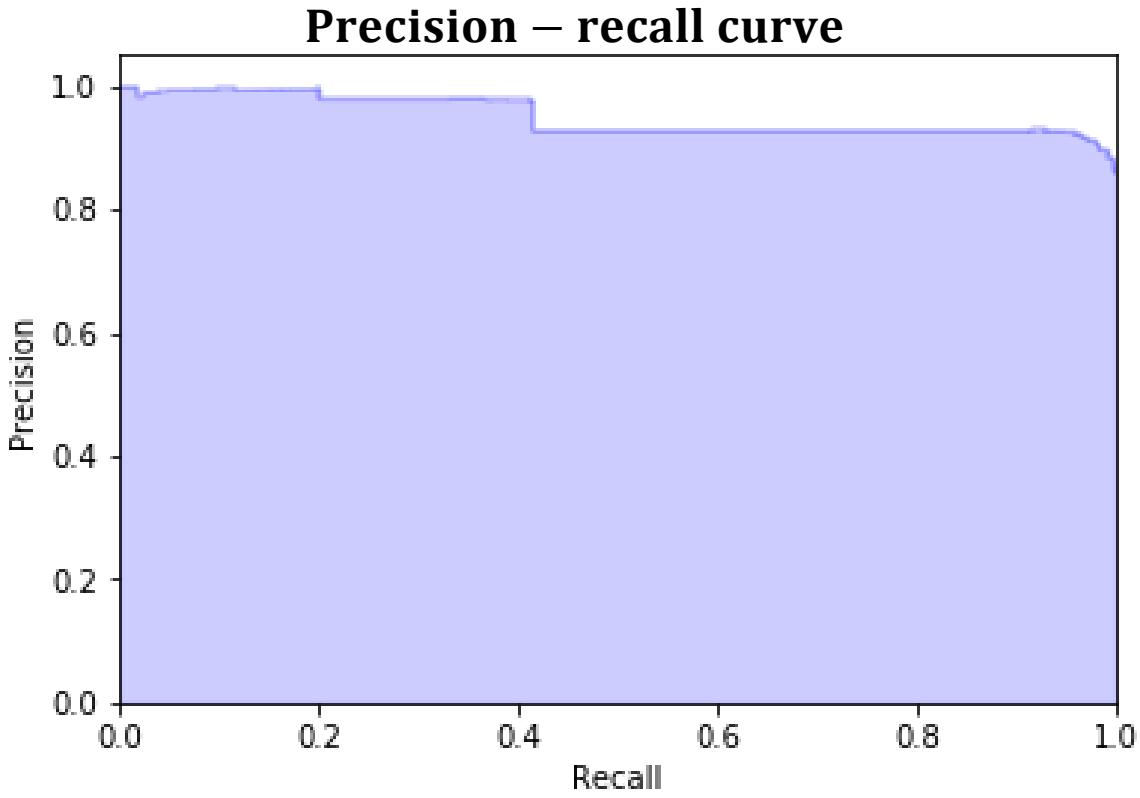
Tf-idf → Chi square feature selection → 5-fold cross validation

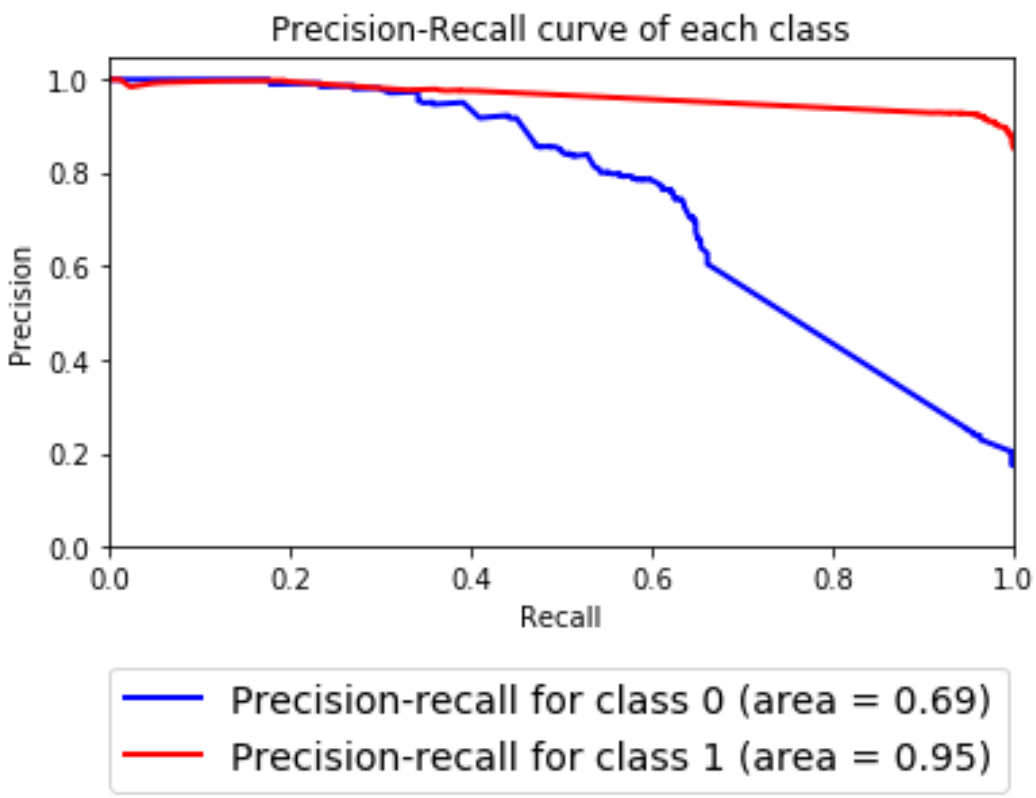
Stages of pipeline	Hyperparameters to tune	Tuning range
Tfidf	n grams	1-3: single word, bigrams, and trigrams 1-4: single word to four grams 1-5: single word to five grams
Feature selection	Number of features to keep	100, 300, all
Classification (5-fold cross validation)		
Naïve Bayes	Smooth factor (alpha)	0.01, 0.1, 1, 10
Random Forest	Number of random trees used to aggregate the predictions	
	Maximum number of features used to grow a tree	20, log2(N), (N)^0.5 N= total number of features
	Maximum number of splits in a tree	8, 10, 40
SVM	Penalty coefficient of the error term	0.01, 0.1, 1, 10
	Kernel coefficient (gamma)	0.01, 0.1, 1, 10
	Kernel	'rbf', 'poly'
Gradient Boosting	Number of trees to assemble	50, 100, 500
	Max number of features used to grow a tree	20, log2(N), (N)^0.5 N= total number of features
	Minimum number of leaves on a tree	3, 10, 20
	Minimum number of split	3, 5, 10

Evaluate tuned classifier



Evaluate tuned classifier





Custom cutoff probabilities for classification

- Unbalanced classes
- Neutral opinions are of less interest
- Choose cutoff probabilities separately for

positive opinions and negative opinions: optimized on F_β

$$F_\beta = (1+\beta)\frac{precision \cdot recall}{\beta \cdot precision + recall}$$

- Cutoff probabilities found for negative opinions is <0.42 and >0.91 for positive opinions

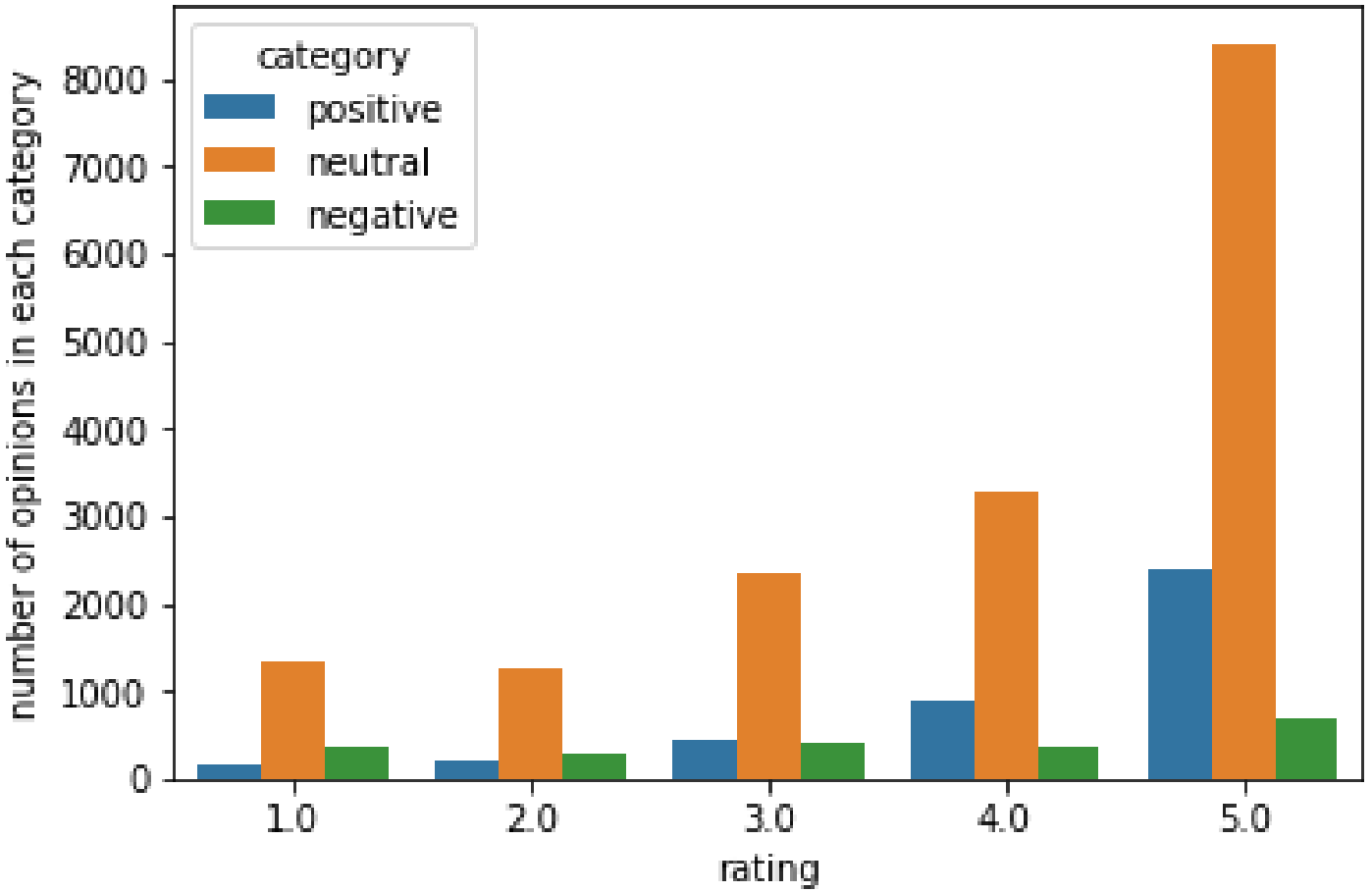
METHODS AND RESULTS

- Data cleaning and integration
- Exploratory data analysis
- Sentiment analysis to identify the positive opinions and negative opinions
- Extracting the relevant product features and representative sentences

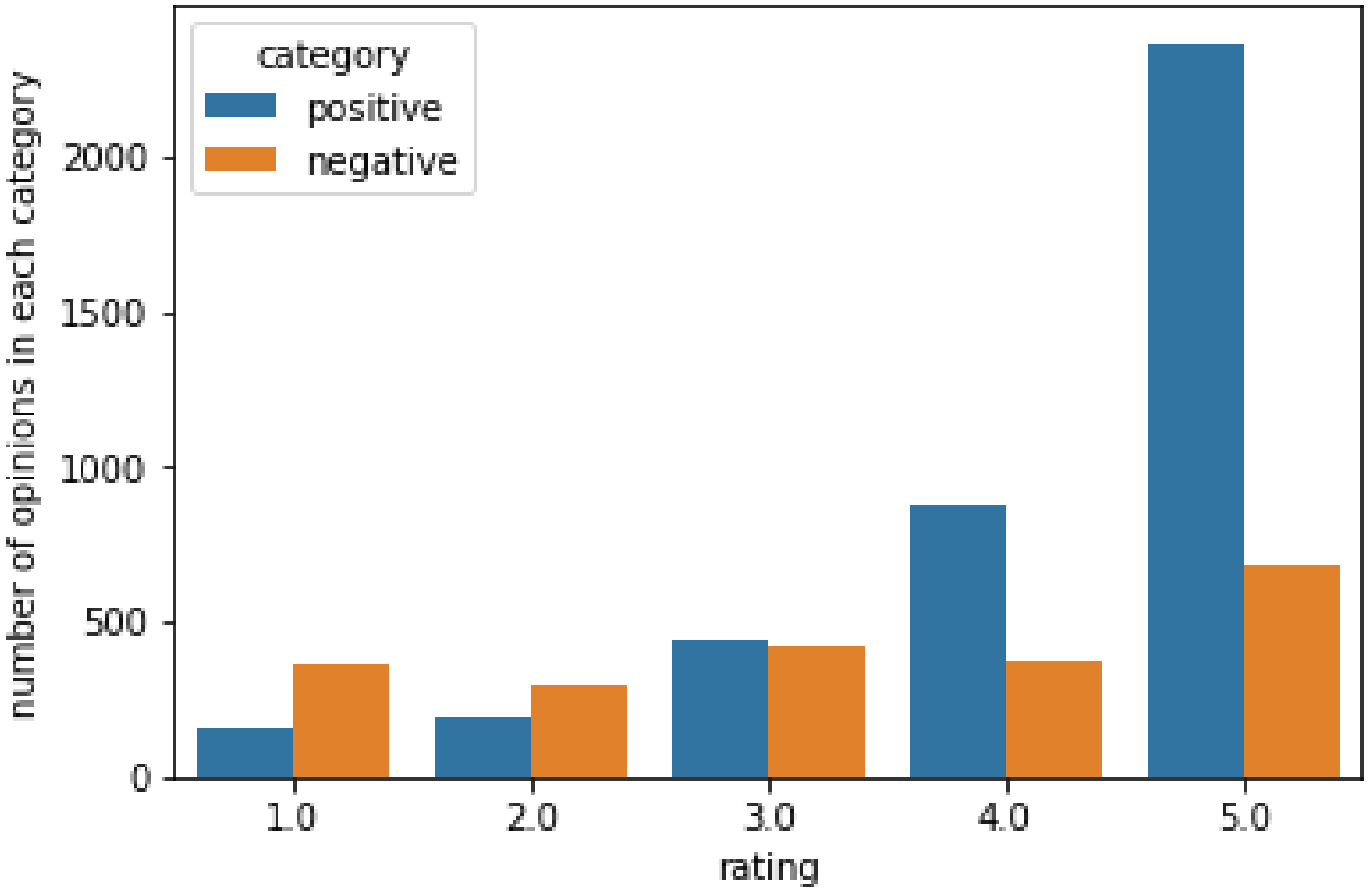
Extracting the relevant product features and representative sentences

- Classifying Unlabeled Data
- Extracting Meaningful Phrases
- Finding the Representative Sentences

Classifying unlabeled data

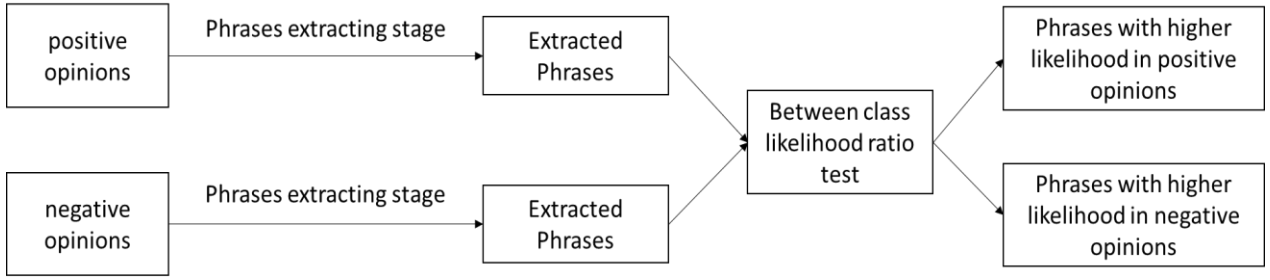


Classifying unlabeled data

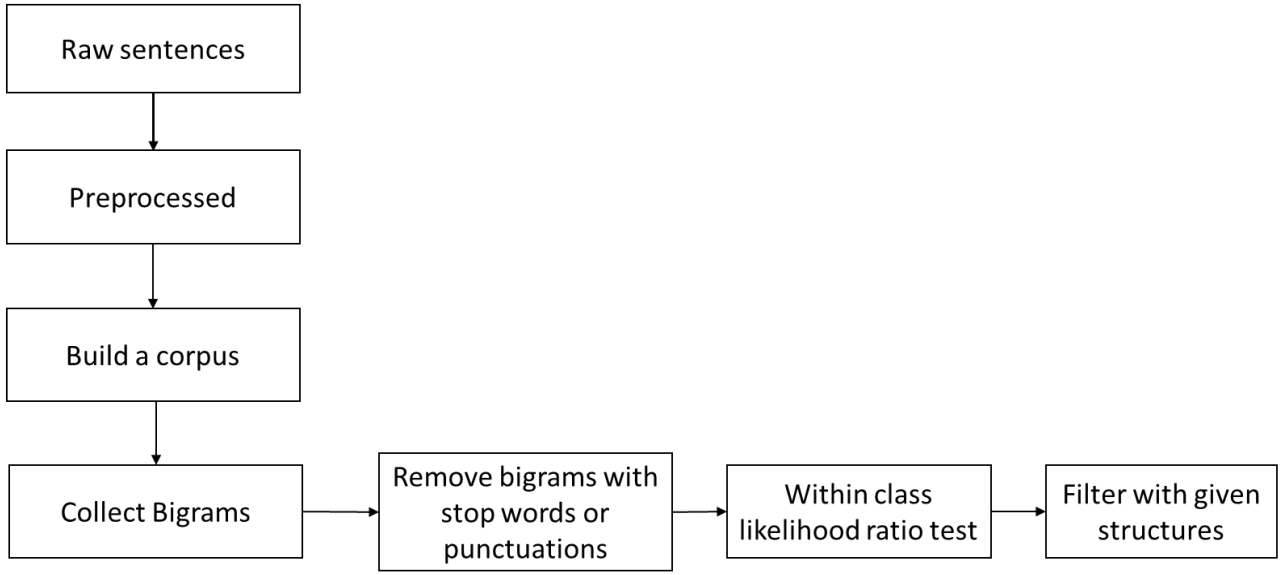


Extracting Meaningful Phrases

Overall work flow to extract phrases for two classes of opinions



Phrases extracting stage



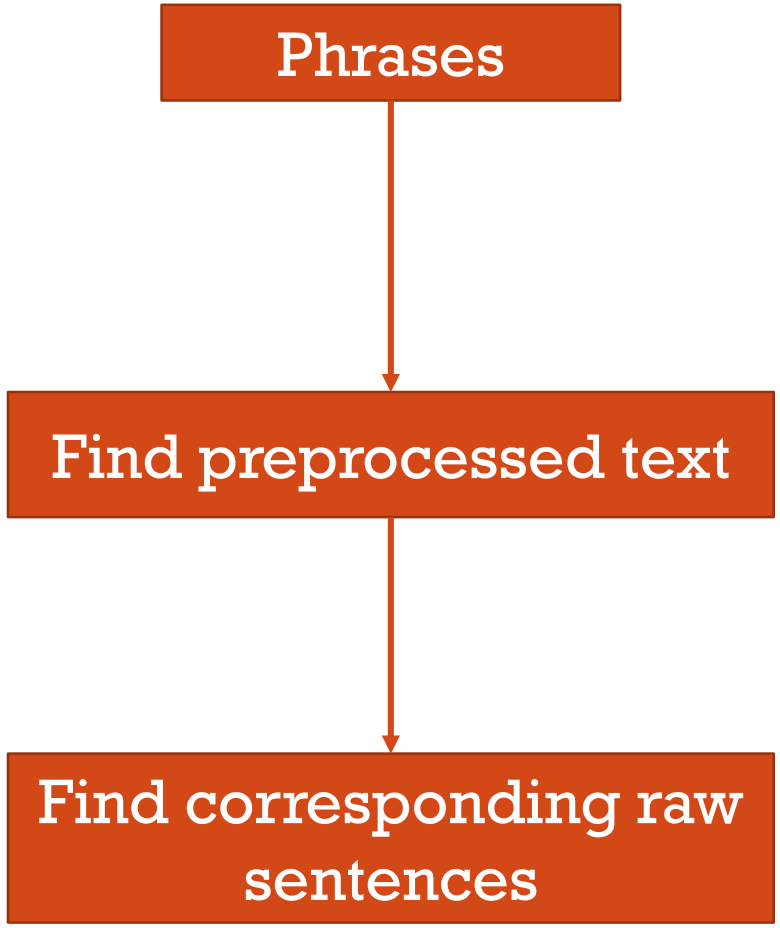
Top selected phrases for positive opinions

Bigrams	Likelihood ratio
sleep mask	inf
fall asleep	3.246981e+215
doe not	7.644850e+160
dream water	1.111200e+159
work well	4.158263e+150
great product	1.323021e+150
work great	2.313019e+133
good night	1.596020e+128
ear plug	5.910845e+121
jet lag	1.146892e+87
doxylamine succinate	1.157488e+69
year old	2.807142e+63
side effect	1.883415e+62
mg sleep	3.214381e+61
next day	4.774278e+60
go sleep	8.225597e+58
time release	5.060111e+58
also sleep	3.014924e+58
eye sleep	2.554948e+58
much better	2.036825e+58

Top selected phrases for negative opinions

Bigrams	Likelihood ratio
not work	inf
not help	4.594080e+70
anything not	1.769477e+67
not not	3.918912e+65
side effects	2.840317e+56
taste not	4.025552e+55
night not	5.350249e+54
find not	1.617184e+54
strap not	1.054834e+53
think not	3.958870e+52
not like	1.248547e+52
thing not	9.385651e+49
not know	5.681983e+48
product not	5.695425e+47
well not	2.455697e+47
sleep aid	2.166788e+45
not sure	1.087599e+41
not notice	1.420797e+37
not feel	1.164237e+33
trying not	1.318783e+30
not use	7.833462e+22
simply not	1.032731e+22
not stay	1.453739e+21
not recommend	7.177878e+20

Extracting Representative Sentences



e.g. doe not

e.g. 'my son doe not wake up at : a. m. anymore! yea! '

e.g. "My son doesn't wake up at 5:30 a.m. anymore! Yea! "

Examples of extracted opinions

Product "Biotab Nutraceuticals Alteril Sleep Aid with L-Tryptophan, Tablets 30 ea"

Positive opinions:

"Works Well for Our 8 Year Old"

Negative opinions:

"I've taken double and triple the recommended dosage and I don't feel a thing, not even a little sleepy. Don't waste your money on this one."

Product "Dream Zone- Earth Therapeutics Sleep Mask, 1ct"

Positive opinions:

"My son doesn't wake up at 5:30 a.m. anymore! Yea!..."

"works well..."

"When I take two tablets, they seem to work well. I'm able to get to sleep within 1/2 hour, and stay asleep for at least 4; many times, longer...."

Negative opinions:

"Strap not comfortable..."

"This product in no way made me sleepy, not to say that it won't work for you...."

"I hate to say this product did not work for me...."

"I'm not sure if it did or did not; I used for about 2-3 months then went back off...."

"That's okay on an rare basis, but would not work nightly."