Predicting User Ratings from Review Text

A Natural Language Processing & Machine Learning Case Study

Springboard Career Track Capstone 1
Heather A. Passmore

Problem

Customer ratings influence sales. Rating errors have big effects.

What companies care?

Problem

Customer ratings influence sales. Rating errors have big effects.

- On average customers read more than two Yelp reviews before deciding to use a business.
- A one-star increase in ratings on Yelp leads to a 5-9% increase in revenue for a business.

Rudolph 2015

User Reviews and Ratings

Ratings matter

Consumers & vendors depend on ratings for:

- Purchase decisions
- Predicting sales

High and low ratings affect sales directly.

Predict rating

Benefits of predicting ratings from text:

- Error catching
- Suggest rating
- Improve consistency
- Flag problematic ratings

Motivation

Build a system to identify positive and negative feedback from customers to give businesses the power to intercede and to improve customer engagement and satisfaction.

Top customer reviews

triple0triple1

★☆☆☆☆ **Get the first edition.**

August 29, 2013

Format: Hardcover | **Verified Purchase**

I have just bought this brand new 3rd edition and compared it with my 1st edition. They are almost exactly the same, even down to the page numbers. This is an old scam by textbook publishers, but usually they will at least change around the table of contents and put in new diagrams, photos and tables. Not so with this title.

And if the 1st and 3rd editions are almost exact copies, then I'm sure the 2nd edition would be fine too.

Also, do not think that you need to buy the "Custom Edition" for your school, it's just the campus bookstore's and publisher's way of keeping you off Amazon and charging you double for an already overpriced book.

Comment | 181 people found this helpful. Was this review helpful to you? Yes No Report abuse

J.T884

★☆☆☆☆ **Beware! Ebook has false advertising about including Mastering Chemistry**

November 7, 2016

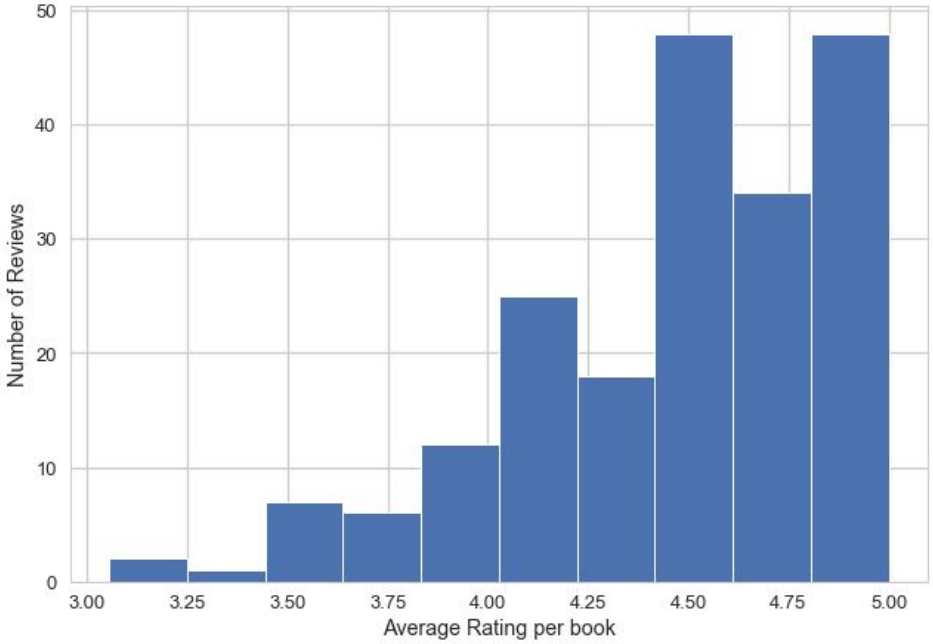
Format: Kindle Edition | **Verified Purchase**

Chemistry: A Molecular Approach

Even though it says twice in the description that the ebook comes with Mastering Chemistry, and it is the first result with an ISBN search that includes Mastering Chemistry, it actually doesn't contain Mastering Chemistry. False adverting at its finest. Chatted with support and all they would do is refund my money so I can order a hardcover. I can't go without a text while in a class. Very bad customer service.

Comment | 7 people found this helpful. Was this review helpful to you? Yes No Report abuse

Reviewers love books! (most ratings are ‘high’)



- Majority of books reviewed by more than 10 reviewers have mean ratings > 4 stars.
- This Science Textbook genre subset of reviews includes reviews for 729 different books in 11546 reviews.
- Binary categories: ‘low' ratings (n = 3884) and ‘high’ ratings (n = 7662).
- Word counts are significantly higher for reviews with ‘low’ ratings.

Machine Learning Highlights

Preprocessing

Specialized stopwords

Remove ‘english’ stopwords except:

- Leave ‘not’ in corpus for bigrams
- Remove ‘book’, books’

Vectorization

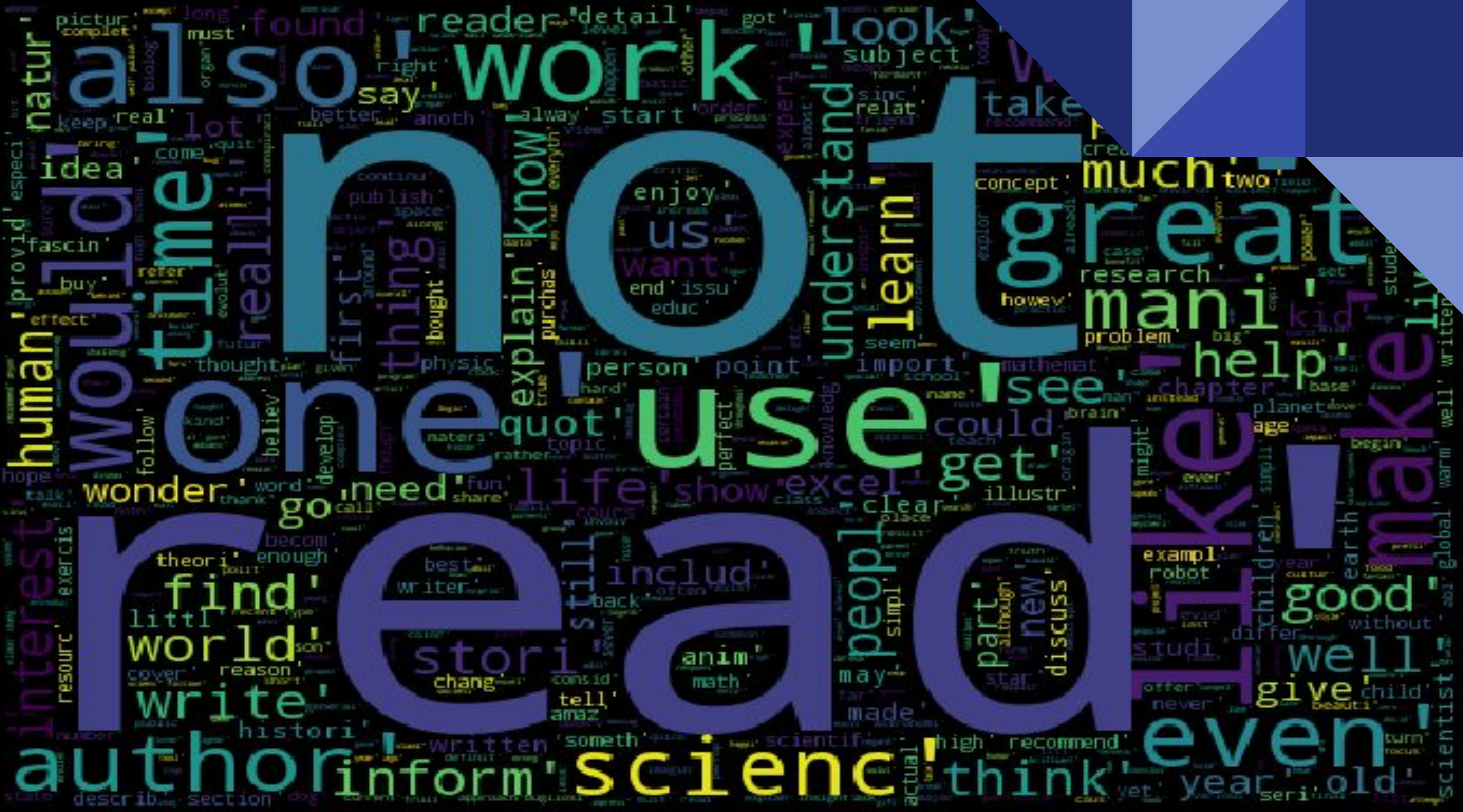
TF-IDF Vectorizer

- Features from single words and bigrams
- Baseline ROC-AUC score = 0.78 with Multinomial Naive Bayes classifier.

Model Tuning

Important Parameters

- Set class_weight to ‘balanced’ since target classes imbalanced.
- Tune with GridSearchCV.
- Compare ROC-AUC scores.


not
read
like
make
great
use
one
would
also
work
time
realli
author
scienc
even
mani
good
well
look
find
world
write
stori
inform
think
know
understand
learn
help
see
get
could
life
show
excel
need
wonder
human
interest
peopl
includ
still
year
old
give
take
much
two
research
kid
chapter
quot
explain
thing
first
us
want
enjoy
found
reader
detail
subject
idea
natur
go
clear
new
discuss
part
anim
math
chang
theori
littl
histori
written
someth
recommend
high
children
studi
child
earth
scientist

Classifier Comparisons

Classifier:	**ROC-AUC:**	**GridSearchCV best params:**
MultinomialNB	0.78	alpha=1, fit_prior=True
RandomForestClassifier	0.75	max_features=750, min_samples_leaf=6
LogisticRegressionCV	0.81	Cs=10, max_iter=100, tol=0.0001
SGDClassifier	0.75	alpha=0.1

Best Classifier

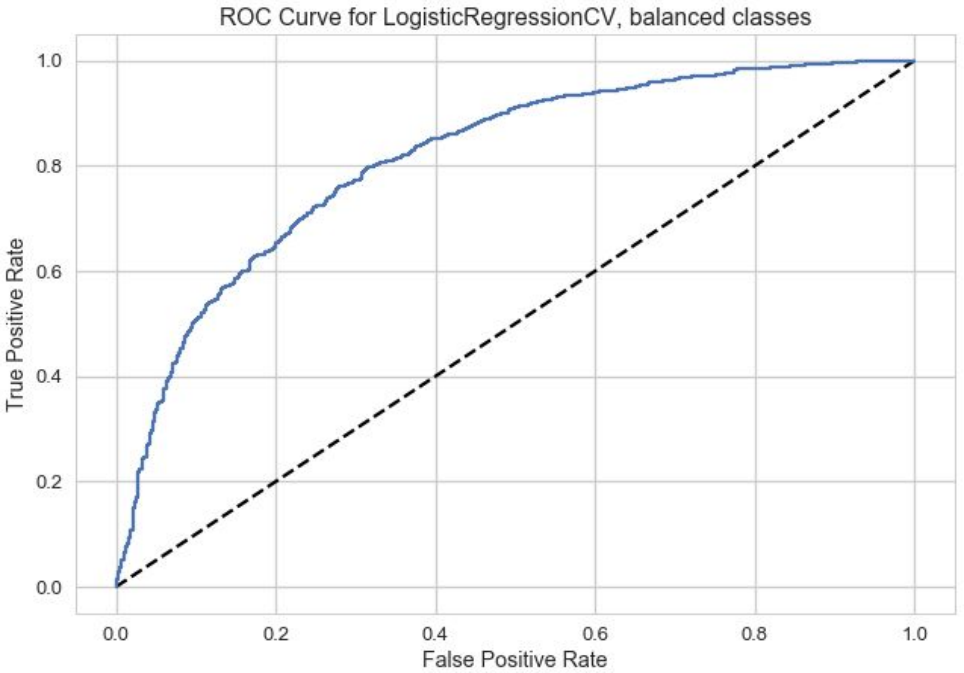
Logistic Regression with Cross Validation

y=1 top features

Weight[?]	Feature
+5.262	excel
+4.659	wonder
+4.364	love
+4.188	great
+3.765	perfect
+3.193	everi
+3.129	conspiraci
+3.125	high
+3.076	high recommend
… 19630 more positive …	
… 17289 more negative …	
-3.064	lack
-3.121	page
-3.123	ok
-3.196	bore
-3.240	good
-3.807	unfortun
-3.847	would
-3.860	seem
-3.992	howev
-5.268	disappoint
-5.320	not

Best Classifier

Logistic Regression with Cross Validation



Improve classification

Goal: Maximize Precision

Precision

Precision (TP/ (TN + FP))

- how often is the classifier correct when a positive value is predicted?
- how "precise" is the classifier when predicting positive instances?

False positives

‘High’ rated reviews are common but ‘low’ ratings are rare.

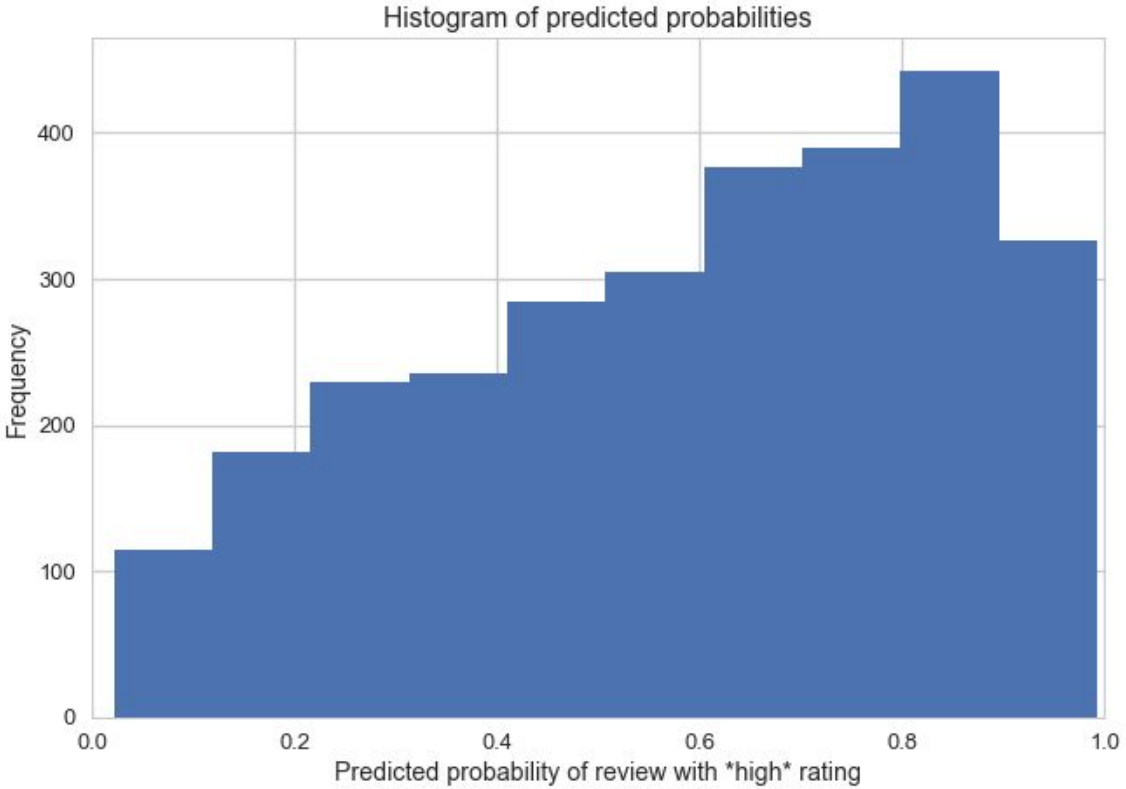
- Minimize number of FP classifications.
- Aim to classify more actual ‘low’ reviews correctly.

Motivation

Customers and retailers benefit from accurate information in review classification.

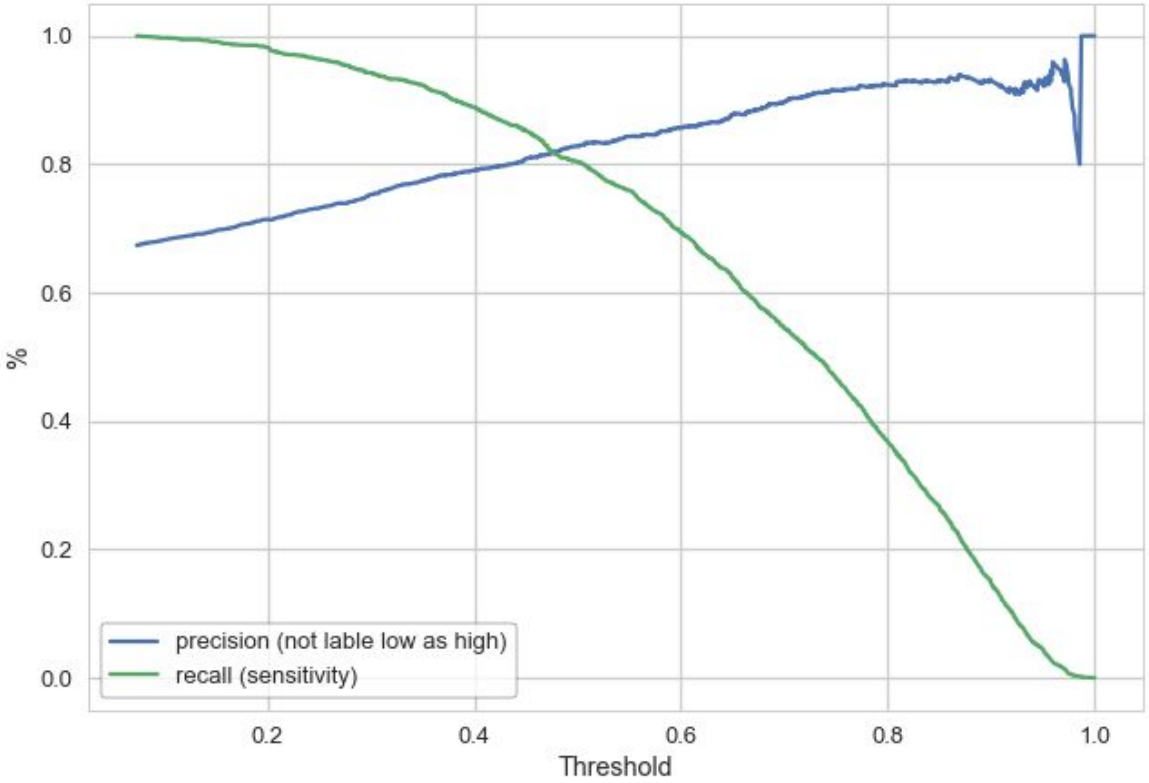
Ex: buying a positively reviewed product instead of an inaccurately classified negatively reviewed product.

Distribution of predicted probabilities



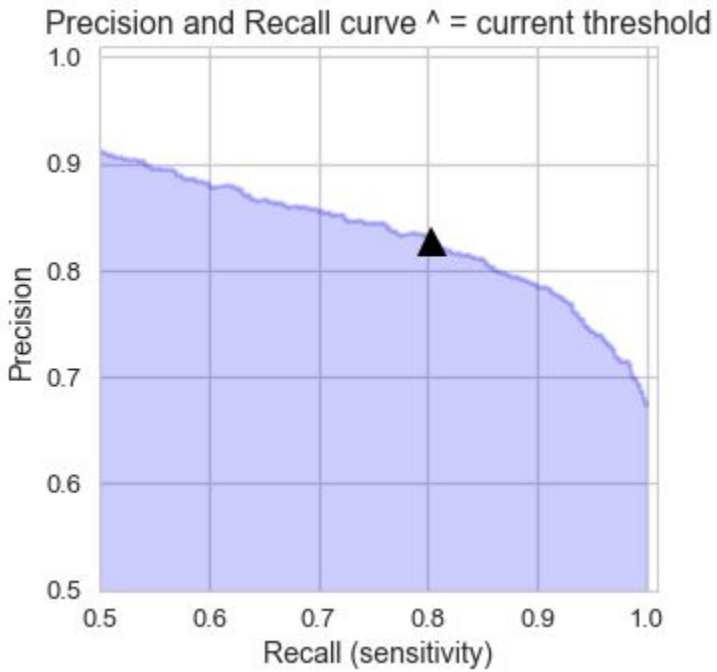
- Review dataset is heavily skewed towards positive ratings.
- Despite efforts to balance classes in the model, the majority of observations have high probability of a ‘high’ rating.

Precision/Recall Tradeoff



- Depending on decision threshold above or below the default
- Low threshold increases sensitivity of classifier (recall)
- High threshold increases precision (accurately labeling ‘low’ reviews)

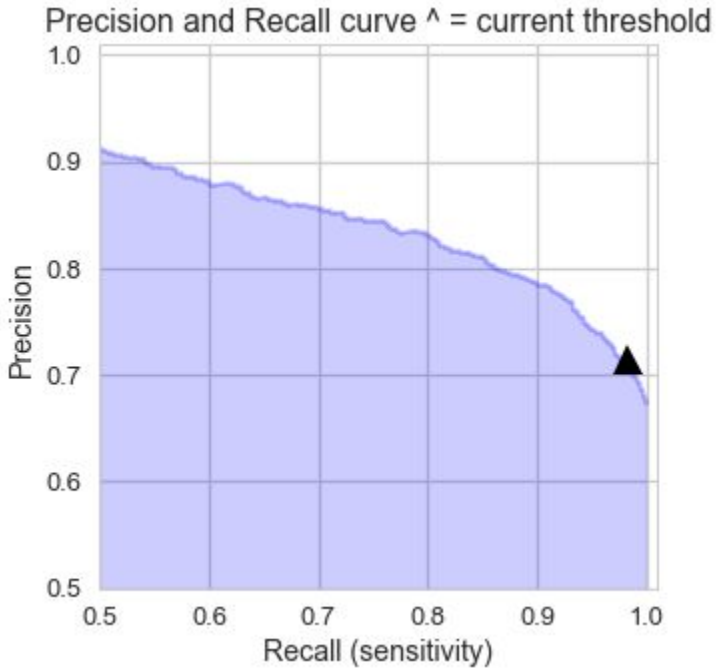
Precision/Recall Tradeoff



n = 2887:	Predicted 'low':	Predicted 'high'
Actual ‘low’	TN = 651	FP = 320
Actual ‘high’	FN = 376	TP = 1540

Threshold = 0.50

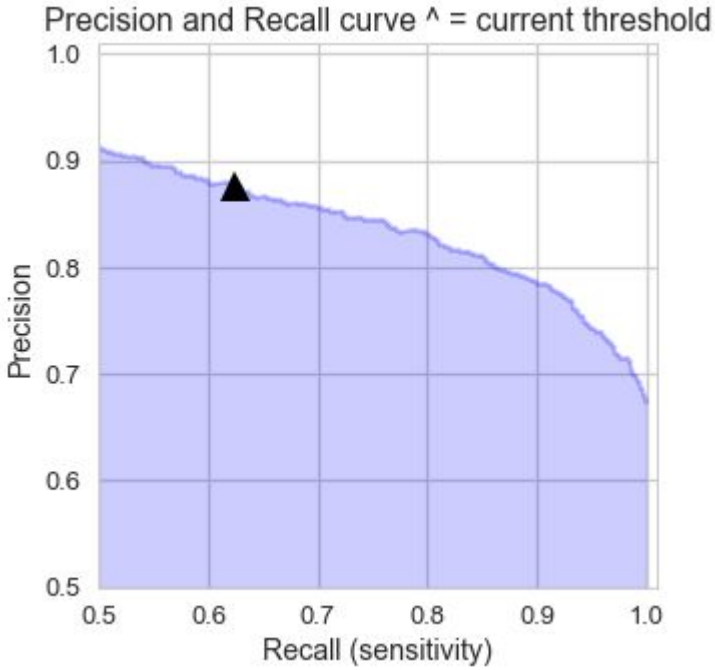
Precision/Recall Tradeoff



n = 2887:	**Predicted 'low':**	**Predicted 'high'**
Actual ‘low’	TN = 219	FP = 752
Actual ‘high’	FN = 34	TP = 1882

Threshold = 0.20,

Precision/Recall Tradeoff

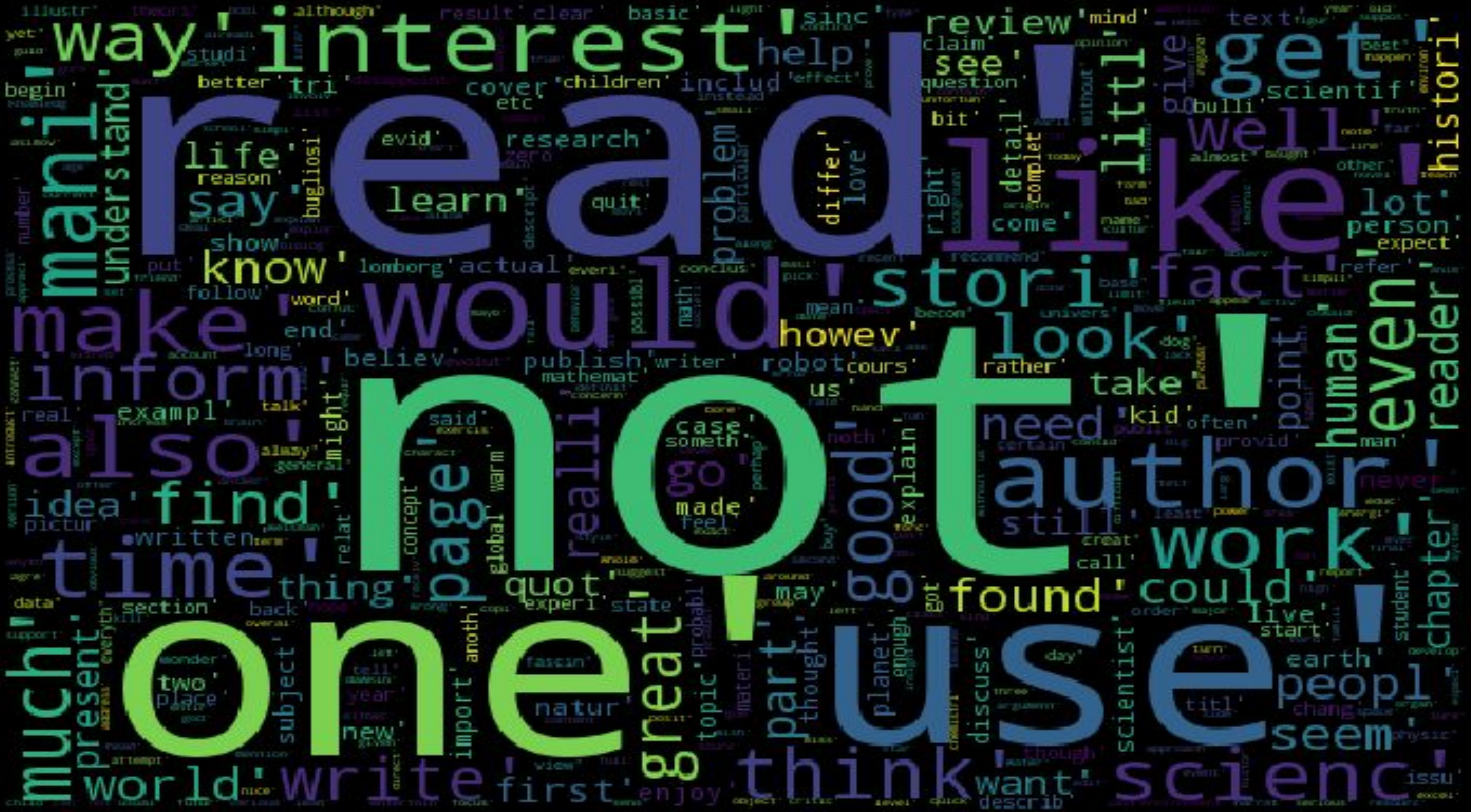


Threshold = 0.65

n = 2887:	**Predicted 'low':**	**Predicted 'high'**
Actual 'low'	TN = 804	FP = 167
Actual 'high'	FN = 723	TP = 1193

Decision Threshold Adjustment:

1. False positives reduced from 320 to 167
2. Precision increased


read
not
one
use
like
would
interest
get
way
author
make
stori
fact
also
inform
time
work
think
scienc
write
world
much
great
good
look
even
reader
human
histori
well
littl
give
mani
understand
page
realli
could
found
part
peopl
seem
need
find
idea
first
want
take
know
say
life
learn
help
see
review
scientif
problem
howev
lot
person
chapter
earth
point
research
show
exampl
publish
believ
robot
thing
quot
present
present
studi
tri
better
begin
cover
children
includ
question
claim
text
mind
basic
clear
result
sinc
although
evid
reason
differ
love
detail
complet
come
right
bugliosi
lomborg
actual
follow
word
end
long
mathemat
writer
cours
rather
kid
often
provid
us
case
someth
made
still
feel
written
pictur
concept
global
warm
relat
back
state
section
experi
may
call
creat
live
start
student
new
two
place
subject
import
natur
topic
probabl
materi
thought
planet
enough
discuss
scientist
titl
chang
enjoy
describ
issu
quit
bit
bulli
expect
refer
everi
conclus
mean
man
never
said
might
go
go
explain
almost
other
alway
general
talk
real
number
put
zero
etc
data
fascin
anoth
got
day
though