Image Help From Yelp

Insights into Yelp’s Open Dataset

Feb 2020

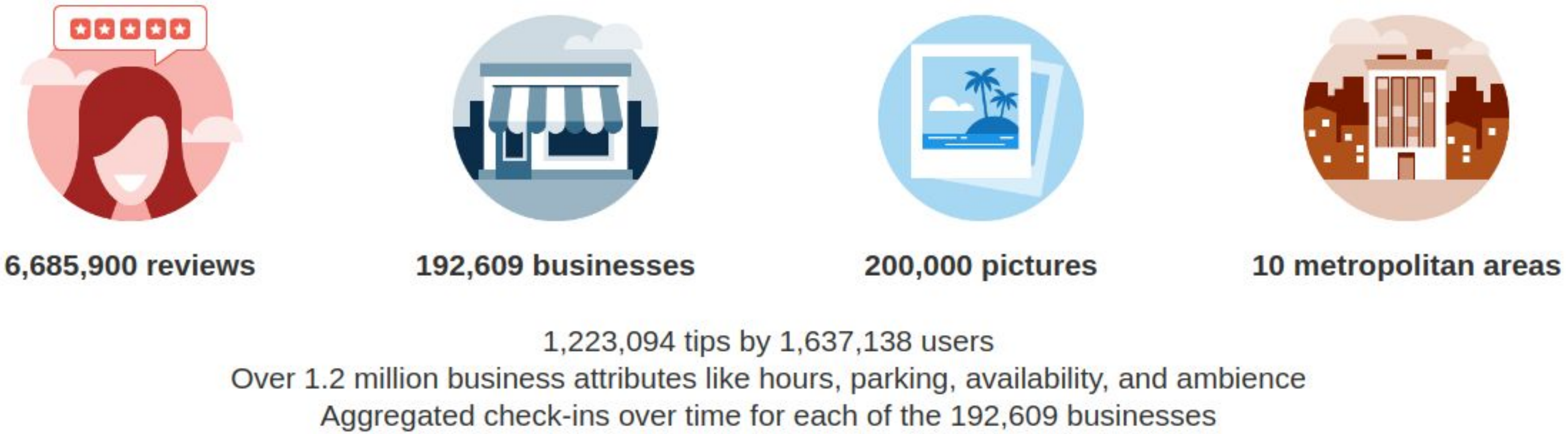
David Yu

Github.com/yuchild/image_help_from_yelp

Case Study Questions

1. Can written reviews ***predict*** ratings?
2. Can the **photos** taken be ***classified***?
3. Can **photos** taken help us ***rate*** an establishment?

Data Source: Yelp Open Dataset



6,685,900 reviews

192,609 businesses

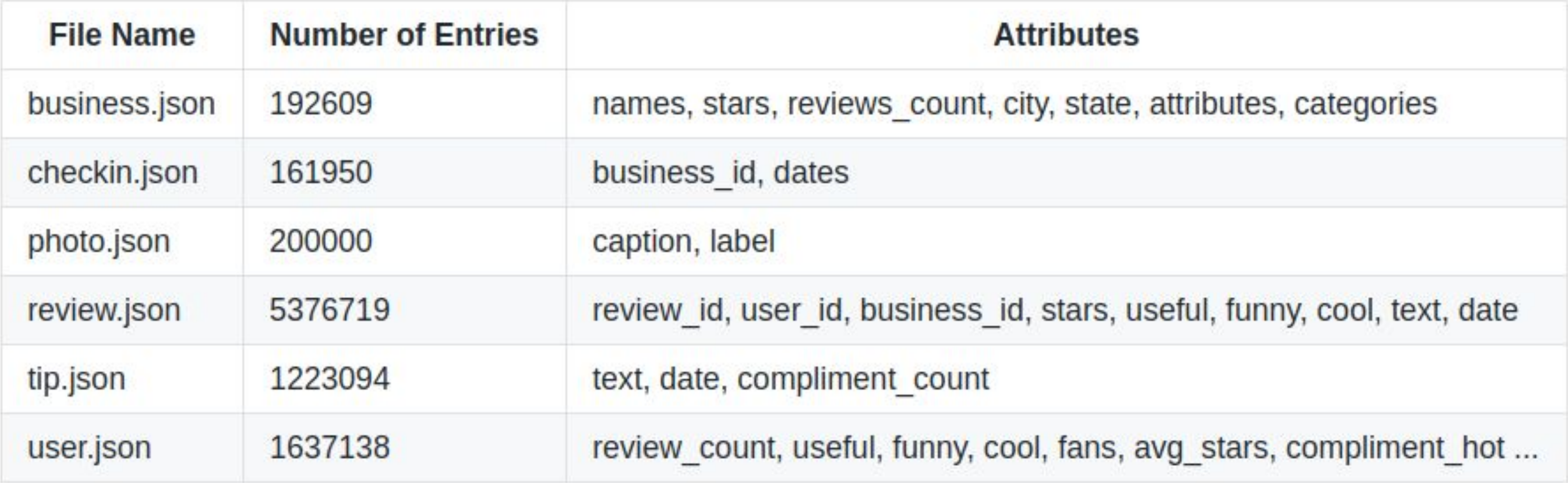
200,000 pictures

10 metropolitan areas

1,223,094 tips by 1,637,138 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 192,609 businesses

Source: yelp.com/dataset

Summary Yelp Open Dataset Used:

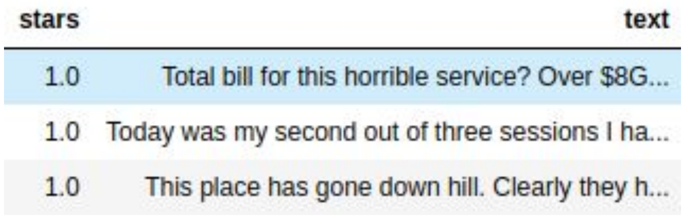


	File Name	Number of Entries	Attributes
★	business.json	192609	names, stars, reviews_count, city, state, attributes, categories
	checkin.json	161950	business_id, dates
★	photo.json	200000	caption, label
★	review.json	5376719	review_id, user_id, business_id, stars, useful, funny, cool, text, date
	tip.json	1223094	text, date, compliment_count
	user.json	1637138	review_count, useful, funny, cool, fans, avg_stars, compliment_hot ...

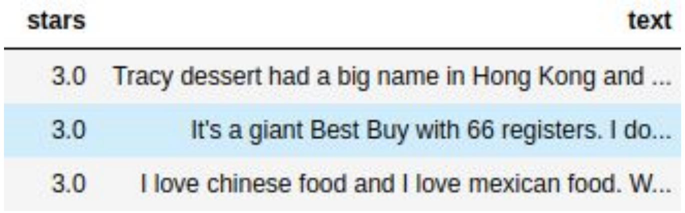
Quick Overview

1. **Written** reviews predicts ratings with **85%** accuracy
2. Classification of **images** was problematic with **SVM** with **45%** accuracy
3. Photos classify business ratings with **62%** accuracy using CNN

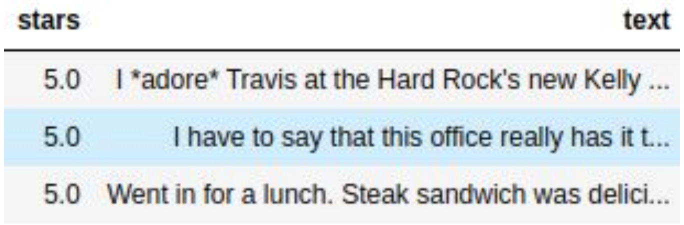
Reviews



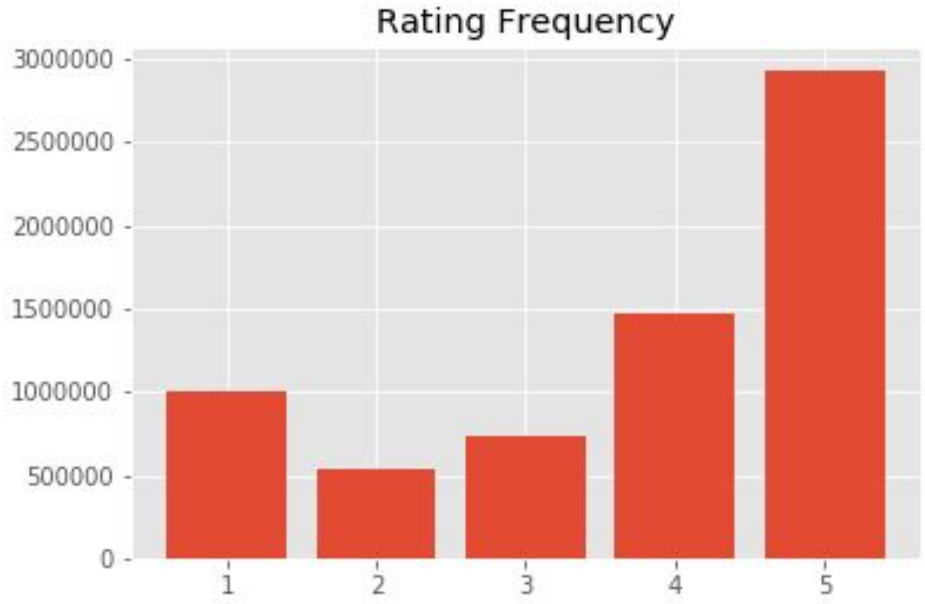
stars	text
1.0	Total bill for this horrible service? Over $8G...
1.0	Today was my second out of three sessions I ha...
1.0	This place has gone down hill. Clearly they h...



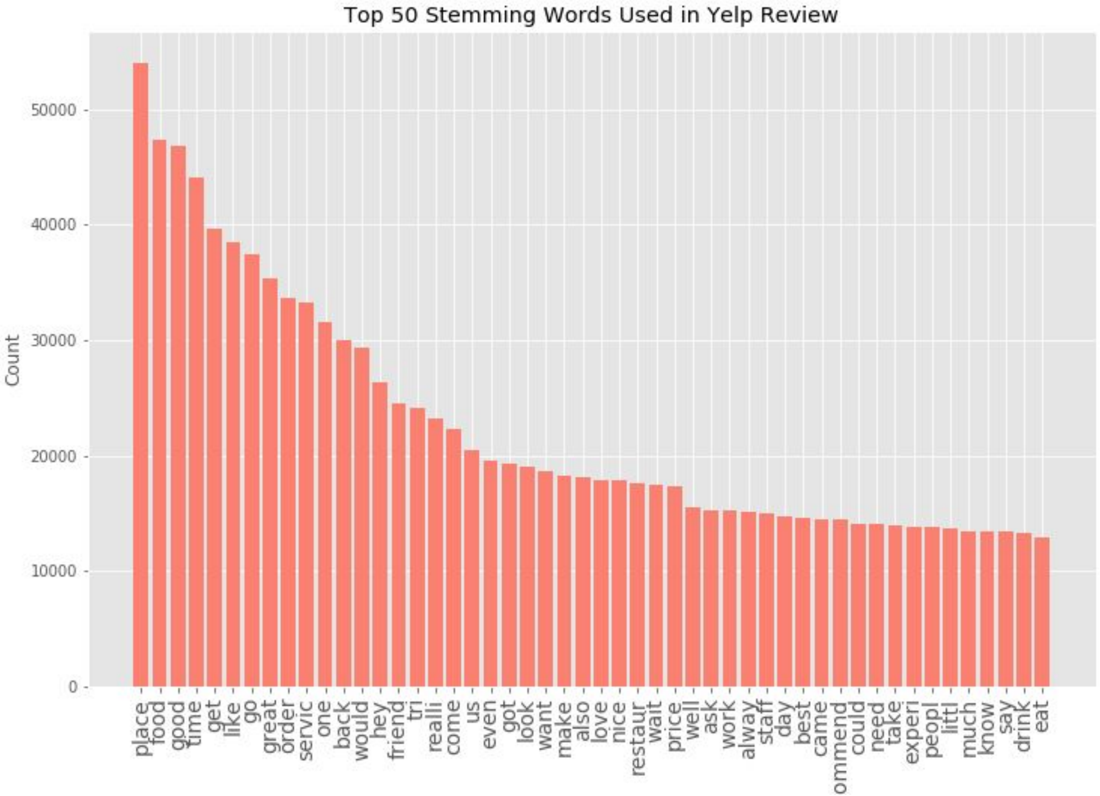
stars	text
3.0	Tracy dessert had a big name in Hong Kong and ...
3.0	It's a giant Best Buy with 66 registers. I do...
3.0	I love chinese food and I love mexican food. W...



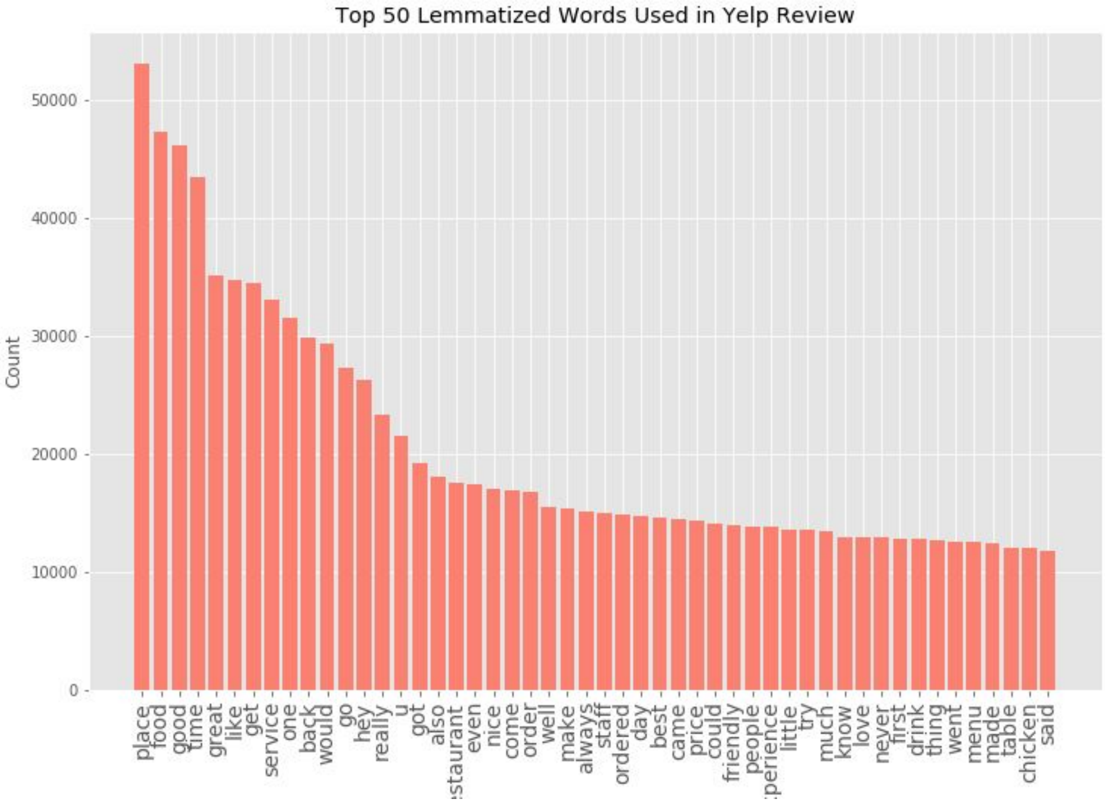
stars	text
5.0	I *adore* Travis at the Hard Rock's new Kelly ...
5.0	I have to say that this office really has it t...
5.0	Went in for a lunch. Steak sandwich was delici...



NLP Top Stemming Words

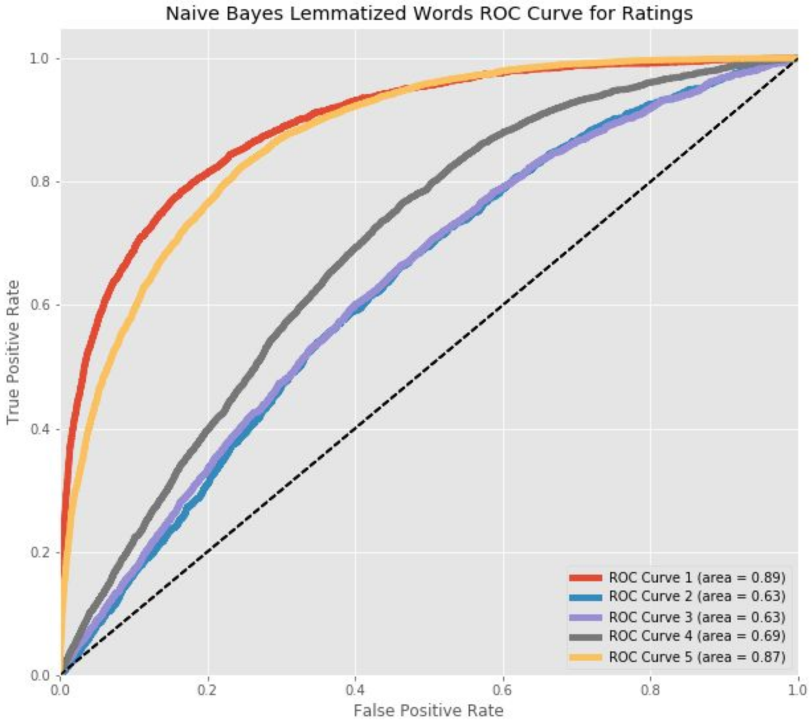


NLP Top Lemmatized Words

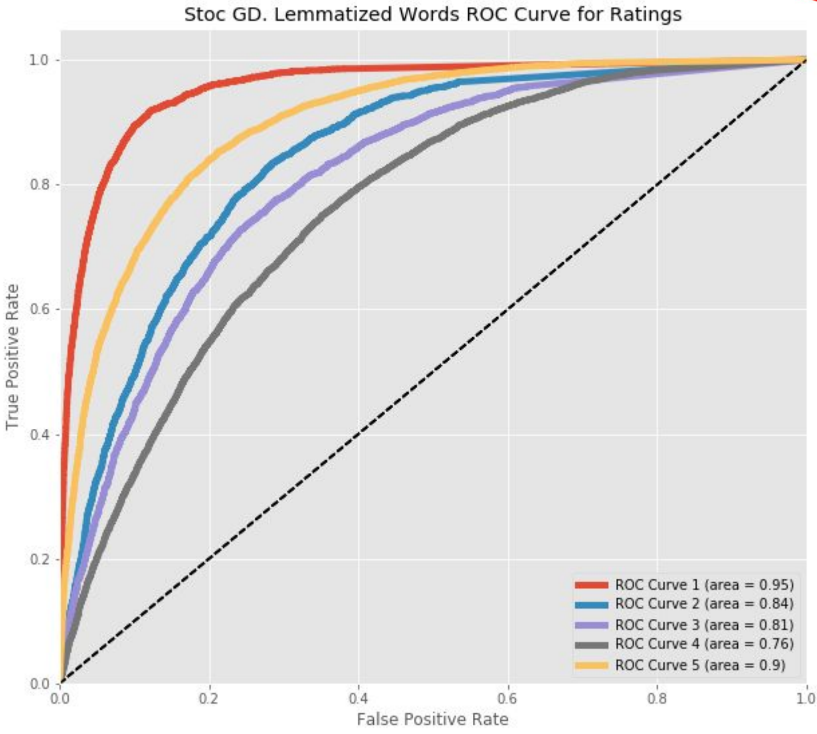


want
first time
amaz
ut
check
go back
expect
store
go
top
well
hat
actual
person
hey
busi
said
tri
left
peopl
nd
lot
see
sinc
make
help
went
perfect
ask
room
use
custom
servic
love
eat
serv
made
nice
one
delici
place
order
way
even
best
drink
need
much
visit
found
restaur
local
come
back
day
got
say
start
show
think
look
told
work
area
still
enjoy
know
flavor

Clash of the Models: **NB** vs **SVM** (SGD)

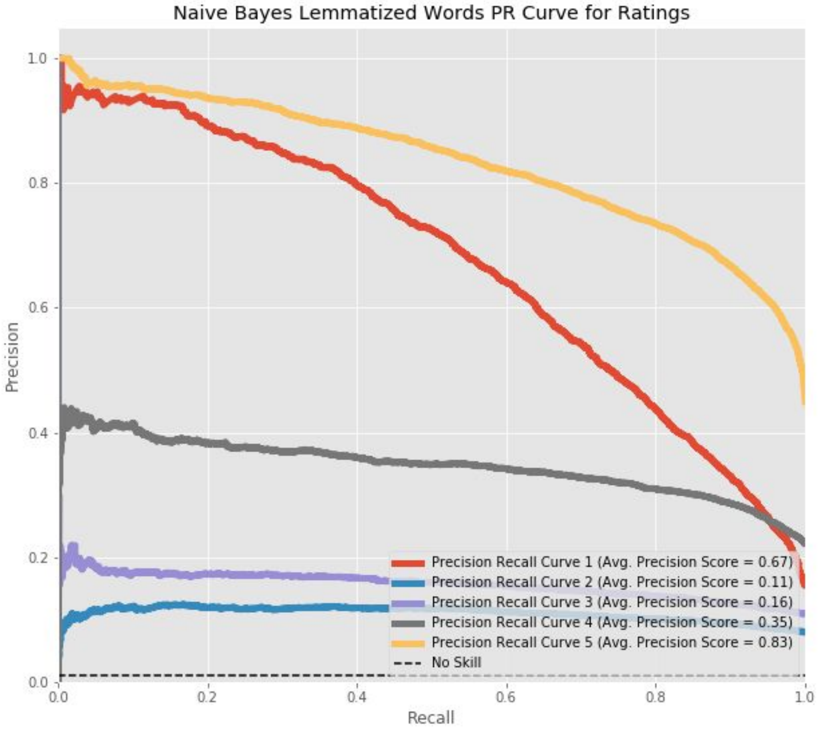


Naive Bayes Lemmatized Words ROC Curve for Ratings

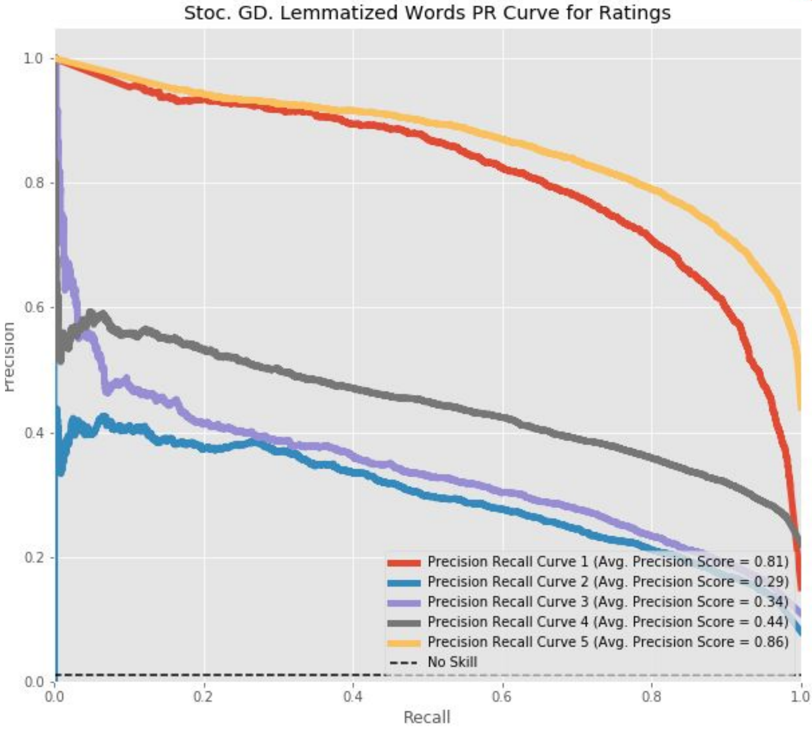


Stoc GD. Lemmatized Words ROC Curve for Ratings

Clash of the Models: **NB** vs **SVM** (SGD)



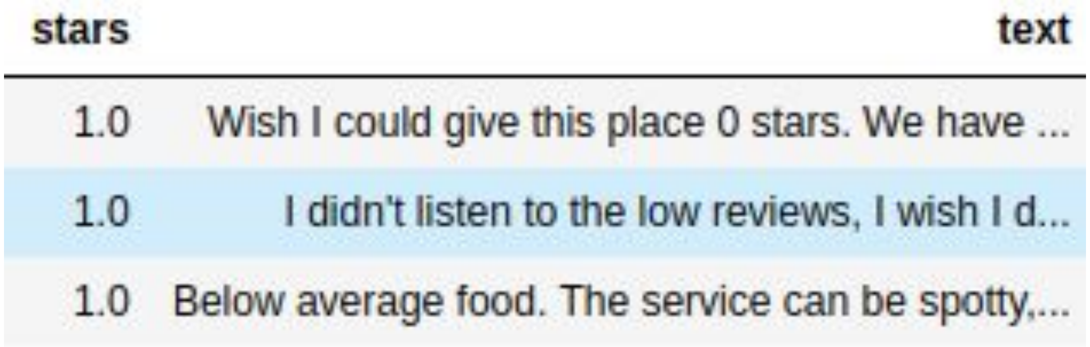
Naive Bayes Lemmatized Words PR Curve for Ratings



Stoc. GD. Lemmatized Words PR Curve for Ratings

Text Ratings Takeaways:

1. **Written** reviews prediction with **85%** accuracy
2. Model is **biased** towards the two ends of the scale: **1** and **5** stars
3. Ratings **2**, **3**, and **4** stars poorly classified even with ***balanced*** training set
4. Only extreme words will trigger a poor rating from the model



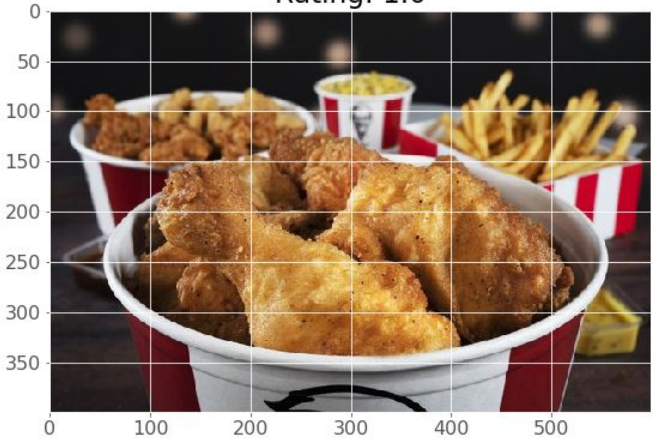
stars	text
1.0	Wish I could give this place 0 stars. We have ...
1.0	I didn't listen to the low reviews, I wish I d...
1.0	Below average food. The service can be spotty,...

Business Rating from Photos

Can a picture tell you ***anything*** about the establishment?

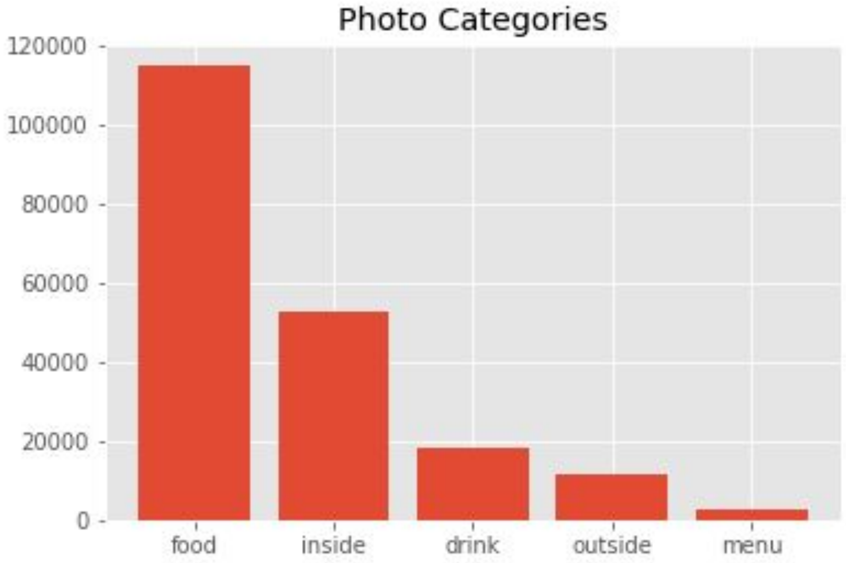
Rating: 4.0

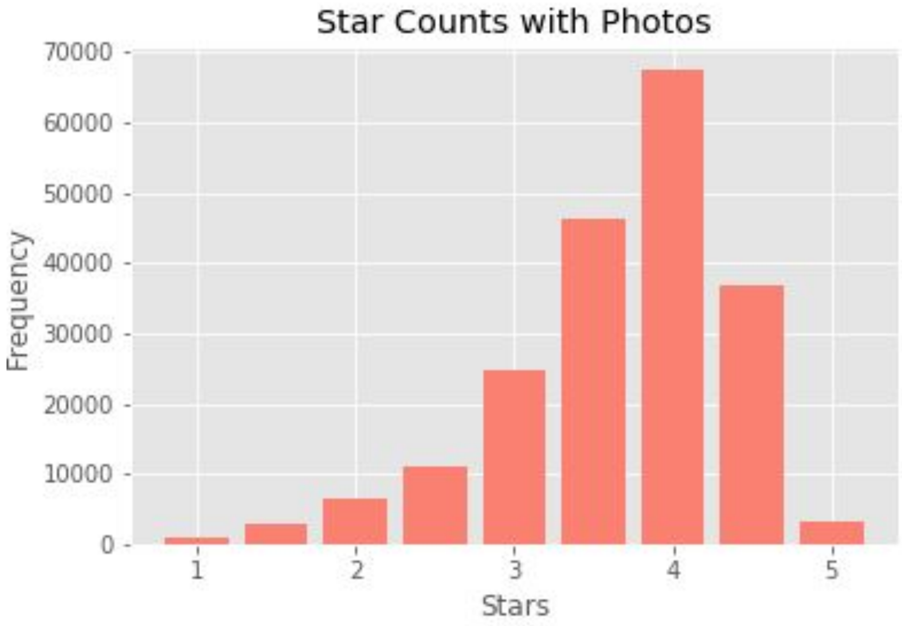
Rating: 5.0



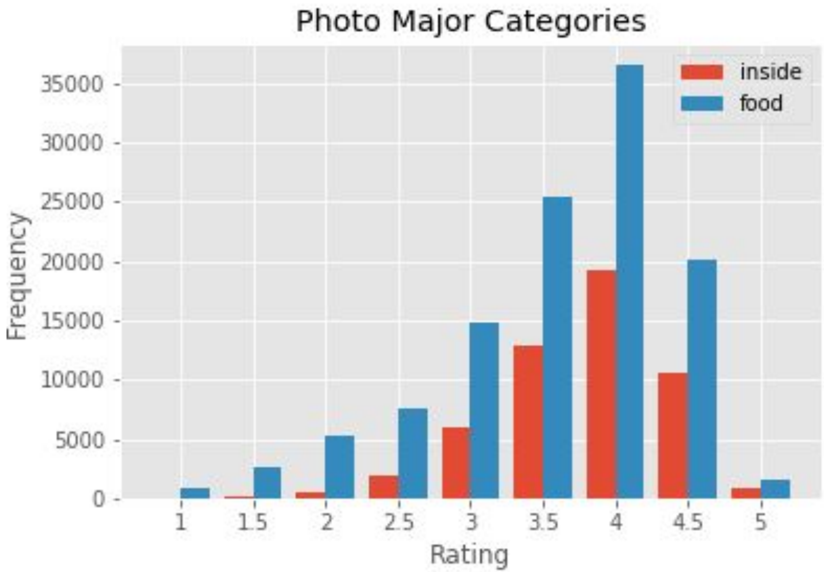
Rating: 1.0

Photo Classification





Photos of What?



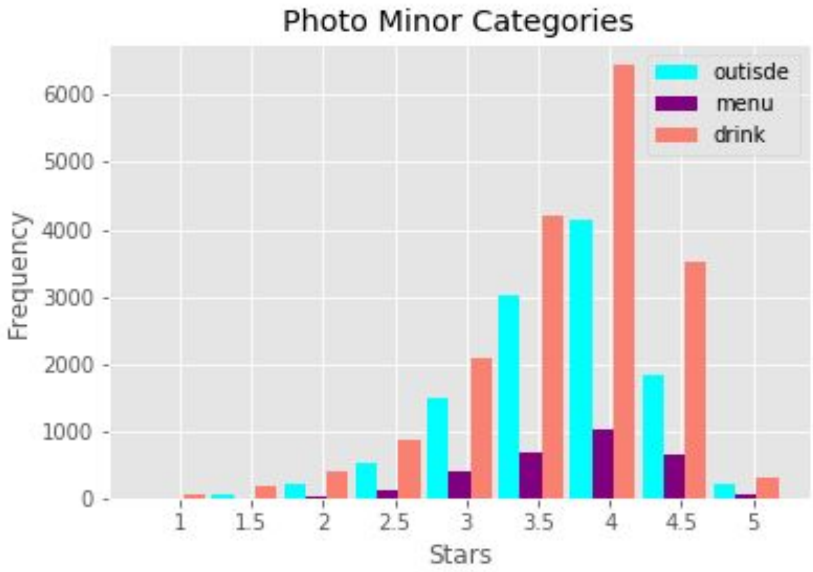
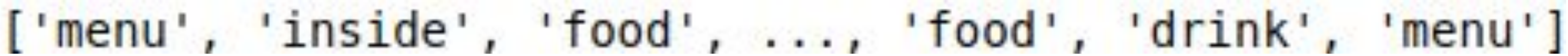
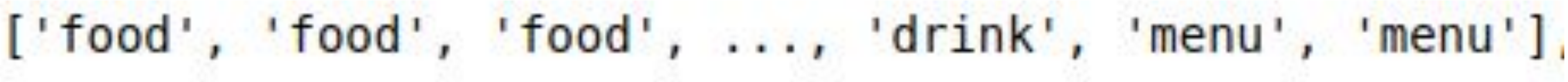


Photo Classifier **SVM** (SGD):

5 fold Cross Validated at **31%** accuracy



```
['menu', 'inside', 'food', ..., 'food', 'drink', 'menu']
```



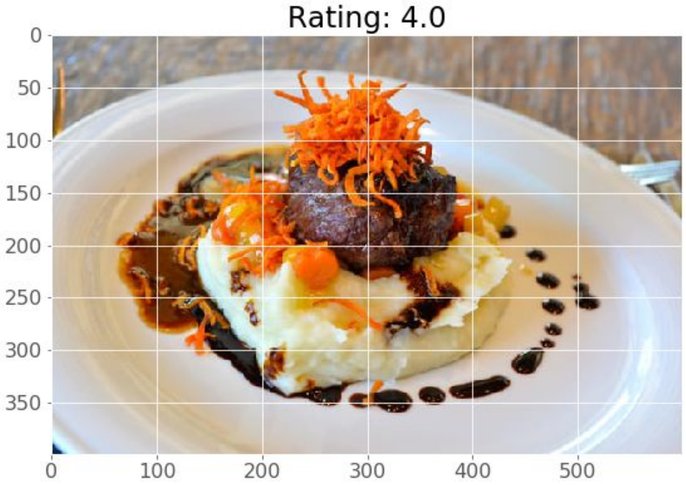
```
['food', 'food', 'food', ..., 'drink', 'menu', 'menu'],
```

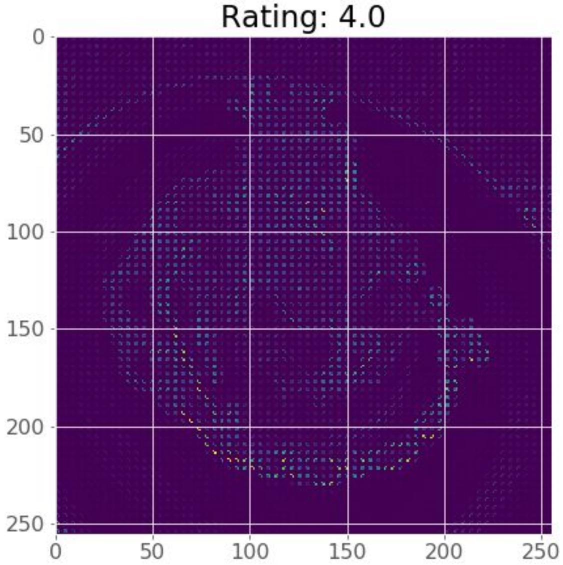
pred_img(file_11)

Photo Classifier **SVM** (SGD):

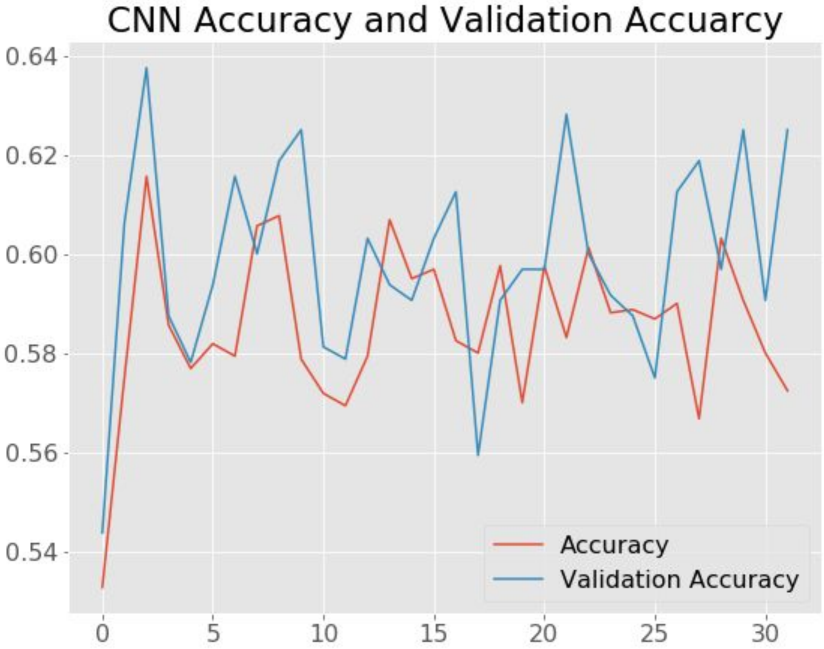
5 fold Cross Validated at **31%** accuracy

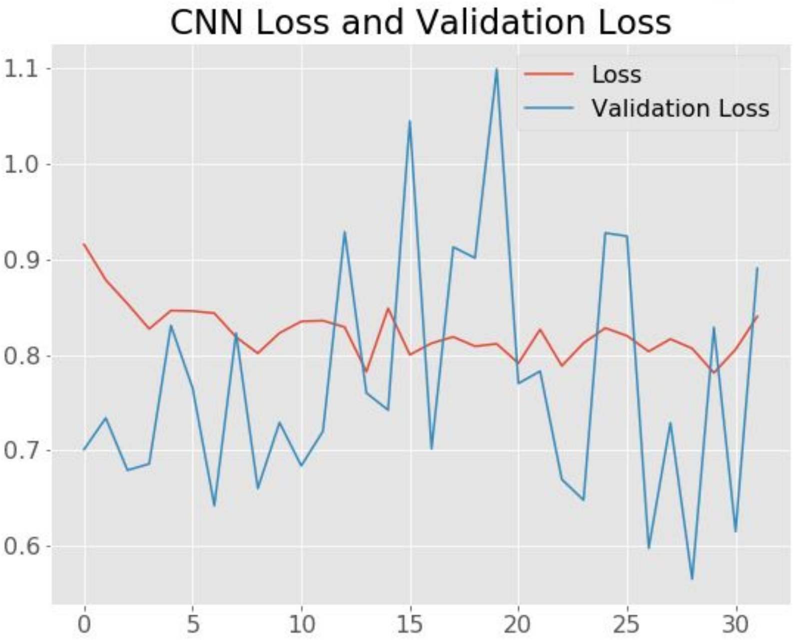
HOG'in the image **increased** accuracy to **42%**



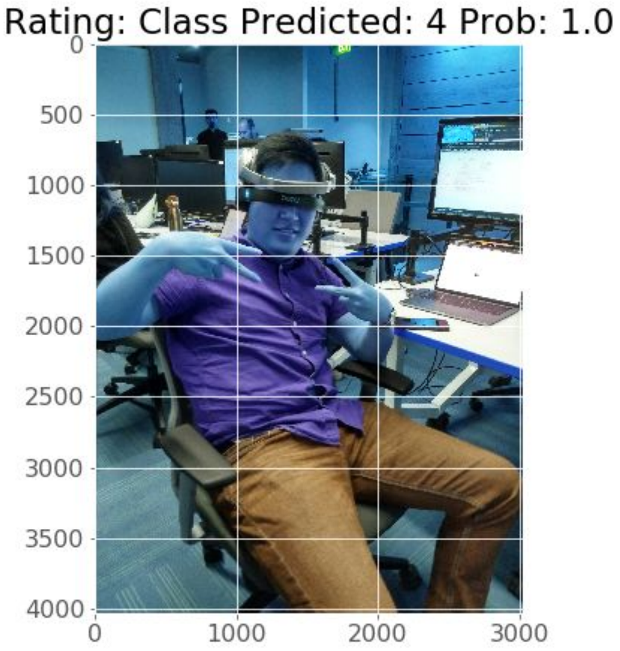


CNN To The Rescue





CNN for Fun, Maybe Not So Much



Rating: Class Predicted: 4 Prob: 1.0

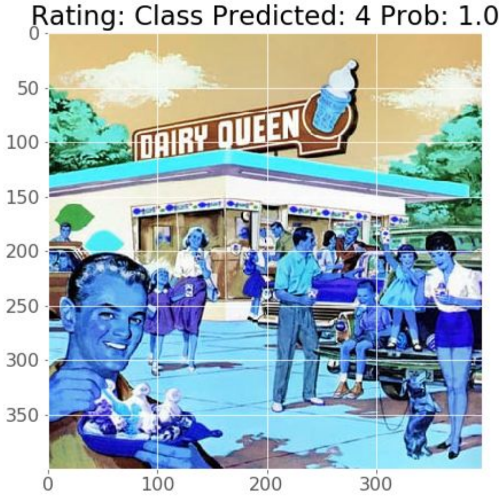
Conclusion

SVM turned to be better at **text** to rating classification than photos

CNN is **better** at classifying photos than SVM by **20%**

Future work includes CNN tuning with **more** HOG or other image preprocessing for a better model

Models are **NOT** operational and will need future retooling


Rating: Class Predicted: 4 Prob: 1.0
DAIRY QUEEN