Hotel Sentiment Classifier

ANNA D'ANGELA | 24 JANUARY 2021

MOD 5 - CAPSTONE PROJECT

Flatiron School, Online Data Science Program

This is my non-technical presentation for my capstone project. The intended audience is my hypothetical stakeholder. This presentation offers a data driven solution to their business problem. Please see my full analysis on GitHub.

Stakeholder: Denver based hospitality software company

End User: Hotel industry guest service managers

Business Case: Reach dissatisfied guests quickly to increase customer retention

01 BUSINESS CASE

For the hospitality industry, reputation is everything. Guests book based on online review scores and word-of-mouth in their community.

Negative reviews can largely impact the booking decisions of future guests.

For negative guest experiences, time and care are crucial for guest recovery.

Finding and attending to an unhappy guest before they post review can change their experience, and their score.

OBJECTIVE

Build a communication management tool to flag dissatisfied guests for rapid recovery

METHOD

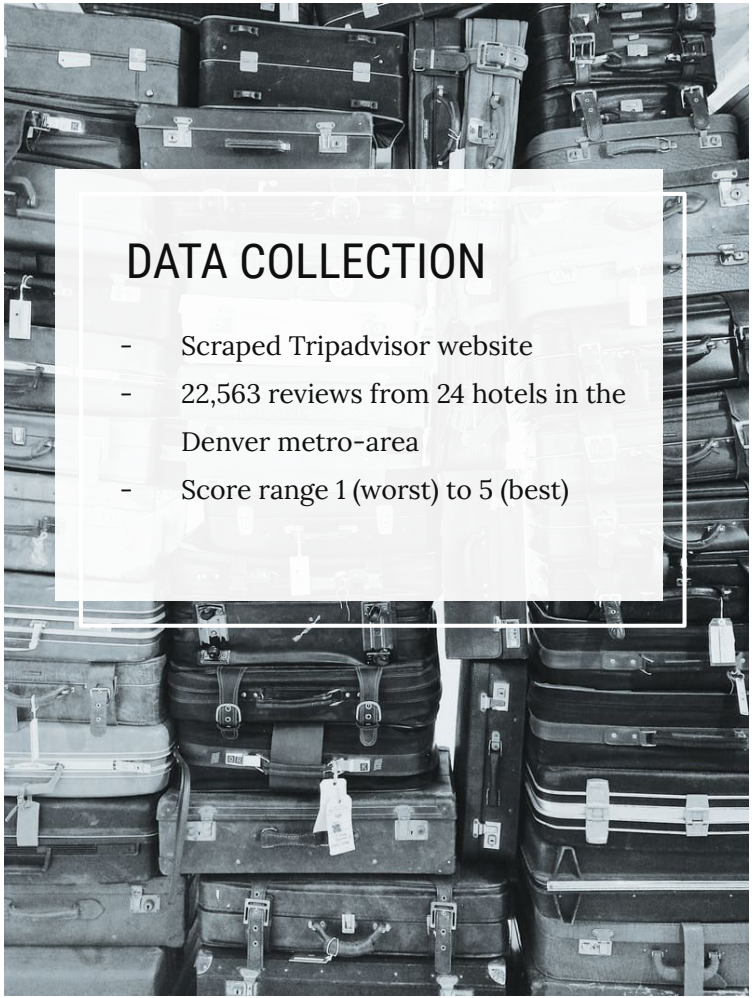
Classify text based sentiment using hotel review data

SUCCESS CRITERIA

Maximize **recall** to catch all target communication

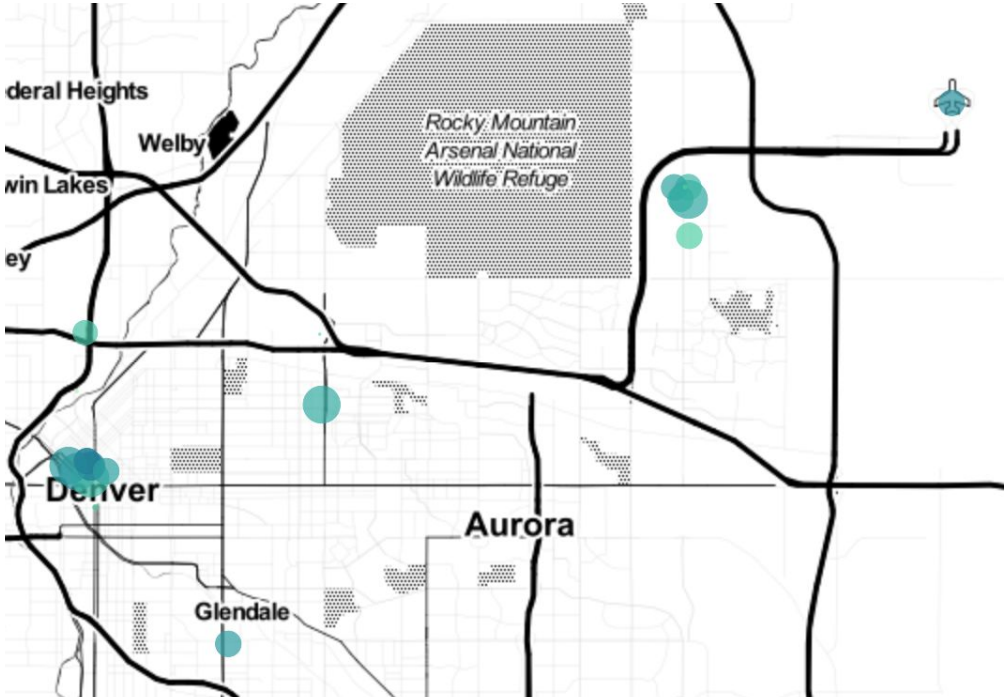
02 METHODOLOGY

- Collect hotel reviews and user rating score
- Process the text data for machine learning
- Train classifier to predict user score, given a text
- Assign sentiment value to the predicted score



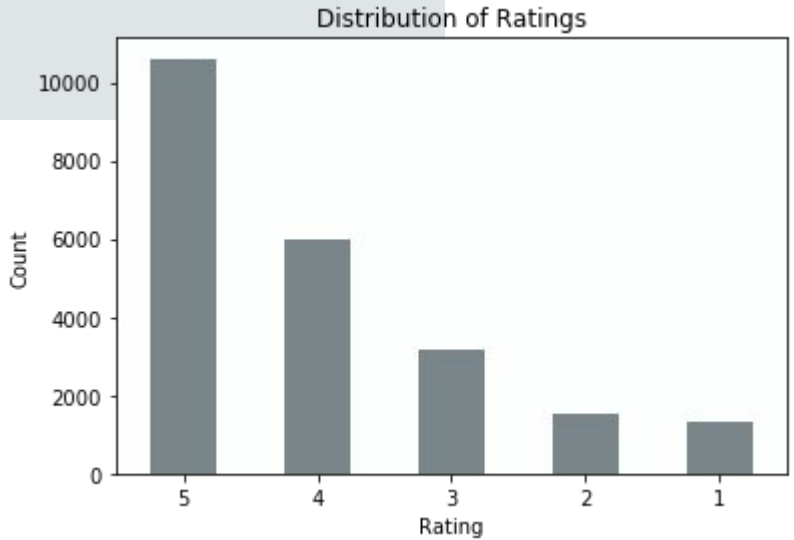
DATA COLLECTION

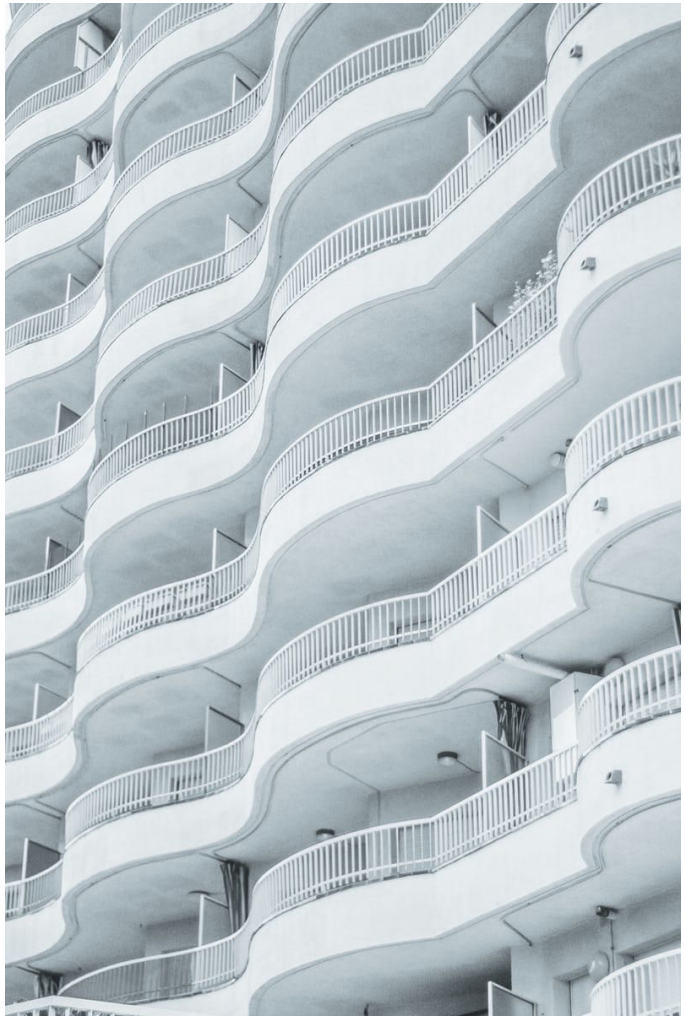
- Scraped Tripadvisor website
- 22,563 reviews from 24 hotels in the Denver metro-area
- Score range 1 (worst) to 5 (best)

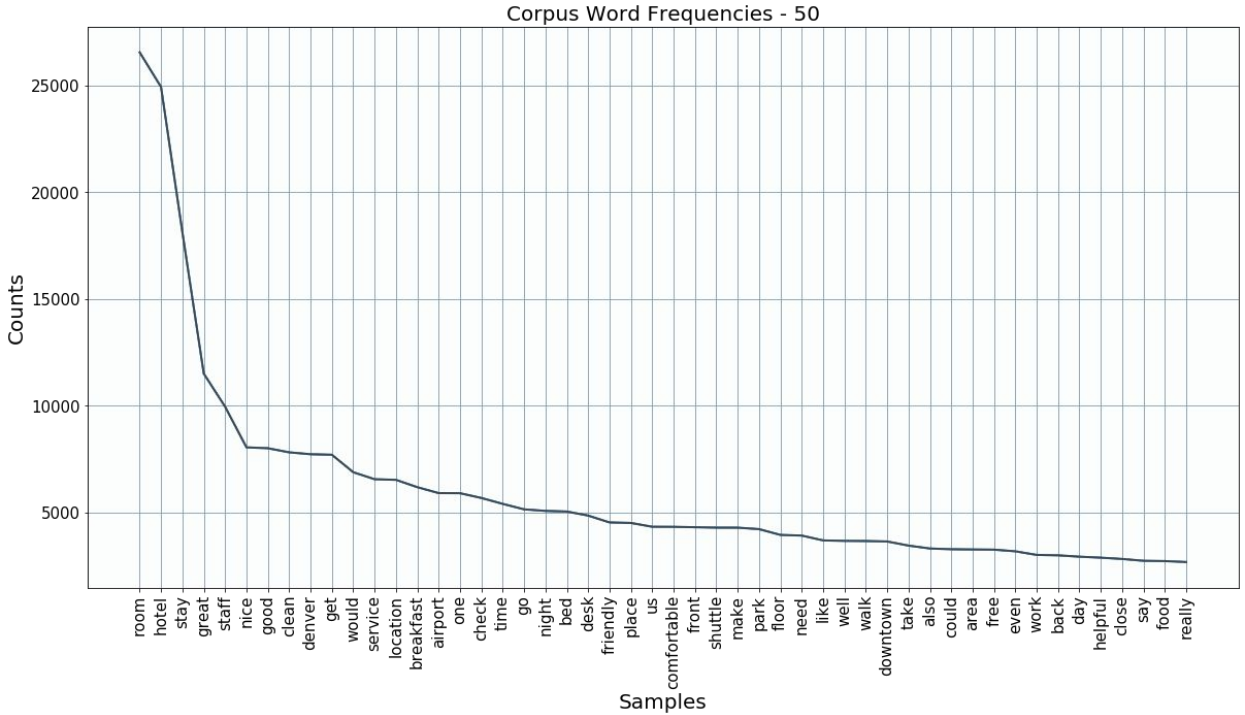


DATA LIMITATIONS

- Location specific vocabulary
- Subjective, user assigned scores, no standardization
- Severe class imbalance







PRE-PROCESSING (NLP)

Removed:

- accents
- punctuation
- stop words
- numbers
- location details
- 'noise': words appearing most frequently across all classes, thus providing no real signal

Lowercase and lemmatize words into tokens

5 Star Reviews

4 Star Reviews

3 Star Reviews

POSITIVE
3 - 5

NEGATIVE
1 - 2

2 Star Reviews

1 Star Reviews

03 MODEL ANALYSIS

89.4%

Overall accuracy

0.90

ROC / AUC (0 - 1)

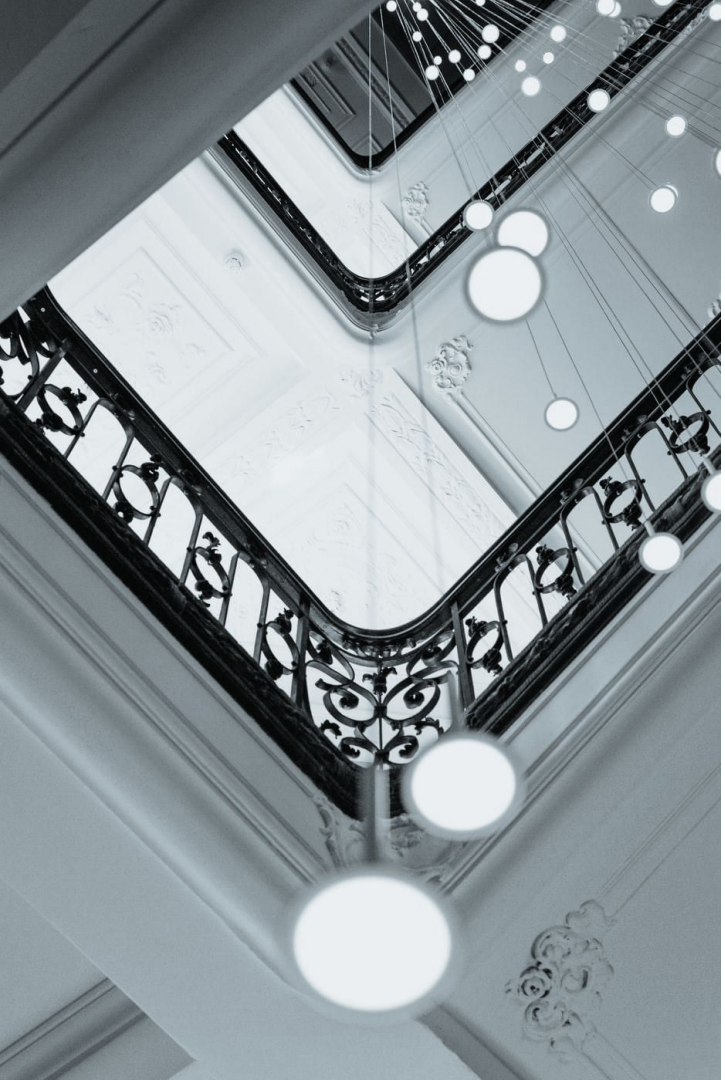
Pipeline

TfidfVectorizer

```
TfidfVectorizer(decode_error='ignore',
                stop_words=['i', 'me', 'my', 'myself', 'we', 'our', 'ours',
                            'ourselves', 'you', "you're", "you've", "you'll",
                            "you'd", 'your', 'yours', 'yourself', 'yourselves',
                            'he', 'him', 'his', 'himself', 'she', "she's",
                            'her', 'hers', 'herself', 'it', "it's", 'its',
                            'itself', ...],
                strip_accents='ascii', token_pattern="([a-zA-Z]+(?:'[a-z]+)?)")
```

SGDClassifier

```
SGDClassifier(class_weight='balanced', loss='log', random_state=619)
```



94%

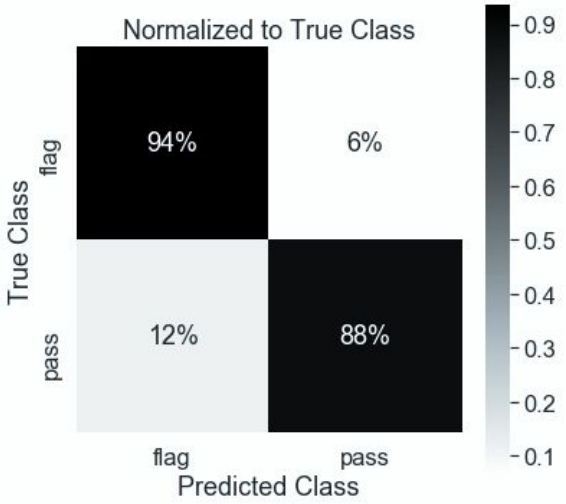
‘Flag’ recall score

6%

Missed ‘flags’

12%

Incorrect ‘flags’



RECALL SCORE

The ability of the classifier to find all the positive samples.

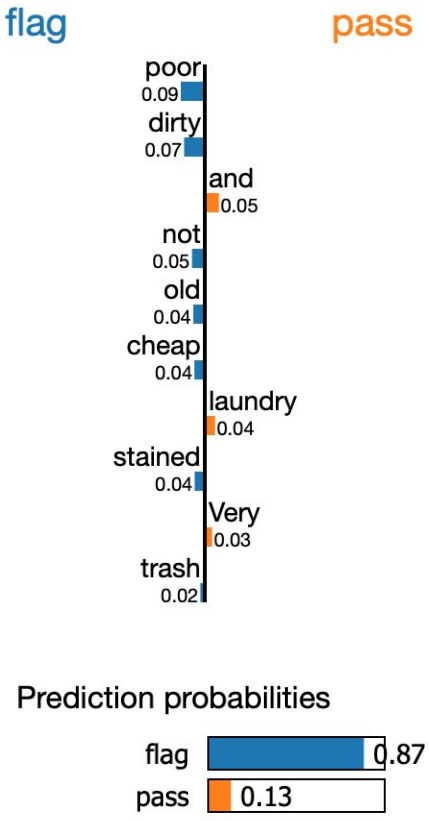
The best value is 1 and the worst value is 0.

Top L to bottom R diagonal.

SAMPLE TEXT EXPLAINED

Text with highlighted words

Very unhygenic. The bathtubs are dirty and stained. The carpet is stained and smelly and so is the linen.Stay away from this propety if you care about hygiene.The customer service is poor. The room did not have laundry sack and when I reported it to the reception they sent me a black trash bag to use os laundry bag.The microwave and refridgerator in room are old and dull too, looks cheap.

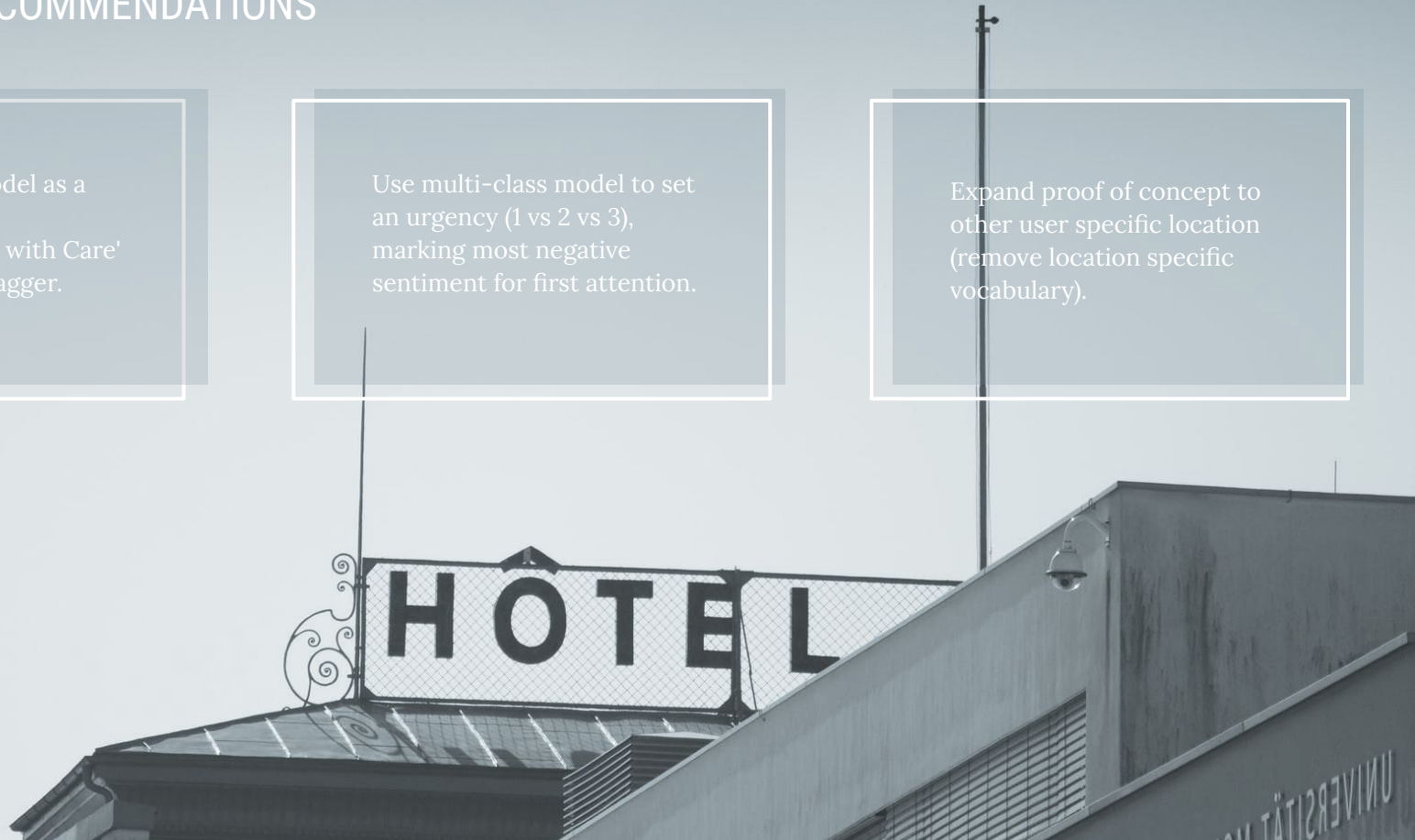


04 RECOMMENDATIONS

Use the binary model as a 'Needs Urgent Attention/Handle with Care' communication flagger.

Use multi-class model to set an urgency (1 vs 2 vs 3), marking most negative sentiment for first attention.

Expand proof of concept to other user specific location (remove location specific vocabulary).



05 FUTURE WORK

- Explore transfer learning with pre-trained NLP models.
- Explore clustering/topic modeling to examine what connects ratings in this industry.
- More locations and thorough scrub of location specific information.

Thank you!

Questions or comments?

Connect with me!

annaadangela@gmail.com

@_dangelaa

anna-dang

Full analysis on GitHub. Photos from Unsplash.