Hospitality Industry Analysis

By: Xianyue Liu, Yuting Qiu, Ruyu Wang, Haochen Zhu, Shiru Xu

Hotel Industry Overview

Overall Trend

The hospitality industry in the US has managed to score a growing number over the past few years. The total values of bookings and expenses value increased from $116 billion dollars in 2009 to $185 billion in 2017.

Market Overview

Clearly, Hotel generates revenue by selling out rooms but more importantly, the other key sources are from food and alcohol sales, service provided in the hotel such as spa, meeting room and etc. to provide customers with better service. Though the occupancy rate fluctuates month to month, the overall revenue generated from this segment keeps increasing, which means that the YoY growth in revenue is driving this segment in the market.

Business Scope

Growth in digital innovation helped customers book hotels more efficiently with different rating and reviews. In this project, we aim to develop ways to improve the rating/review system. For hotel side, we would like to do sentiment analysis of customer reviews and help them improve service in order to attract and retain customers. For customers, we will develop the rating and accommodation system to classify spam/non-spam reviews to provide them with more accurate ratings and reviews.

Executive Summary

Overview

Hospitality Industry Overall Trend

How do Hotels make money

Business scope

EDA

Dataset is yelp hotel reviews in North America. It consists of reviews and labels for spam/non-spam reviews

Sentiment Analysis + Topic Modeling

Used LSTM with GloVe Embedding to categorize the reviews into positive and negative and cluster the reviews into different topics

Spam/Non Spam Classification

Used various models to categorize the reviews into two categories which are fake and genuine

Business Insights

Revenue of hotels come from accommodation and food service mainly. There are huge potential lies in this industry.

ROI

Calculate ROI for hospitality industry based on 3 conditions: total dataset, nospam, and nospam & high rating.

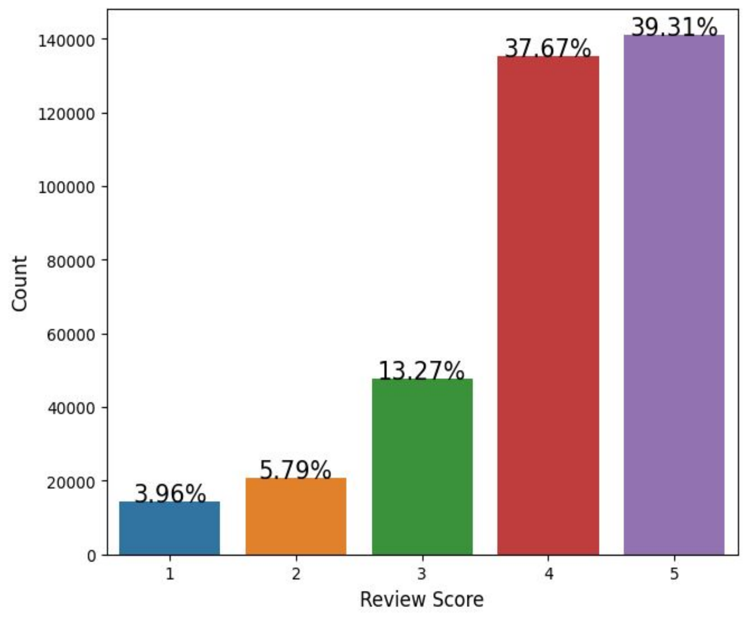
Basic Analysis

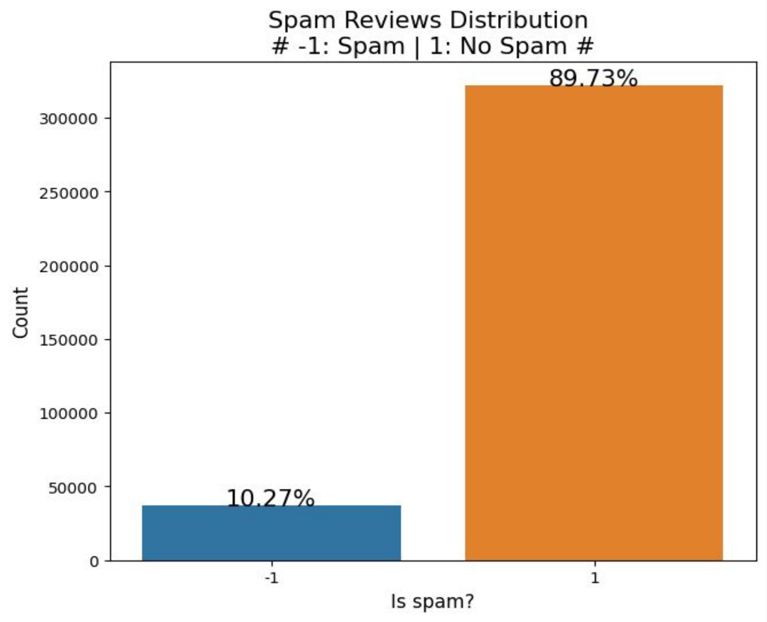
EDA & Text Preprocessing

Dataset Overview

- This dataset contains 6 columns and 359052 rows.
- 6 columns are user_id, product_id, rating score (rating scores range from 1-5), date, review, labels (-1 stands for spam and 1 stands for non-spam)
- Dataset is 100% populated and we do not need to fill in any missing values

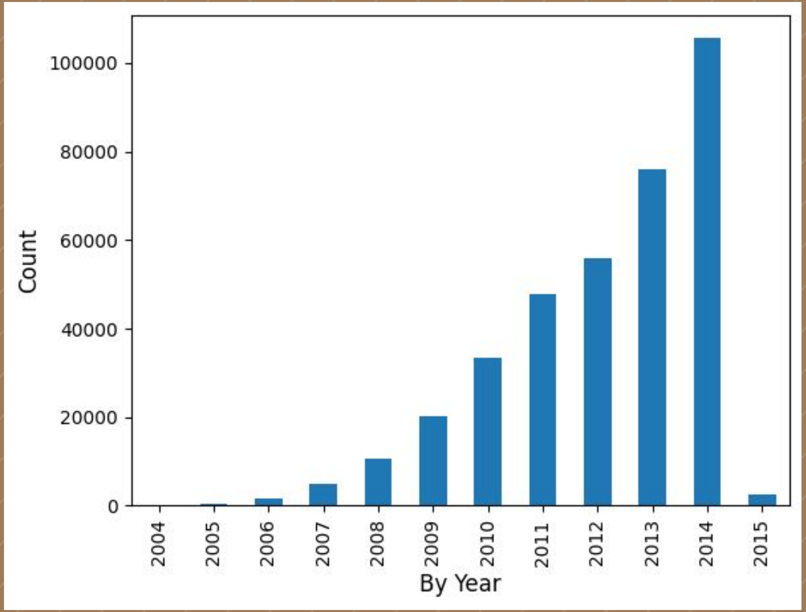
Rating Distribution vs. Spam/Non-Spam Distribution





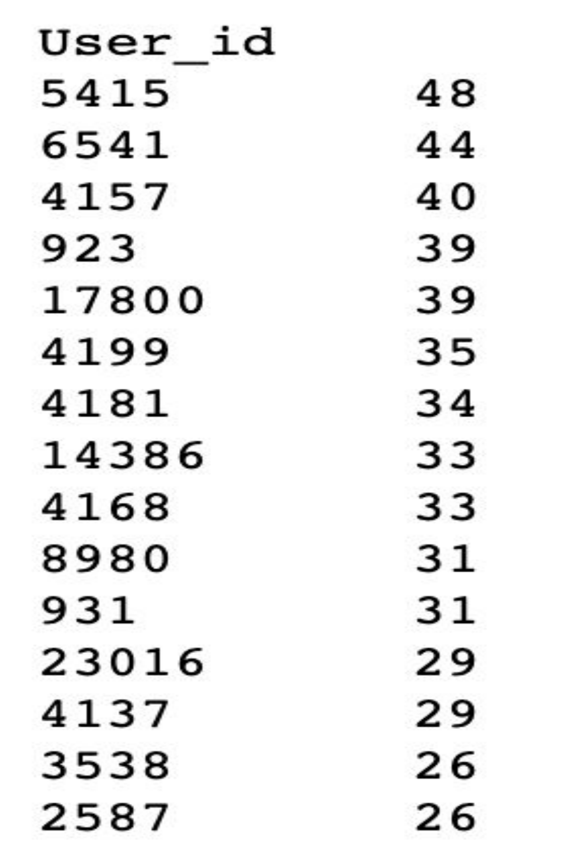
EDA Insights

- The number of users leave reviews for hotel is constantly increasing from year 2004 to 2014, and it reaches a total number of 105480 reviews in 2014



User_id with Spam Reviews

- We also discovered that these user_ids give a large number of spam reviews. We can do some more research on that to find out the characteristics of these account and try to flag out these spam reviews

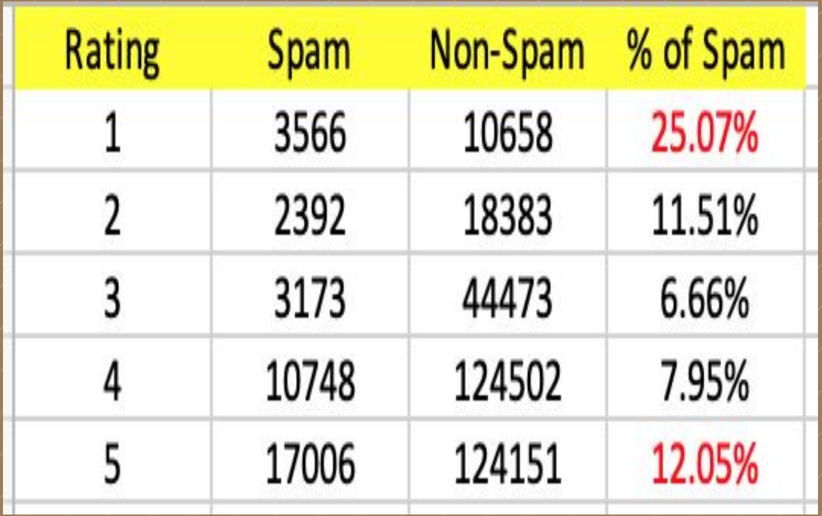


```
User_id
5415        48
6541        44
4157        40
923         39
17800       39
4199        35
4181        34
14386       33
4168        33
8980        31
931         31
23016       29
4137        29
3538        26
2587        26
```

EDA Insights

- Number of spam/non-spam postings per review score
- We can see that the percentage of spam review is higher in rating score 1 and 5, which means that these spam-reviews are not moderate and will significantly affect the overall ratings

% of Spam Overall



Rating	Spam	Non-Spam	% of Spam
1	3566	10658	25.07%
2	2392	18383	11.51%
3	3173	44473	6.66%
4	10748	124502	7.95%
5	17006	124151	12.05%

Text Preprocessing Steps

Regex

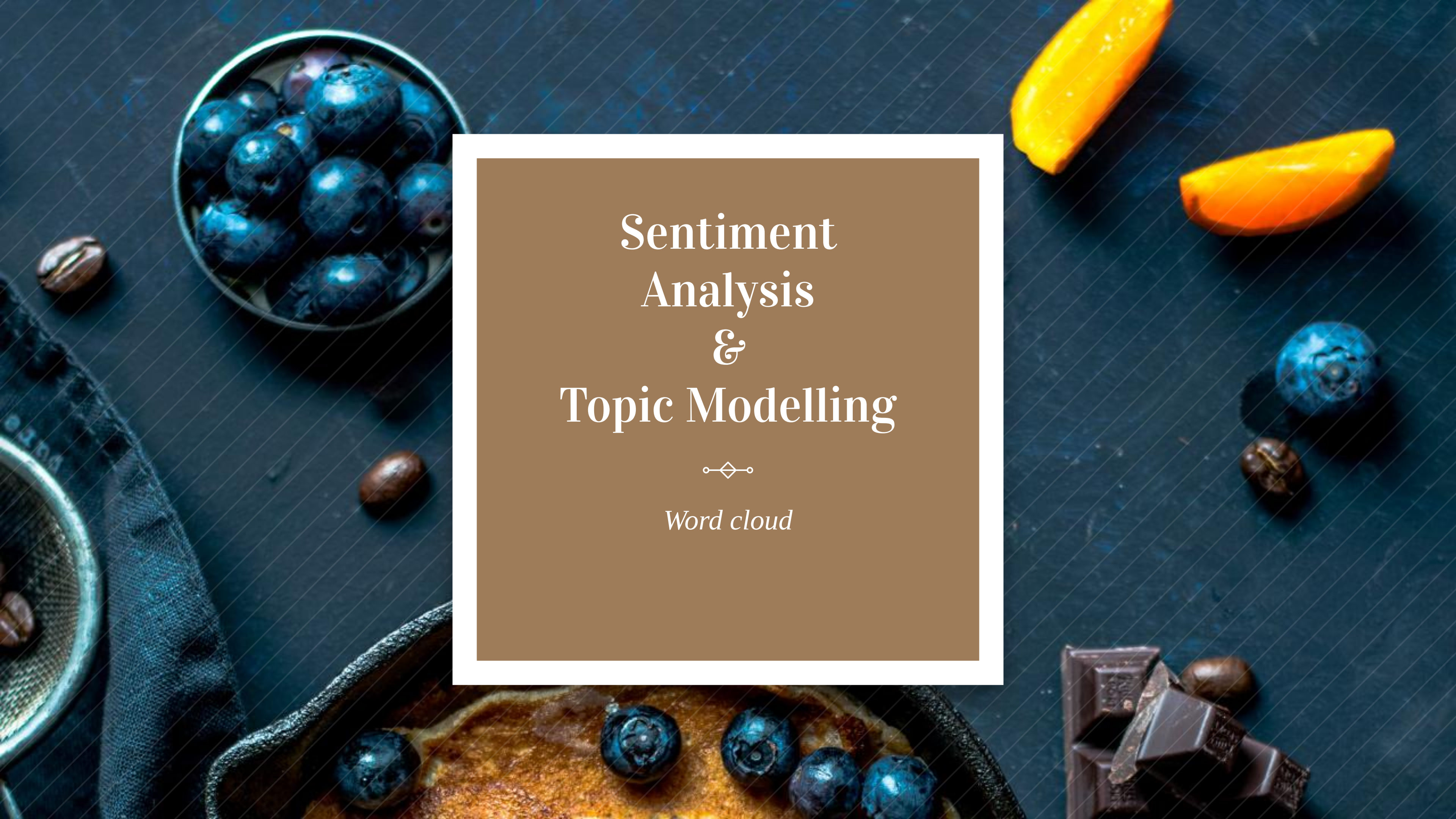
- Format text to lower cases
- Remove punctuations, urls, hashtags, emojis and etc

Stopwords Removal

- Removed the most common words (such as the, a, an, etc) because they do not add actual meaning to analysis
- Created custom stopwords

Lemmatization

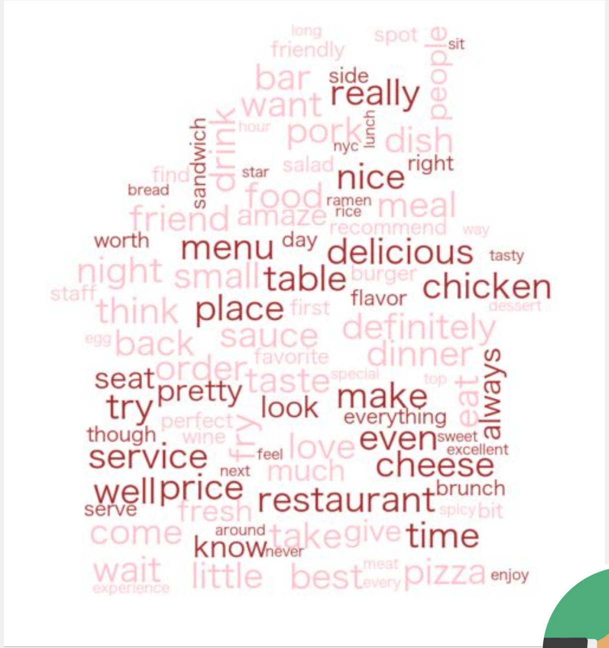
- We choose to use lemmatization over stemming in our analysis to keep original meaning of our texts as much as possible



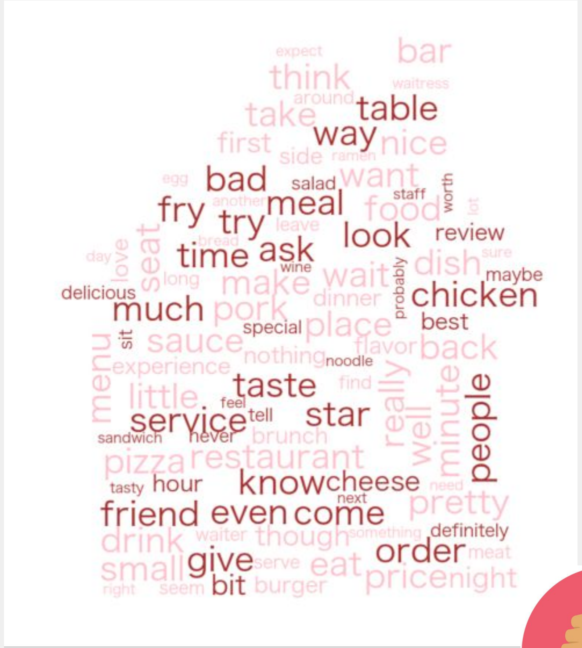
Sentiment Analysis & Topic Modelling

Word cloud

Exploratory Sentiment Analysis



Positive Reviews



Negative Reviews

Models

Sentiment Analysis Model

Aim: Predicts whether each review's sentiment is positive or negative (we set 1-3 ratings to be negative and 4-5 ratings to be positive)

Architecture: Recurrent Neural Network with a Long-Term Short Memory (LTSM) layer. The Embedding used pre-trained Glove.

Output: Gives a score between 0 and 1 for all the genuine records. The high scores (0.5 or higher) indicate that the review is more positive, and low scores (0.5 or lower) indicate that the review is more negative

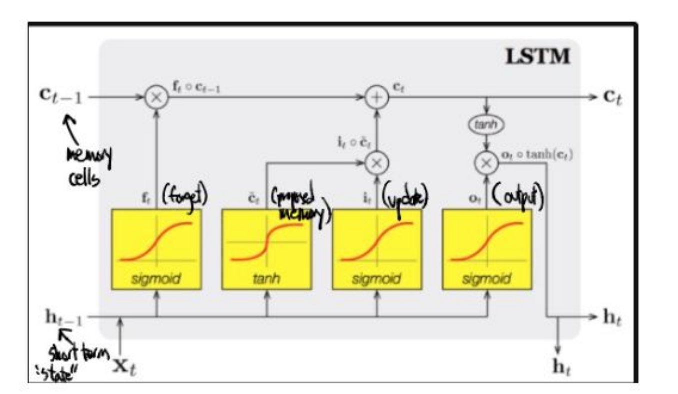
Topic Modeling Model

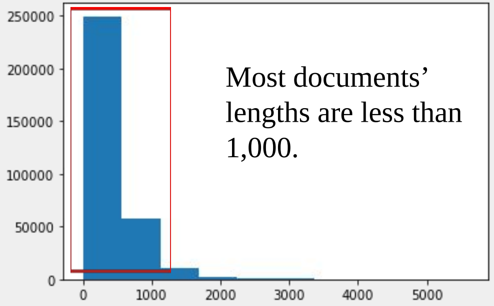
Aim: Cluster the reviews into different topics for positive reviews and negative reviews

Architecture: Non-Negative Matrix Factorization

Output: Gives the topic that each review belongs to

Sentiment Analysis Model Architecture





Most documents' lengths are less than 1,000.

Architecture

The model was Long short term memory with pre-trained Glove vector embedding. LSTM model is an extension of RNN. It considers the previous output (feedback) and stores in its memory for a short period of time (short-term memory). The model outputs the probability(using a sigmoid function) that each review was positive or not.

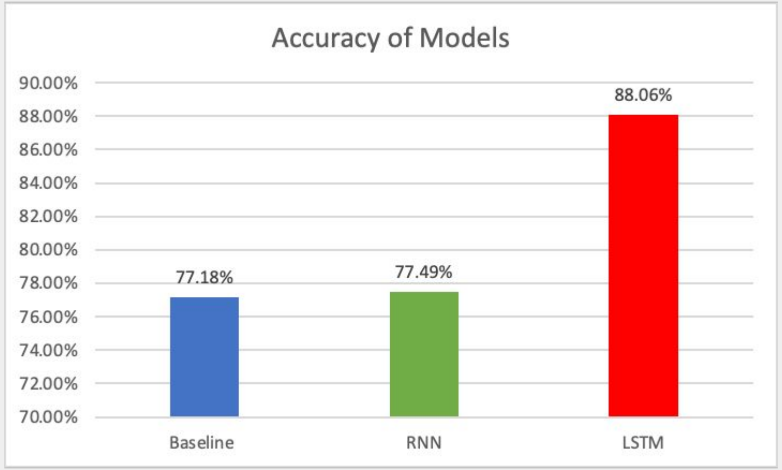
Hyperparameters

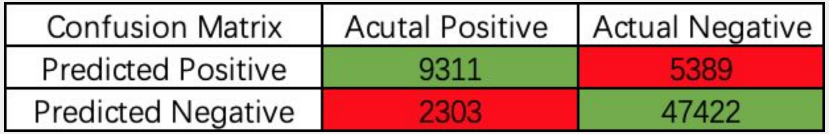
Vocab Size: 122320, Max_length:100,

MAX_SEQUENCE_LENGTH: 121321

Layers: Embedding (Glove), Masking, LSTM (32 units), Dense (16 Units), Dense (2 Unit Sigmoid), Activation: 'softmax', epochs: 3

Sentiment Analysis Model Performance



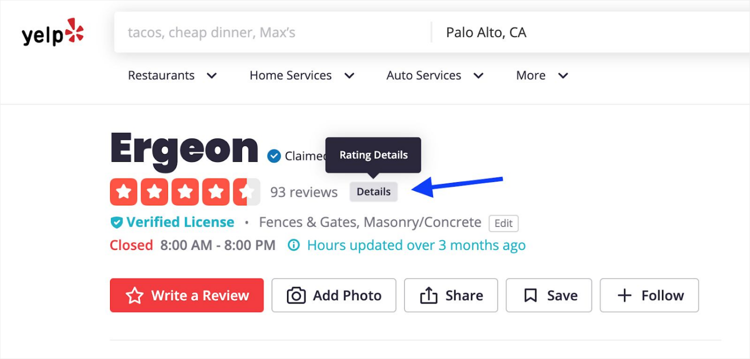


Confusion Matrix	Acutal Positive	Actual Negative
Predicted Positive	9311	5389
Predicted Negative	2303	47422

- Our chosen LSTM model performs better than baseline accuracy: 88.17% v.s. 77.18%.

- The precision was 63.34%.
- The recall was 80.17%.
- Therefore, the model was more good at predicting positive review than negative review.
- The reason might be that sometimes in the negative reviews, reviewer will mention some good points of the hotel as well. This will make the sentiment confusing and biased.

Sentiment Analysis Business Use Cases



yelp

tacos, cheap dinner, Max's

Palo Alto, CA

Restaurants

Home Services

Auto Services

More

Ergeon

Claimed

Rating Details

93 reviews

Details

Verified License • Fences & Gates, Masonry/Concrete Edit

Closed 8:00 AM - 8:00 PM

Hours updated over 3 months ago

Write a Review

Add Photo

Share

Save

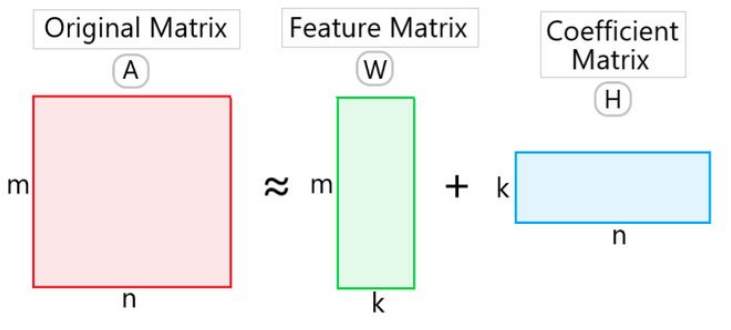
Follow

Improved Rating System

Original Yelp rating system used 1-5 scale with reviewers' comments. The ratings is helpful in helping customers understand the overall goodness of the hotels.

However, our model can help create a new type of rating, like the recommend/ not recommend system so that users can have a more direct ideas about what percentage of users will recommend the hotel, which can help users understand the goodness of hotel in another way.

Topic Modelling Architecture



Models

Approach 1: Non-Negative Matrix Factorization **(Chosen)**

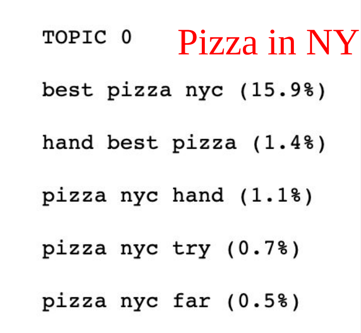
Approach 2: LSA (Latent Semantic Analysis)

Approach 3: Latent Dirichlet Allocation

Hyperparameters

```
vectorizer = TfidfVectorizer(ngram_range=(2,2),
                             binary=True,
                         token_pattern=r'\b[a-zA-Z]{3,}\b',
                             stop_words="english",
                             min_df=4,
                             max_df=0.5)


nmf = NMF(n_components=4)
```

Topic Modelling Results- Positive Reviews

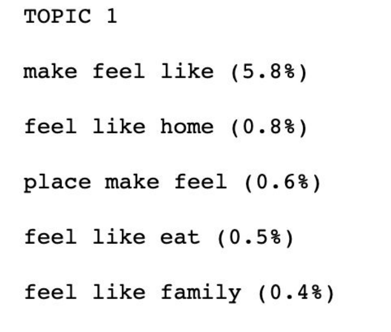


```
TOPIC 0

best pizza nyc (15.9%)

hand best pizza (1.4%)

pizza nyc hand (1.1%)

pizza nyc try (0.7%)

pizza nyc far (0.5%)
```

Pizza in NYC

- one of the best pizzas in nyc :) its not cheap though but worth it
- called the best pizza in nyc. undecided if it's better or different than #chicago #pizza.
- a must visit for the best pizza in nyc. also have great breakfast sandwiches and burgers too. best food for the money in the city.
- the best pizza i have had in new york city!
- at least one of if not the best pizza in new york

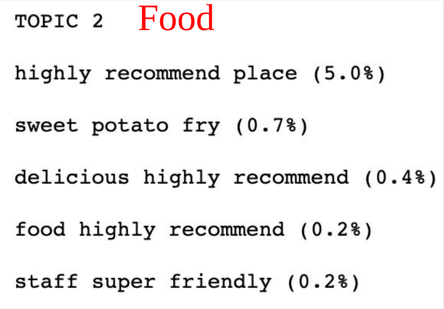
Living/Eating Feelings



```
TOPIC 1

make feel like (5.8%)

feel like home (0.8%)

place make feel (0.6%)

feel like eat (0.5%)

feel like family (0.4%)
```

- to be read in the poshest british accent you can manage* mmm! ah! i say, lord drunklington, have you met my good friend lillie? she's a delightful old girl, very beautiful inside and out, splendidly arrayed at all times in red velvet and marble. she's positively a class act! or, as we say at oxford, a proper chum! and garn, if she doesn't make a smashing cocktail! i'm quite partial to the diamond fizz, absolut citron with lemon juice and prosecco, and makes me feel like i'm a proper heiress with diamonds on. lord drunklington, you simply must let me introduce you to lillie. she'll be so charmed to make your acquaintance!
- great lunch place- working in soho this is a great place to escape the hustle while making you feel like your on fiesta in spain for a hour. yummy salads that's a good portion and delish tapas! and you get to try the samples! :)
- gandhi is a wonderful place. i lived in india for some time, and typically avoid any sort of "indian" food in america. gandhi, however, makes you feel like a member of the family. the soap operas remind me of the farm i lived on in sirsa, where the farmers wife would watch them every day. traditional, tasty, authentic, reasonable, and the ã□â friendliest people in the world. go now if you live anywhere nearby.
- great for lunch, takeout, delivery...open 24/7. empanadas are amazing and the arroz con pollo makes you feel like you're home.

Topic Modelling Results- Positive Reviews

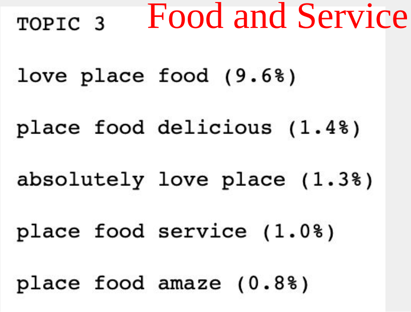


Food

```
TOPIC 2

highly recommend place (5.0%)

sweet potato fry (0.7%)

delicious highly recommend (0.4%)

food highly recommend (0.2%)

staff super friendly (0.2%)
```

- ate there last night, and it was delicious! my swordfish was huge!! salad also big! i highly recommend this place.
- easily the best sliders in the city. i get a double when i'm hungry. the shakes are good too. i highly recommend this place.
- met an old colleague/ friend for lunch here last week and heart this place. he always picks good places to meet up but this place was a great pick. cute, great lunch menu and in my dream neighb (i'd give my left foot to live on this particular block!), highly recommend this place. it's sorta romantic too (though my lunch date was not romantic haha!).
- i had lunch here yesterday and i loved it. it was crazy busy, but the service was good. i had one of the best sandwiches ever and a glass of sangiovese to go with. i highly recommend this place.

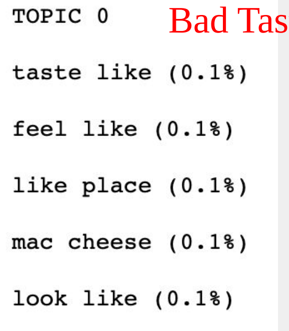
- i love this place. great food, professional staff and great mojitos. and very romantic atmosphere.
- i love this place. the food's delicous, the servers are sweet, and it's a killer space. my one tiny gripe is two out of the four times i've been there, they've been playing radiohead, but that's just me. i fucking hate radiohead.
- love this place, the food is good and the happu hour is awesome!!!
- llove this place great food i get the mixed kofta and chicken in rice because f having to choose great ppl too
- i love this place. the food was great!!!!



Food and Service

```
TOPIC 3

love place food (9.6%)

place food delicious (1.4%)

absolutely love place (1.3%)

place food service (1.0%)

place food amaze (0.8%)
```

Topic Modelling Results- Negative Reviews

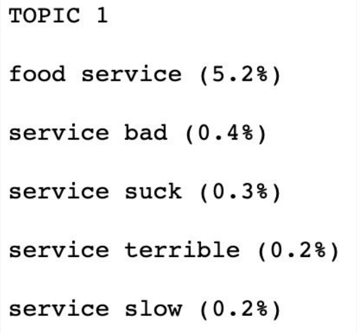


```
TOPIC 0

taste like (0.1%)

feel like (0.1%)

like place (0.1%)

mac cheese (0.1%)

look like (0.1%)
```

Bad Taste

- came here for lunch really wanting to like this place. but the food simply sucks and lacks flavor
- this wasn't anything great. ã□â i was really excited to try this place out. ã□â i feel like this place is over hyped.
- i like this place...except when it is tremendously busy...then it's not so much fun.
- i'm with everyone else... get the mac and cheese.
- the ingredients didn't meld. tasted like bread and cream together.

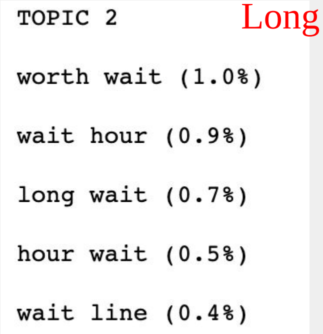
Unsatisfied Service



```
TOPIC 1

food service (5.2%)

service bad (0.4%)

service suck (0.3%)

service terrible (0.2%)

service slow (0.2%)
```

- ok food, ok service.
- overrated. the concept of organic surpasses the food and the service. don't see the cohesion of food.
- food was good, service was dissapointing!
- ts great food is better than service oh well.
- food is great , the service isn't!

Topic Modelling Results- Negative Reviews

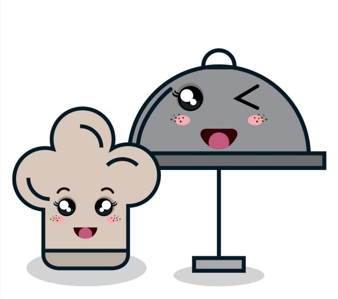


```
TOPIC 2

worth wait (1.0%)

wait hour (0.9%)

long wait (0.7%)

hour wait (0.5%)

wait line (0.4%)
```

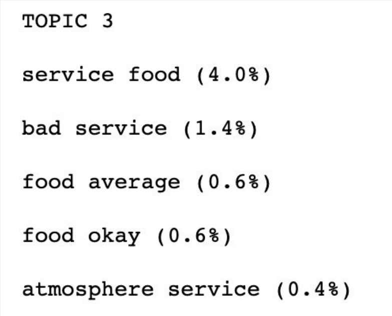
Long Waiting Time

- nothing special and certainly not worth a wait.
- ok food. definitely not worth waiting in line for.
- really not worth the wait.
- not worth waiting in line for hours.
- good pizza- not worth the wait.

Unsatisfied Food and Service



- the service was great but the food...not so much
- bad service. ã□â good food.
- not bad. good service, and food is above average.
- great service but food is just okay.
- service good. food and drink no good.



```
TOPIC 3

service food (4.0%)

bad service (1.4%)

food average (0.6%)

food okay (0.6%)

atmosphere service (0.4%)
```

Topic Modeling Performance

Overall the topics can be diversified to food, service, waiting time and feelings.

However, there are some overlaps between topics.

In negative reviews. Topic 0 is about food taste, topic 1 is about service but topic 3 contains food and service.

In positive reviews, Topic 2 is about food, Topic 3 is about food and service.

Topic Modeling Use Cases

Reporting System

Major Hotels in Yelp will often want feedback on the customer sentiment and the topics they care about, especially for the negative topics, therefore they can know what areas to maintain and what areas to improve in order to better serve customers.

Our model can therefore help to generate the reports about customers’ frequently mentioned positive and negative topics and help to guide hotels’ further directions of improvements.

Fake/Genuine Review Classification Model

Linear model, Random Forest, Naive Bayes, Lstm, RNN

Preprocessing and Models

Data Preprocessing and embeddings:

1.Add customized stop words after performing wordcloud for spam and non spam reviews

2.Use TF-IDF to determine relatively phrases with ngrams = 2

3.GloVe Embedding for RNN and LSTM models

Models:

Logistic regression

Multinomial Naive Bayes

Random Forest

LSTM

RNN

Model Performance

Logistic Regression

Model Performance:

Accuracy:

97.79%

Predicted Accuracy:

72.27%

Naive Bayes

Hyperparameters:
alpha=0.2

Model Performance:

Accuracy:

99.49%

Predicted Accuracy:

72.4%

Random Forest

Hyperparameters:
n_estimators=250,
random_state=50,max_depth =
30

Model Performance:

Accuracy:

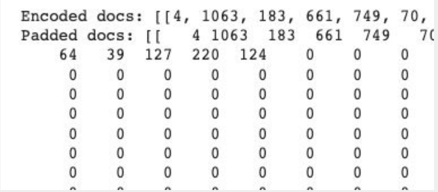
76.41%

Predicted Accuracy:

67.62%

Input documents → Integer indices

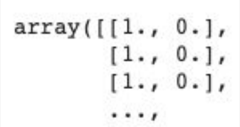
- Use spacy pipeline to tokenize and remove stopwords
- Use tokenizer from keras to encode texts to integers
- Pad the docs to max sequence length



```
Encoded docs: [[4, 1063, 183, 661, 749, 70,
Padded docs: [[    4 1063  183  661  749   70
    64   39  127  220  124    0    0    0
     0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0
```

Encode labels as 1 and 0

- Use sklearn LabelEcoder to encode the labels as:
  - 0: genuine review (not-spam)
  - 1: fake review (spam)



```
array([[1., 0.],
       [1., 0.],
       [1., 0.],
       ...,
```

Create Embedding Matrix

- Download and load in GloVe vectors (the embeddings trained by GloVe)

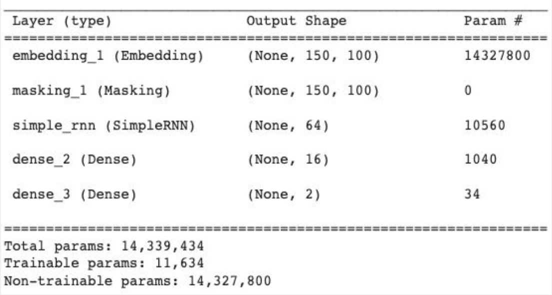
Define RNN Modle

- First Layer: integer indices and the embedding matrix
- Masking Layer: make place where no embeddings to be 0
- Units: make the dimension for hidden state to be 64
- Activation function: ‘softmax’
- Test size: 0.2
- Epochs: 3
- Output:
  - Loss function: ‘categorical_crossentropy’
  - Metric: accuracy

Define LSTM Modle

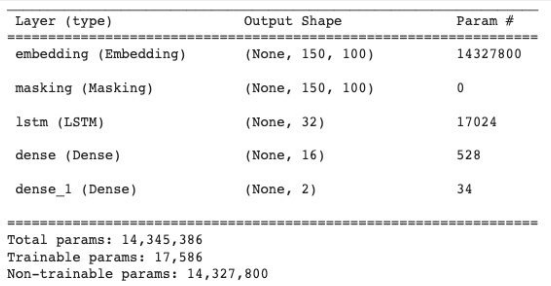
- First Layer: integer indices and the embedding matrix
- Masking Layer: make place where no embeddings to be 0
- Units: make the memory cells to be 32
- Activation function: ‘softmax’
- Test size: 0.1
- Epochs: 3
- Output:
  - Loss function: ‘categorical_crossentropy’
  - Metric: accuracy

RNN Model



```
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 150, 100)          14327800

masking_1 (Masking)          (None, 150, 100)          0

simple_rnn (SimpleRNN)       (None, 64)                10560

dense_2 (Dense)              (None, 16)                1040

dense_3 (Dense)              (None, 2)                 34

=================================================================
Total params: 14,339,434
Trainable params: 11,634
Non-trainable params: 14,327,800
```

LSTM Model



```
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 150, 100)          14327800

masking (Masking)            (None, 150, 100)          0

lstm (LSTM)                  (None, 32)                17024

dense (Dense)                (None, 16)                528

dense_1 (Dense)              (None, 2)                 34

=================================================================
Total params: 14,345,386
Trainable params: 17,586
Non-trainable params: 14,327,800
```

RNN Architecture

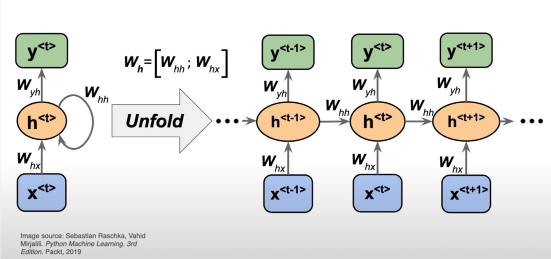
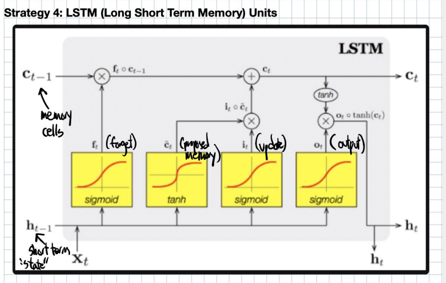


Image source: Sebastian Raschka, Vahid Mirjalili. *Python Machine Learning. 3rd Edition.* Packt, 2019

LSTM Architecture



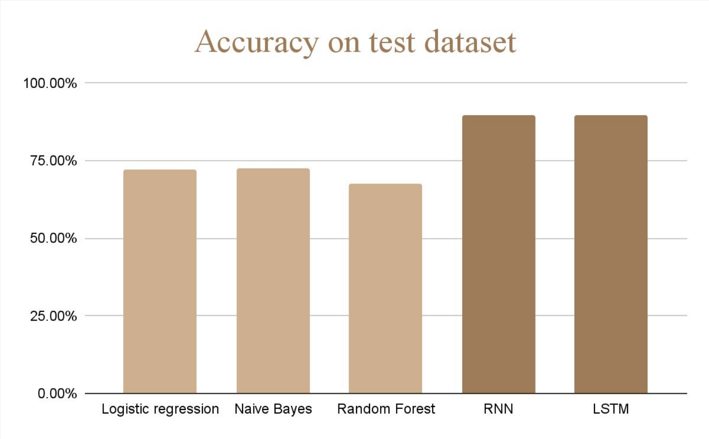
Strategy 4: LSTM (Long Short Term Memory) Units

RNN & LSTM | Tokenization & Embedding | Model | Accuracy

Test dataset Accuracy

- RNN: 89.739%
- LSTM: 89.741%

Use supervised classification Lstm and RNN model with GloVe embeddings to compare the models we previously built



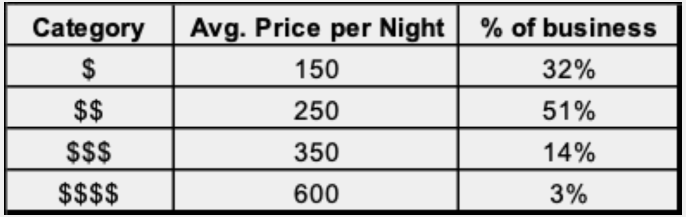
Business Insights

Market size and Revenue model,
ROI
&
Use case for spam classification

Market Size & Revenue Model

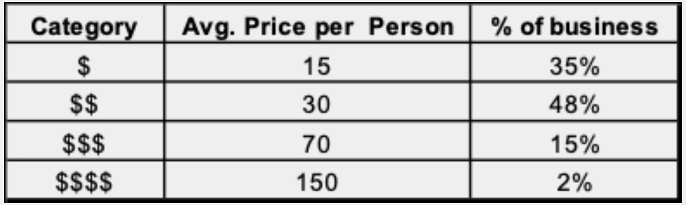
- The market size of Hotels and Motels industry in US is $177.6bn in 2022.
- Revenues of hotels mainly come from Accommodation and Food Services. According to listings on Yelp, hotels can be classified into 4 categories in terms of average price.

Accommodation



Category	Avg. Price per Night	% of business
$	150	32%
$$	250	51%
$$$	350	14%
$$$$	600	3%

Food Service:



Category	Avg. Price per Person	% of business
$	15	35%
$$	30	48%
$$$	70	15%
$$$$	150	2%

Return On Investment (Metrics)

1.How much investment is in reviews in hospitality inventory(per month)

2.Number of leads from reviews (per month)

3.Average customer spend in hotels (per month)

4.Visits per customer per month

5.Lead conversion rate

*Current Invest Value = Total Leads * Average Customer Spending * Number of Visits Per Customer * Conversion Rate

*Leads:someone who is much more likely to be a new customer or become one soon

Return On Investment (Calculation)

- Assume $3000/month investment on reviews, 1000 leads per month, spend about $200 per visit, 400 converted leads(Only for one hotel)

- Calculate visits per customer per month from data set based on (total dataset(2), non spam(2.2) and both non spam & higher rating(2.2))

- Leads conversion rate: converted leads/total leads = 4%

  After filtering out spams and higher rating, the conversion rate increased to 6%

- ROI(Total) = (Current Investment Value - Investment Cost) / Investment Cost

  = (4%*250*2*1K - 3K)/3K =567%

- ROI(non spam) = ROI(non spam & higher rating) = (6%*250*2.2*1K-3K)/3K =1000%

- **Conclusion**: Updated spam filter system and more positive rating would increase ROI

Classification Model Business Use Case

Create better environment for Yelp review

- By letting the fake/genuine classification model automatically delete the fake reviews, Yelp can ensure the genuineness of the reviews
- Punish the accounts which often post fake reviews
- Yelp can also use this model on its other review platforms
- Use this model as the first step of its review data process, so Yelp can offer more accurate recommendations for customers

Selling the model

- Sell this model to other similar platforms
    - For example, it can sell to Twitter to distinguish fake tweets

Conclusion

Improvement
&
Next steps

Sentiment Analysis & Topic Modeling Next Step

- Implement dimensionality reduction before modeling to get rid of correlated features and overfitting problem.
- Considering more variables such as location and star of hotels as inputs for the model.
- Try on more advanced models such as transformers (BERT) to save training time and realize dynamic predictions.

Classification Model Next Step

Tune some hyper-parameters

- We can try to change max_sequence_length in the step of tokenization
- This time we fit in 3 epochs, next time we will use k-means to decide the best number of epochs
- We can also add some layers in NN

Try other algorithms

- If we have enough computing power, we can train our own embeddings
- We can try other advanced algorithms like BERT

Thank you

References

https://blog.yelp.com/businesses/understanding-yelp-metrics/
https://www.mordorintelligence.com/industry-reports/hospitality-industry-in-the-united-states
https://wpbusinessreviews.com/roi-review-marketing/