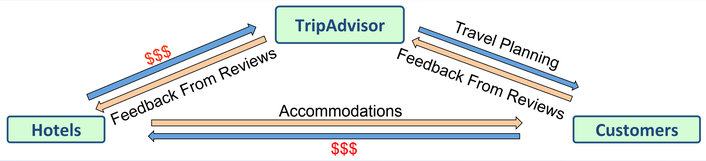
Opinion Mining and Sentiment Analysis of TripAdvisor Hotel Reviews

Yaxiong Cai, Toby Du, Peiheng Hu, Arjun Sanghvi

Instructor: Pavlos Protopapas, TF: Rahul Dave

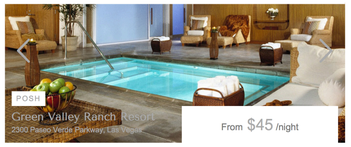
Capstone • Spring 2015

Motivation



Original Reviews and Expected Outcome

Original Reviews on Jetsetter



What we love

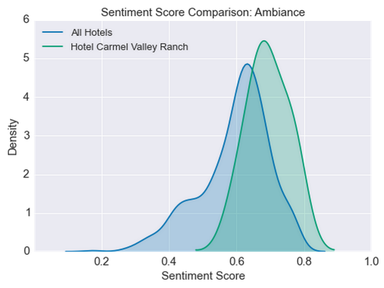
- Jetsetter rate includes one $50 value pack (valid once per stay) that comes with perks like one free cocktail at any casino bar, one free bingo blue pack, one free $1 three-spot keno ticket, one free dessert with entree purchase at Grand Cafe, one free bowling game at Red Rock Lanes, one $10 table-games match play, one $30 poker room buy-in for $20 and a 10 percent discount at the gift shop
- Modern, Mediterranean-inspired rooms with thoughtful amenities
- Locally revered, sexy steakhouse Hank's for meat and martinis
- A smart spa you won't want to leave
- Wide array of scheduled entertainment, including A-list headliners
- Family fun, including 10 cinemas, a full arcade and an expansive pool
- Airport transfers and WiFi are included in the daily resort fee

What to know

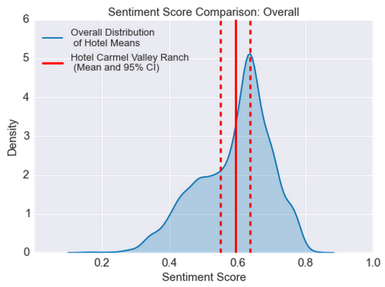
- A resort fee of $25 per night will be collected upon checkout from Jetsetter
- There is a two-night minimum stay for Deluxe Rooms
- Green Valley Ranch is approximately 10 miles (15–20 minutes) from the Las Vegas Strip
- Other than the District, not much is within walking distance
- Like most Las Vegas resorts, access to many amenities and restaurants is through the casino floor
- No scheduled family or kids activities (though the District offers some)

Expected Outcome

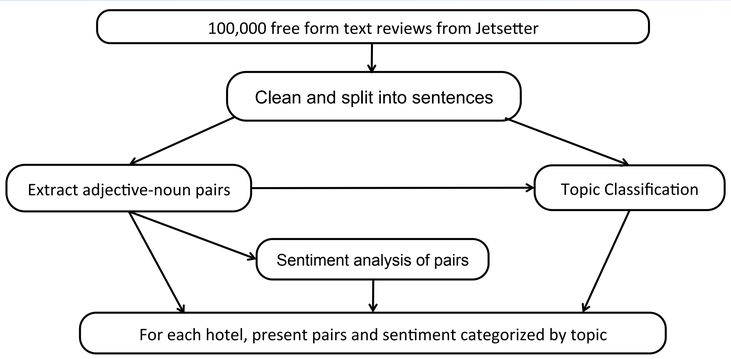
```
Category: Room
    nice bathroom
    delicious room service
    same room
    spacious room
    lovely room
  Average sentiment score of  Room is 0.65
--
  Category: Service
    front desk employee
    smaller room
    little working
    helpful staff
  Average sentiment score of  Service is 0.51
--
Category: Ambiance
    peaceful environment
    comfortable suites
    beautiful property
    wonderful scenery
  Average sentiment score of  Ambiance is 0.69
```




```
--
Overall sentiment score of hotel is 0.593
```

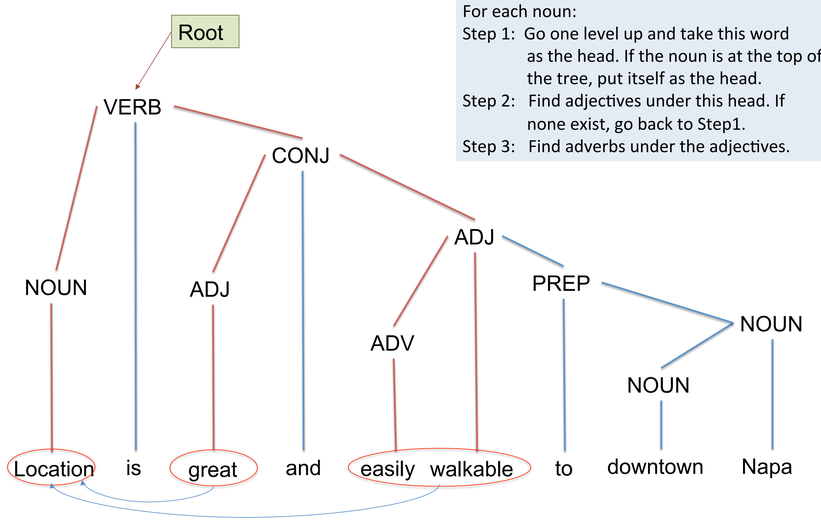


Methods and Framework



Extraction of Adjective-Noun Pairs

Dependency trees provide a representation of grammatical relations between words in a sentence.



For each noun:
Step 1: Go one level up and take this word as the head. If the noun is at the top of the tree, put itself as the head.
Step 2: Find adjectives under this head. If none exist, go back to Step1.
Step 3: Find adverbs under the adjectives.

Using this method, we can extract both pairs:

1. (Great, location)
2. (Easily walkable, location)

Topic Classification

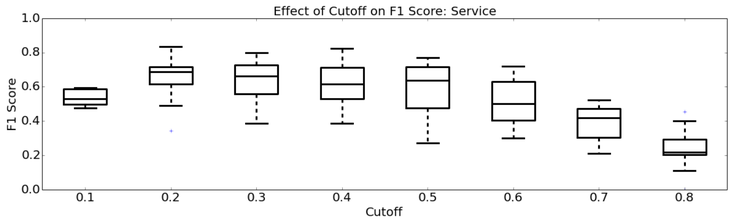
- Hotel rating systems tend to focus on a few, specific categories
- To be consistent with this framework, we classify each sentence and adjective-noun pair into one of several pre-defined categories

"The price was great and the staff was excellent"

- Eleven categories were identified based on:
  - most common topics discussed in the reviews (blue)
  - main categories in the TripAdvisor rating system (green)

Ambiance Amenities Cleanliness Food Jet Setter
Location Property Room Service Sleep Quality Value

- Classification procedure:
  - Create a training set by manually classifying 1500 sentences into topics
  - Train one logistic regression classifier for each topic (11 classifiers total)
  - Since some topics appear more frequently than others, we tested various probability level cutoffs with 10-fold cross validation and selected the cutoff that achieved the best F1 score for each topic



Classification performance by topic:

Topic	Ambiance	Amenities	Cleanliness	Food	Location	Property	Room	Service	Sleep Quality
F1 Score	0.55	0.6	0.67	0.71	0.53	0.66	0.59	0.65	0.65

Sentiment Analysis

- Goal: use the adjective-noun pairs in each topic to evaluate the general sentiment of the topic

Pair: "beautiful views" Sentiment: 0.725
⇨ 72.5% probability that the text is positive

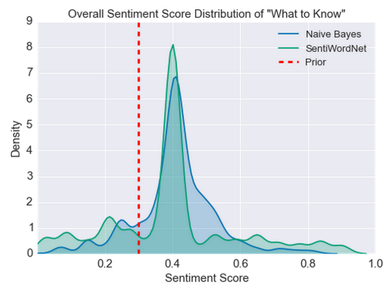
Pair: "not attentive staff" Sentiment: 0.339
⇨ 33.9% probability that the text is positive

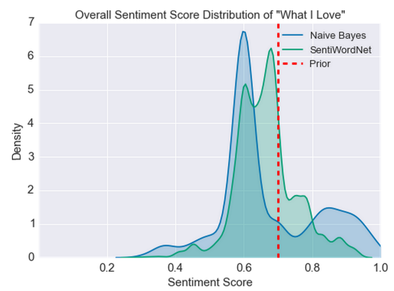
- Procedure:
  1. Calculate the sentiment score of each pair using 3 different approaches
  2. Assume a Bayesian model to combine the results
  3. Average the sentiment scores over all pairs in a topic
- Prior
  - Each review is a response to one of two prompts: "what I loved" or "what to know"
  - We assume reviews from "what I loved" are positive with 0.7 probability
  - We assume reviews from "what to know" are positive with 0.3 probability
- Observation 1: Unsupervised Learning
  - Look up each word in SentiWordnet: a database of words with multiple definitions of each word, and the manually labeled sentiment of each definition
  - To determine the correct definition of the word given the context, we use the Lesk algorithm, a popular algorithm for word sense disambiguation
- Observation 2: Supervised Learning
  - Train a Naïve Bayes classifier on sentiment-labeled movie review data

Assuming a Normal-Normal Bayesian model, the posterior is calculated as:

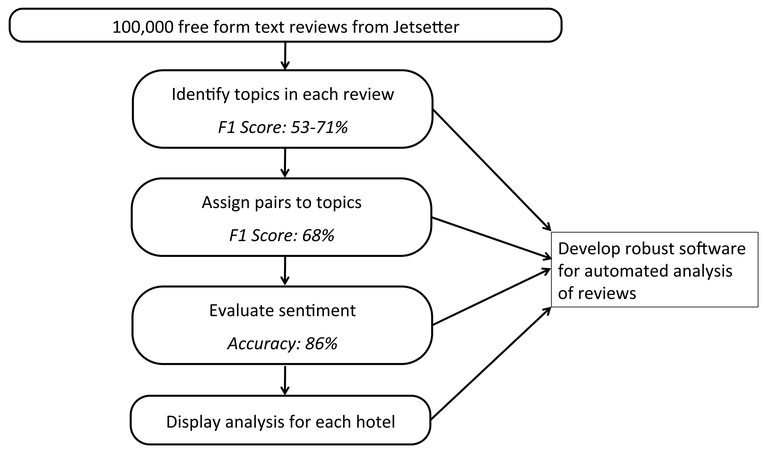
$$\frac{\sigma^2_{prior}\mu_{prior} + 2\tau^2_{obs}\mu_{obs}}{\sigma^2_{prior} + 2\tau^2_{obs}}$$

Under the assumption of equal variance, this simplifies to the average of the prior and observations.





Summary



Ongoing work: account for edge cases to improve metrics

Acknowledgements

We thank Ryan Amari, Bryan Balin, and Chris Stasonis for providing the data and for their support.