Dirty Comments and Clean Plates

An analysis of Yelp comments and their (possible) effect on restaurant inspections

Claire Boyd, Raul Castellanos, Jack Gibson, Benjamin Leiva

Ever see one of these notices? Gross!

“The Chicago Department of Public Health closed **Seoul Taco**, 1321 E. 57th St., on Nov. 14 for the restaurant's continued noncompliance in controlling its infestation with mice.

A recent inspection found 140 droppings in the restaurant…” (Hyde Park Herald)

Who is handling this?

Public health departments across the country often conduct **yearly health inspections** on food facilities, focusing on preventing foodborne illnesses and educating operators on proper food handling.

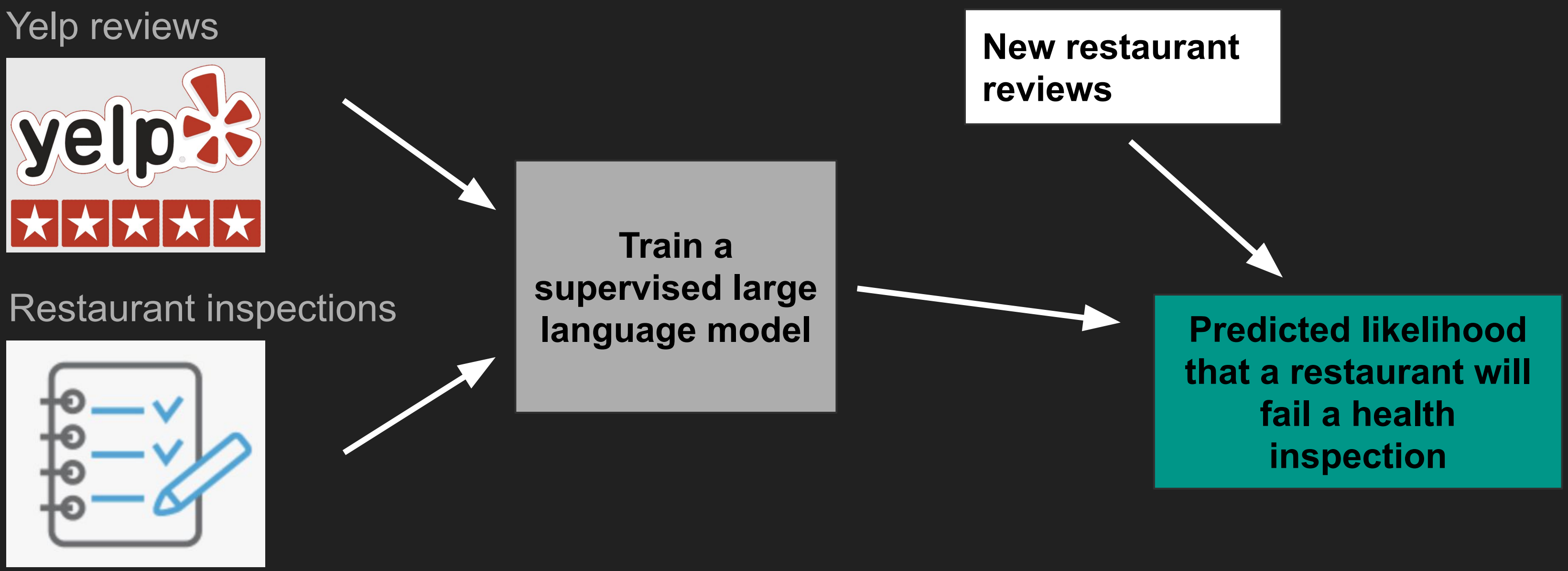
But, these departments are often **under-resourced and are unable to conduct inspections of all restaurants in their jurisdictions**. They often have to randomly select a sample of restaurants to inspect, but this approach has a few issues:

- The random assignment of inspections **is inefficient** when taking into account commute and shift schedules, the conditions of a restaurant can change drastically from the time of a complaint to time of inspection (Hutton)
- Health inspectors complete only 3 to 4 inspections a day on average, with a single inspection taking **anywhere from an hour to several hours** depending on the conditions (Krishna)

Using consumer reviews to flag restaurants potentially in violation of health code makes **this process more efficient** by acting as a filter for the random inspection system. (Hutton)

Our task

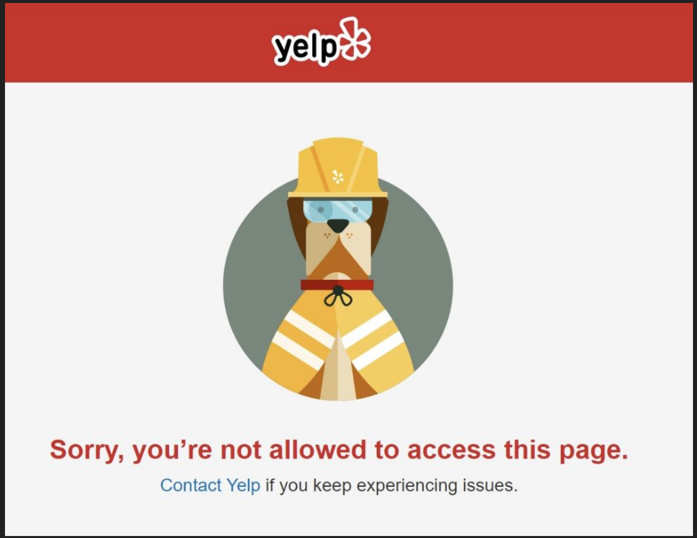
Use a corpus of **yelp reviews** of restaurants that have had health inspections to train a model to classify if a restaurant is likely to fail a health inspection



Data Collection

Early Data Collection Challenges

- Our initial goal was to crawl Yelp to get reviews for a representative sample of restaurants in Chicago
- However….we got blocked by Yelp ;(
- So…we pivoted to Philadelphia!



Available Data

Yelp reviews

- Yelp is a social networking that connects consumers with local businesses.
- The Yelp **academic dataset** is a subset of Yelp businesses, reviews, and user data for use in connection with academic research.
    - Businesses: 150,346 rows (attributes, categories, review count)
    - Reviews: 6,990,280 rows (text, stars, useful, funny, cool, date)
    - Users: 1,987,897 rows (# friends, # fans, average stars, +more)

Restaurant inspections

- **Region:** Pennsylvania
- 112,000 inspections (organization name, location, inspection reason, overall compliance)

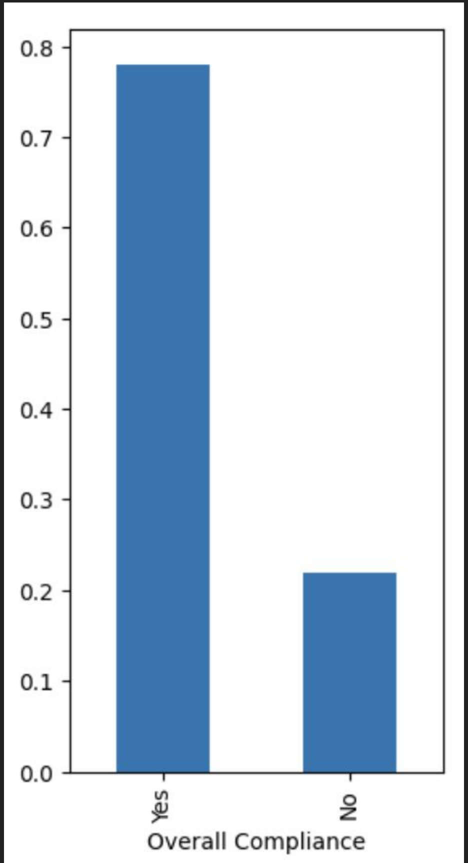
What do we have to work with?

After merging inspections with relevant reviews, we were left with the following:

2,165 inspections, for **1,057** unique restaurants with an average of **7.8 reviews per restaurant** *(16,897 total reviews)*

Outcome variable: "Overall Compliance" with health guidelines: Yes *(78%)*, No *(22%)*

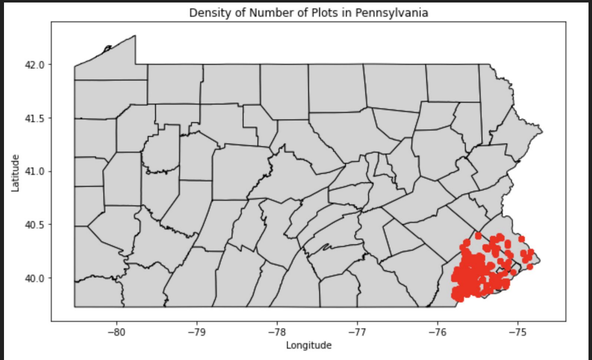
Train (80%, 1754), Val (10%, 216),Test (10%, 202)

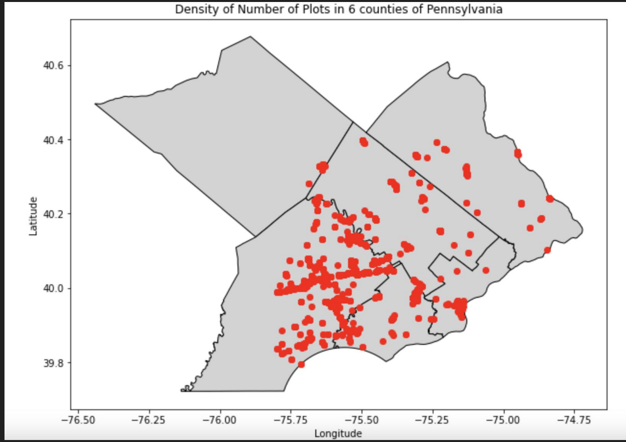


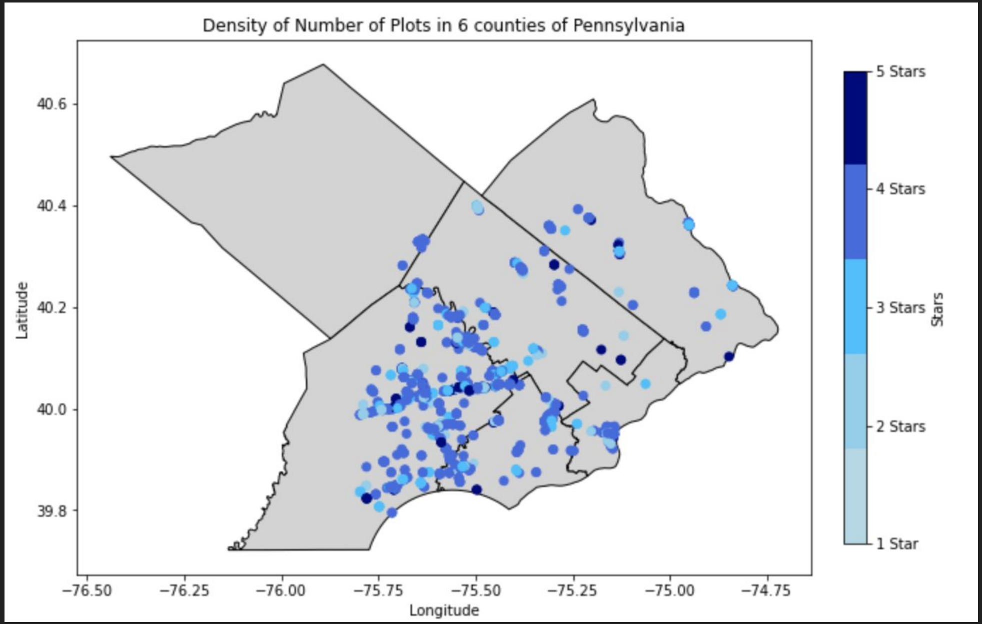
There are 10 **different types** of health inspections in Philadelphia.

Inspection Reason Type	Inspections	% Passed	% Failed
All Inspections	***2165***	***78.06***	***21.94***
Regular	1649	75.01	24.98
Follow-up	311	88.42	11.57
Opening	76	96.05	3.947
Complaint	69	82.60	17.39
Change of Owner	40	77.5	22.50
Emergency Response	10	80	20
2nd Follow Up	6	100	0
Type 2 Follow-up	2	100	0
Foodborne Investigation	1	0	100

Where is our data collected from?







Example Reviews

"DONT EAT THERE !!! Food taste awful and over priced !! They will RIP YOU OFF !! I ordered chicken masalah with little extra sauce..he charge me 21.95 for one meal !! NO SIDES OR BREAD !! and when I questioned him he was soooo RUDE and kept arguing with me so I just left !! I definitely don't recommend going there !!!"

"Incredible Indian food. It's the best in the area. The owners are extremely nice and the service is great.",

"This is my favorite Indian restaurant in my area. The food here is excellent and made to your desired heat and spiciness level. Very casual atmosphere and a friendly wait staff that makes the dining experience so much better. Try the tandoori bone in chicken, any of the lamb curries, samosa and don't forget to order bread Naan, roti or the puffy puri. The owners here are great people with cute kids. A good family run establishment",

[Aman's Indian Bistro]

"Our first visit here and probably won't be back. Everything had a very unusual taste. The fried cauliflower was sour as though it was seasoned with vinegar. I ordered nachos....also sour. The chicken was peppery. Not my fav.",

"I would give it 3 stars for food and 4 for drinks. I went for happy hour and the drinks were cheap and yummy. The food was okay but nothing special. It was packed during happy hour and there was one bartender. She held her own but that's a lot of people for just her. Also, after it died down I was just hanging with friends and you had to flag down someone for just water even though it was empty and not much to do. I definitely would come back for drinks and even food but I would never pay full price."

'We had a great burger here. Juicy, thick and wonderfully tasty. Nice atmosphere. Attentive staff. Appetizers were so yummy! Cheese curds and nachos were satisfying and filling. Will return !'

'Terrible food that is overpriced and takes forever to come out. There are several better options in west chester than this.'

[Stove & Tap]

Example Reviews

"DONT EAT THERE !!! Food taste awful and over priced !! They will RIP YOU OFF !! I ordered chicken masalah with little extra sauce..he charge me 21.95 for one meal !! NO SIDES OR BREAD !! and when I questioned him he was soooo RUDE and kept arguing with me so I just left !! I definitely don't recommend going there !!!"

"Incredible Indian food. It's the best in the area. The owners are extremely nice and the service is great.",

"This is my favorite Indian restaurant in my area. The food here is excellent and made to your desired heat and spiciness level. Very casual atmosphere and a friendly wait staff that makes the dining experience so much better. Try the tandoori bone in chicken, any of the lamb curries, samosa and don't forget to order bread Naan, roti or the puffy puri. The owners here are great people with cute kids. A good family run establishment",

Aman's Indian Bistro

Reviews before a failed inspection 👎

"Our first visit here and probably won't be back. Everything had a very unusual taste. The fried cauliflower was sour as though it was seasoned with vinegar. I ordered nachos....also sour. The chicken was peppery. Not my fav.",

"I would give it 3 stars for food and 4 for drinks. I went for happy hour and the drinks were cheap and yummy. The food was okay but nothing special. It was packed during happy hour and there was one bartender. She held her own but that's a lot of people for just her. Also, after it died down I was just hanging with friends and you had to flag down someone for just water even though it was empty and not much to do. I definitely would come back for drinks and even food but I would never pay full price."

'We had a great burger here. Juicy, thick and wonderfully tasty. Nice atmosphere. Attentive staff. Appetizers were so yummy! Cheese curds and nachos were satisfying and filling. Will return !'

'Terrible food that is overpriced and takes forever to come out. There are several better options in west chester than this.'

Stove & Tap

Reviews before a passed inspection 👍

Modeling Approaches

Data Transformations

Data cleaning: standard cleaning (splitting by word, removing newlines, removing alphanumeric characters, punctuation, etc.) and removing stopwords

Feature engineering: created categoricals of characteristics we thought would be important to the final prediction including:

- **Location of restaurant** (boolean): "Chester", "Bucks", "Philadelphia", "Delaware", "Montgomery", "Berks"
- **Review characteristics** (numeric): "stars", "review_count", "is_open", "n_reviews", "avg_rating"
- **Inspection characteristics** (boolean): Regular, Follow_up, Other
- **Restaurant characteristics** (boolean): 25 top cuisine types (e.g. “Mexican”, “Sushi”, “Nightlife”, “Juice Bars & Smoothies”, “Cheesesteaks”)

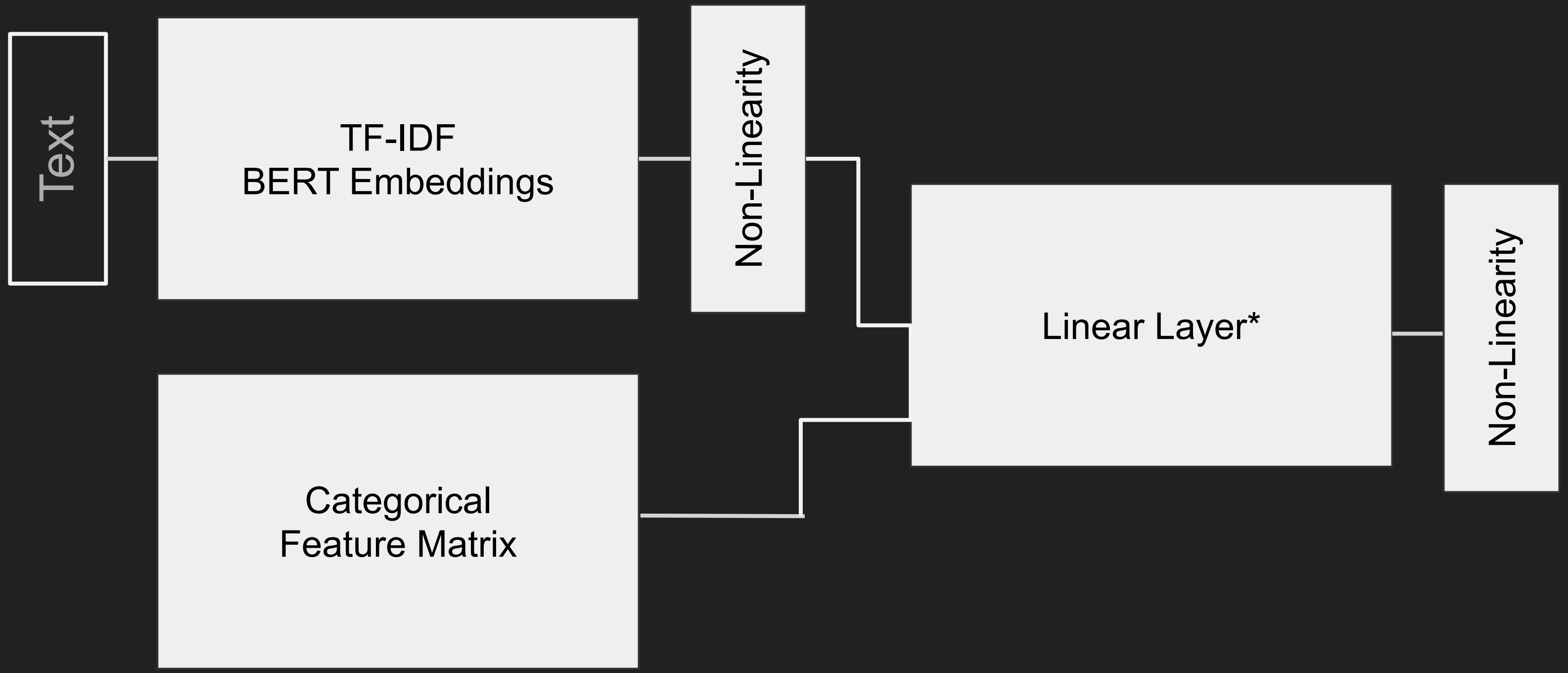
Weighted sampling (tried re-sampling “no”s to get a 50/50 split, but got worse results)

Unit of observation: Review, inspection period

All models

Neural net approaches	Transformer-based language learning models
Logistic Regression (text only)	distilBERT (text only)
Logistic Regression (text, features)	distilBERT (text, features)
SVM (text only)	

Merging Categorical + Text Features



*Embeddings and feature matrices appended together

Results

	By inspection period			By review		
Model	accuracy	f1	recall	accuracy	f1	recall
Logistic Regression (text only)	0.745	0.035	0.018	0.759	0.000	0.000
Logistic Regression (text and features)	0.718	0.032	0.018	0.754	0.016	0.008
Support Vector Machines (text only)	0.741	0.000	0.000	0.759	0.000	0.000
Fine-tuned BERT (text only)	0.806	0.892	1.000	0.786	0.880	1.000
Fine-tuned BERT (text and features)	0.800	0.622	0.539	0.562	0.716	0.702

Majority class label accuracy: 0.74074

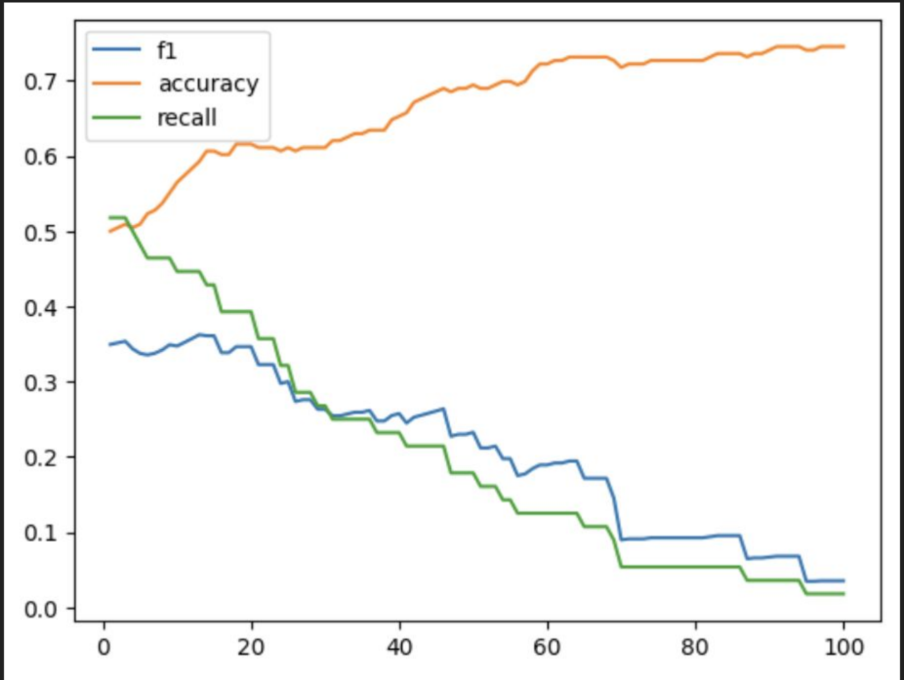
Limitations

Dataset is too small to learn anything substantive

Even with categorical features, the model didn’t learn enough to beat the majority class label.

…let’s try something else!

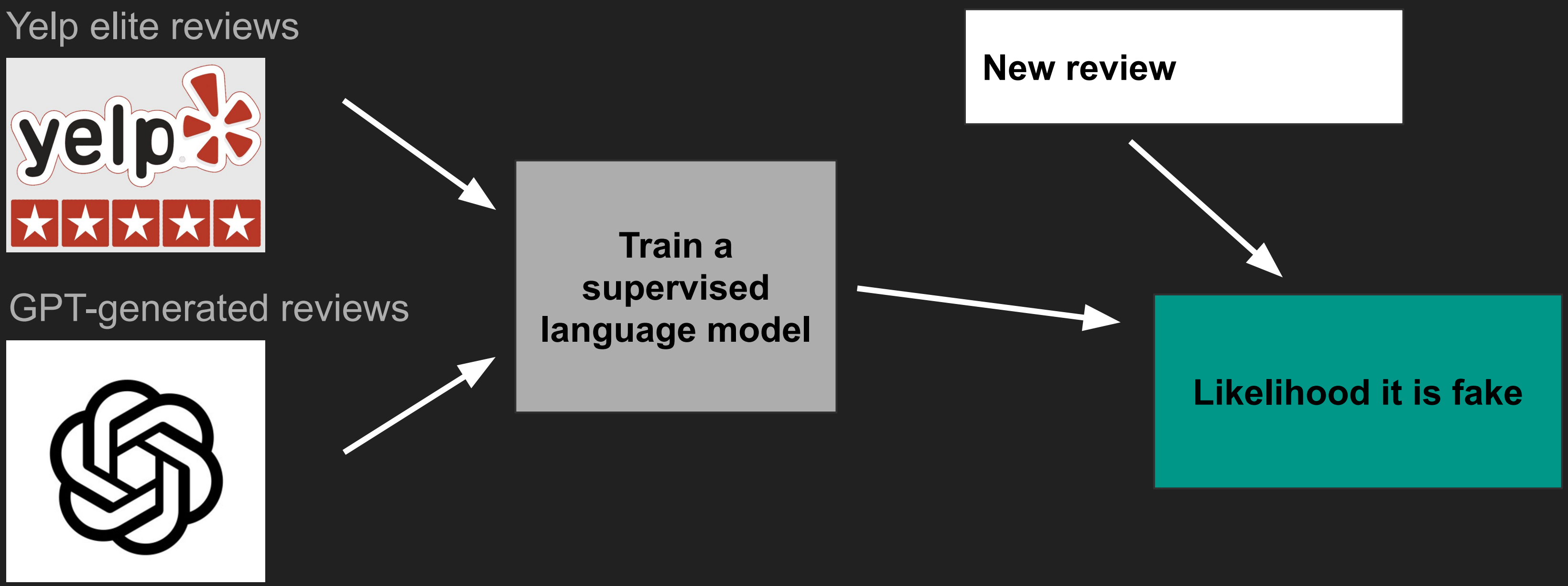
Logistic Regression (only text)



Classifying Fake Reviews

Our NEW task

Create a labeled dataset of fake reviews with GPT and real reviews with yelp to train a classification model on fake reviews.



Generating labeled dataset

Yelp elite reviews

Use all yelp reviews by “elite” members (verified as real people by the app) as labeled “real” reviews

GPT-generated reviews

Use GPT 3.5, 4.0 to generate fake reviews with zero-shot and few-shot prompts

Prompt Engineering

Base prompt:

- You attended a restaurant named REST_NAME.
- You rated your experience with NUM_STARS stars out of 5.
- Write a review of NUM_CHAR characters describing your experience.

Input	**Generation**
Name of restaurant (REST_NAME)	Simple random sampling from verified dataset
Rating (NUM_STARS)	Non-uniform random sampling (used empirical distrib. probs)
Review length (NUM_CHAR)	Simple random sampling from verified dataset

Prompt Engineering, zero-shot

Example #1:

- You attended a restaurant named Lerua's Fine Mexican Food.
- You rated your experience with 4 stars out of 5.
- Write a review of 1248 characters describing your experience.

Example #2:

- You attended a restaurant named Legal Sea Foods.
- You rated your experience with 1 stars out of 5.
- Write a review of 754 characters describing your experience.

Prompt Engineering, few-shot

You attended a restaurant named Broad Ripple Brewpub.
You rated your experience with 4 stars out of 5.
Write a review of 258 characters describing your experience.

Example 1: Name = Broad Ripple Brewpub; Number of characters = 640; Number of stars = 3.0.

- **Review:** Upon walking into this dimly lit pub, you wouldn't necessarily get an accurate idea of what the menu is made up of. Sure, there are classic pub staples like fish and chips, but alongside that there is a vegan counterpart! This place is truly a vegetarian/vegan haven, boasting an exciting and expansive menu geared towards this demographic. One major downfall is service, however. Come here expecting slow, and often rude, interactions with your server. Unless you happen to be lucky enough to get Kristi as your server-then you're in luck. Oh, and keep an eye out for their jalapeno beer-after trying it just once, I think about it often!

Example 2: Name = Broad Ripple Brewpub; Number of characters = 228; Number of stars = 4.0

- **Review**: love the laid back vibe and the patio that faces the monon. they have several good things. i like the drunken ravioli, pub ploughmans platter, and beer crock cheese. my kiddos love their chicken finger meals and everything is reasonably priced. wait staff is always great.

Fake Reviews

API call specifications:

- Temperature: 1.0
- Number of shots: 0, 1-3
- Models: GPT-3.5, GPT-4
- System role: “You are a person that writes restaurant reviews”

Examples:

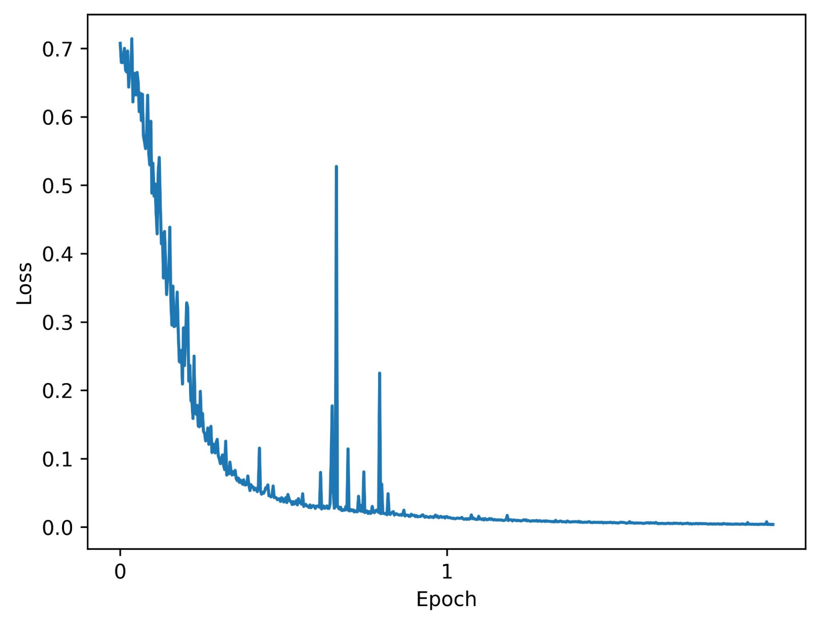
- Review: Loved my experience at Jimmy J's Cafe! The cozy atmosphere and delicious food made it worth the wait. Highlights were the blueberry brandy with Brie cheese french toast and the unique shrimp & garlic aioli french toast. Can't wait to come back! #Foodie #Deliciousness

- Disappointing experience at Baby Blues BBQ Philly. Food was average and service was lacking. Uninspiring flavors and slow service left me unimpressed. Would not recommend. ⭐⭐

Results

	Per review		
Model	accuracy	f1	recall
RNN	0.792	0.841	0.990
Fine-tuned BERT (text only)	0.998	0.998	0.997

Majority class label accuracy: ~0.5



Fined-tuned BERT loss per epoch

BERT Classification

Source	Review	Human	GPT
Elite Yelp User	“Drinks were ok but pricey for what you get. Menu was rather interesting but place is next door to a pizza restaraunt we favor. Nice location but better other options in the area.”	0.9974	0.0032
gpt-3.5	“I had an incredible dining experience at Snarf’s Sandwiches Skinker From start to finish everything was absolutely perfect. The staff was friendly and welcoming making me feel right at home. The menu had a fantastic variety of sandwiches to choose from and I went with the Italian combo mouthwatering combination of meats cheeses and veggies. The sandwich was generously sized and packed with flavor…”	0.0029	0.9968

What did we learn from this process?

- Don’t scrape from yelp ;)
- Ensure that you have a robust enough dataset before embarking on a modeling strategy
- Reviews aren’t as predictive as we thought - makes sense because individual preferences vary widely

Questions?

Thank you

Bibliography

- Predicting Restaurant Health Violations Using Yelp Reviews: A Machine Learning Approach https://www.harlanhutton.com/yelp.pdf
- Yelp Data https://www.yelp.com/dataset
- Open Data Pennsylvania https://data.pa.gov/Public-Safety/Public-Food-Inspections-last-24-months-County-Agri/etb6-jzdg/about_data

Neural Net Model Architecture

Preprocessing

Text Processing & Embedding:

- Removed newlines, used regex to remove any non-letter or number
- **Used a TF-IDF Vectorizer** to capture the relative importance and frequency of unigrams and bigrams across the corpus of reviews

Sampling:

- WeightRandomSampler (tested using a sampler to account for the class imbalance, but this approach generated worse results)

Batch-Size: 16

Model Architecture

- Logistic Regression (text only):
    - Linear layer (2000, 2)
    - Sigmoid (2)
    - Output (2)
- Logistic Regression (with features)
    - Linear layer (2000, 2)
    - Sigmoid (2)
    - Concatenate with features
    - Linear Layer (features+2, 2)
    - Sigmoid
- SVM
    - Linear layer (2000, 2)

RNN Architecture

Preprocessing	Model Architecture

Text Embeddings:

- Glove embeddings, dim=300
- Sequence length: 150, truncating if longer, padding if shorter

Hyperparameters

- Hidden layer size: 256
- Batch size: 32
- Epochs: 25
- Learning rate: 1-e3
- Dropout: 0.5
- Weight hidden layer initialization: 0.0

Steps (per batch)

- Input (150x300)
- LSTM (300, 256)
- Linear (hidden, 2)
- Sigmoid (2)

BERT Model Architecture

Preprocessing

Text Embeddings:

- Encoder model
- Uses first 512 tokens of each review to get embedding matrix (pads if it’s not that long, truncates if it is over that length)
- Dimensions: batch size X 512

Sampling:

- [None]

Training Batch-Size of 16, Validation/Testing Batch-Size of 8

Model Architecture

- Text only:
  - Input Layer (batch_size, 512)
  - BERT layer(batch_size, 2)
  - Sigmoid
- Text + Features
  - Input layer (batch_size, 512)
  - BERT layer (batch_size, 2)
  - Sigmoid
  - Linear layer (batch_size, 2 + n-features)
  - Sigmoid