CAPP 30255 Project Final Report - FreshDash

Using Yelp reviews to optimize restaurant inspection

Ran Bi, Shambhavi Mohan, Minjia Zhu

1. Summary

This project proposes a machine learning system to assistant public health agencies to optimize food establishment inspection leveraging on information gathered from the social media. In particular, we apply natural language processing techniques to restaurant reviews from Yelp to identify restaurants with high risk of hygiene code violation. We generate two types of text-based features sets, i.e. keywords obtained from ngram, and topic distribution obtained from topic modeling, combine them with business and inspection record features, and use them to train four classification models. Our empirical analysis shows that these review text-based features can predict hygiene code violation with 60.1% accuracy with Ada Boosting model, which is 2% improvement over baseline results. To note that our baseline model is a simple machine learning model excluding text features.

2. Problem Description and Related Works

2.1 Motivation

According to the Centers for Disease Control, more than 48 million Americans become sick, 128,000 are hospitalized, and 3,000 die from food related issues per year. An estimated 75% of the outbreaks came from food prepared by caterers, delis, and restaurants. Such illness is prevented by inspection and surveillance conducted by health departments across America. However, the surveillance is insufficient due to lack of inspectors and inefficient inspection process. In Chicago, for example, three dozen inspectors are responsible for over 15,000 food establishments across the city. Moreover, some cities conduct annual inspection only, which only captures a small window of time and cannot reflect the actual operations of food establishments all year round.

Therefore, it is critical to accurately identify restaurants at high risk of code violation and inspect accordingly. Novel approaches such as mining social media data and predictive analytics are gaining momentum to enhance public health surveillance. For instance, the City of Boston held competition to use Yelp reviews to model hygiene violations. The City of Chicago is pioneering a prediction model based on inspection records and 311 requests. Building upon those successful attempts, we seek to apply NLP techniques on Yelp review, together with other known predictors of food inspection outcomes, to optimize the prediction model and optimize inspection efforts.

2.1 Related Works

Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews (Kang, Kuznetsova, Luca, Choi)

It is one the first works to use reviews as a predictive tool for assessing hygiene of restaurants. It uses SVM and SVR to predict restaurants with high unhygene scores (inspection health score >=50). One of the best things about the paper is that it uses first only individual features such as previous inspection scores, review counts, average rating, etc and shows that including review sentiment improves the prediction score to 82%.

Using Yelp Data to Predict Restaurant Closure (Michail Alifierakis)

In this blog report, the author uses yelp data to predict the probability of a restaurant closure. The author used both the yelp api and the google api to get details of all the restaurants that were open/closed as of 2013. The machine learning model used for this task was logistic regression optimised for precision of open restaurants. The final precision of open restaurants was 91%. This means that among the restaurants that are recognized as open by the model, 91% of them actually remained open. The remaining 9% are false positives. However, the model was not very good at predicting closure of restaurants. One of the suggestions was to use health inspection scores to improve the accuracy of predictions, which further cemented our decision to use both the yelp reviews and other features like previous inspection records to predict probability of a restaurant being a high health inspection scorer.

Predicting Restaurant Health Inspection Penalty Score from Yelp Reviews (Uppoor, Balakrishna)

This paper penalty score as a sum of minor, major and severe violators. This paper is also inspired by *Kang, Kuznetsov,etc* paper and uses ridge regression to come up with the best model to predict restaurant health inspection penalty score. Few things that we are going to use from this paper include the size of the word will be directly proportional to the the Ridge regression coefficient for that word. Hence, larger the word, higher the penalty. We will also consider global average penalty score as our base case.

3. Data

3.1 Data Source

We use the datasets that integrate restaurant information on Yelp with health inspection result in Seattle, and further combine it with Food Establishment Inspection Record from open data portal.

Yelp dataset in Seattle is obtained from paper *Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews*. (available at http://cs.stonybrook.edu/~junkang/hygiene). The dataset contains Yelp reviews written for restaurants in Seattle over the period of 2006 to 2013, and public inspection records of the same period. The dataset contains more than 1,200 businesses, 13,299 inspections, and 162,310 reviews.

Food establishment inspection record is obtained from King County Open Data Portal (https://data.kingcounty.gov/Health-Wellness/Food-Establishment-Inspection-Data/f29f-zza5). There are 265,340 rows of inspection record over the period of 2006 to 2013. The dataset contains detailed inspection information, including description, program identifier, inspection results, violation type, etc. After integrating the two datasets, there are 7,705 unique inspection instances.

3.2 Data Exploration

Following are exploratory analysis on the Seattle dataset.

- Limited coverage of hygiene inspections geographically and seasonally. Geographically, more than 50% of the restaurants listed under Yelp did not have inspection records. The heatmap [Figure 1] of inspections also suggests unbalanced inspection efforts. Seasonally, though inspection instances increased from 700 in 2006 to 3,175 in 2012 [Figure 2], the pattern of inspection shows consistent seasonality over years – inspections were the most frequent in the first quarter, and the least frequent in the third quarter. [Figure 3]

- Our dataset is skewed as most of the restaurants have 0 code violation (36%). As of 2012, only 47% of the restaurants inspected had high violations. [Figure 4]

- Review rating is not a good predictor of hygiene code violation. Following the convention in Kang et al. paper, we define an "inspection period" of each inspection record as the period of time starting from the day after the previous inspection to the day of the current inspection. If there is no previous inspection, then the period stretches to the past 6 months in time. Then we aggregate Yelp reviews generated during the inspection period by review counts and average rating. The histograms of review rating for those passed inspection and those with penalty score higher than 50 present very similar pattern. Additionally, we created compared the code violation type vs rating, most of the restaurants are rated 4,5 and hence is not informative.[Figure 5].

- Review count is probably a weak predictor of hygiene code violation. Similar to review rating, we calculate the Spearman's rank correlation coefficient for review count during inspection period and inspection penalty score at different cutoffs. They seem to be positively correlated, with highest coefficient of 0.1 when cutoff is set at 30. [Figure 6] There seems to be weak correlation between violation codes and review counts.

Those observations justify the necessity of our project, i.e. using review content-based features to predict hygiene code violation.

4. Approach

4.1 High Level Overview

Our approach follows four steps:

1. Generate non-text-based features, including: previous penalty score, cuisine type, review length, average review rating, review count.
2. Apply LDA to Yelp review text to generate topic distribution features.
3. Apply N-gram to Yelp review text to generate keywords features.
4. Use text-based features together with other features to train the machine learning model.

4.2 Yelp Review Based Features - Topic Modeling

In order to mine hygiene information from Yelp reviews, we need to discern reviews commenting on health-related topics from other topics that are not useful in predicting hygiene code violation, for instance, cuisine authenticity, service quality, etc. Therefore, we use Latent Dirichlet Allocation (LDA) to achieve this goal by extracting the main latent topics from review texts.

We use review texts of Seattle restaurants as training texts and pre-process them with the following steps:

1. Tokenize the texts and remove punctuations using *gensim.utils.simple_preprocess* module;
2. Remove stop words, create bigrams and lemmatize tokens using gensim and spaCy libraries;
3. Create the corpuses and dictionaries required for LDA model training.

During model training phase, we underwent several iterations on appropriate training set. We started out with extracting topics from reviews with strong indicator of potential hygiene code violation, i.e. reviews with star rating less than 3 and inspection penalty score higher than 60. It turned out that the topic extracted were very close to each other, and it was hard to discern health specific items. We realized that the corpus itself was too narrow to extract meaningful subtopics.

Secondly, we decided to expand the training set to 162,310 reviews all together. The latent topics were well-distributed on the inter-topic distance map [Figure 7], but a close scrutiny of the subtopics revealed that they centered around restaurant type, service quality and comments on the food itself, without a clear subtopic on hygiene complaint.

Thirdly, we segmented reviews based on their star ratings in order to reveal underlying sentiment aspects of the reviews, which would be otherwise lost when all reviews were considered together. After some trial and error on different segmentation rules, we decided to separate all reviews into two groups: (1) positive reviews, i.e. reviews with 5-star rating (49,183 in total); and (2) negative reviews, i.e. reviews with 1-star rating (9,378 in total). Two LDA models were trained on respective review groups. The number of topics (k) was determined experimentally. We trained LDA models using k=20,30, and 40 and used coherence

value to determine the optimal model. Interestingly, number of topics and coherence value correlated negatively for negative group [Figure 8] but positively for the positive group [Figure 9]. We tried k=15 and 20 respectively, and k=20 gave more human-interpretable results. So we used LDA model with k = 20 for positive and negative models.

Finally, we experimented to train LDA model with reviews combined during each inspection period, instead of individual reviews to avoid information leakage when aggregating individual topic distributions back to inspection period. However, the coherence score of such method is the lowest among previous models. After carefully examine the top words in each topic, we found there were too many overlapping words, i.e. the topic excludeness is too low. So we carefully curated a customized stop words (removing high frequency words particular to Yelp context, e.g. restaurant, food, lunch, place, etc.), and use *id2word.filter_extremes* function in gensim to filter out tokens appear in less than 10 documents, or above 50% the documents. The coherence score improved from 0.2893 to 0.4143. We applied the same customized stop words to individual reviews in the second step, and the coherence score improved from 0.4096 to 0.4468. A detailed comparison is presented in **[**Table 1**]**

In addition, we also took advantage of "labels" associated with the review text to experiment LabeledLDA, a simplified topic model for use when documents have labels, and labels correspond to topics. LabeledLDA asserts a one-to-one relationship between labels and topics, and every document is a mixture of the words associated with its labels. Such assumption simplifies inference of LDA, and it is particularly useful when each document is associated with multiple labels. In our case, each review text can be labelled by its rating, its cuisine type and its penalty score. By running labeledLDA, we could figure out whether the review talks more about overall feeling, or about hygiene issues particularly. To simplify the case, we labelled each review text with its star rating, and code_violation (no_violation if penalty score = 0, severe_violation if penalty score > 50, minor_violation if penalty score is in between). The model is trained using mallet package in Java. Details can be found at http://www.mimno.org/articles/labelsandpatterns/

Table 2 and Table 3 show the represented words for selected topics. Table 4 shows the top words using LabeledLDA. We have following observations:

1. It seems that LDA models well capture the negative words in 1-star reviews (e.g. depressing, hygiene, paper_thin, endless, grossly), and positive words in 5-star reviews (e.g. great, friendly, excellent, awesome).

2. In positive reviews, people rarely talk about cleanness of the restaurants, but more about dishes, service quality, and the environment. In negative reviews, people explicitly complain about hygiene of the restaurant.

3. Topics extracted from negative reviews seem less human-interpretable. Though the model successfully extracts hygiene-related topic, more work should be done to improve the distinction among topics beyond using customized stopwords. Figure 10 shows distribution of words for hygiene-related topic from LDA model trained on negative reviews.

4. LabeledLDA results exhibit similar pattern as positive/negative modeling. 5-star reviews center around positive praises, whereas 1-star reviews are mostly about bad service. Moreover, representative words for code violation labels are mainly associated with cuisine type. We identify more Asian food in severe violation labels.

Eventually we obtained 40 features for each review text, among them are 20 subtopics from LDA model trained on the positive reviews, and 20 subtopics from the negative reviews. The values for subtopic features are the percentage contribution of the corresponding subtopic. Then we aggregate reviews to inspection period level by adding subtopics weight, and divide by count of reviews in each inspection period.

4.3 Yelp Review Based Features - Top Frequently Occuring Keywords

To understand how useful Yelp reviews are in predicting restaurant code violations, we get the top occuring keywords in all the reviews.

To better select the keywords, we divided the dataset into good restaurants and bad restaurants based on review rating (>4 or <3, respectively) and code violation (0 and 1 respectively).

To maintain consistency across different review based features, we maintained the same process as that of topic modelling to get lemmatized tokens.

We used the below methods to determine the keywords:

1. Ngrams

   We first tried using unigrams to determine keywords related to good reviews and bad reviews. However, we find that unigrams is unable to take care of cases like “not great”, etc. Hence we tried bigrams which resulted in better results. With bigrams we get sample keyword set for good reviews: ('much', 'better'), ('not', 'bad'), ('one', 'favorite'), ('little', 'bit'), ('late', 'night'), ('will', 'definitely) and for bad reviews as: ('long', 'time'), ('not', 'great'), ('go', 'wrong'). There are few overlaps between the two keyword sets, but we hope that the big sample feature set will help overcome the error caused by the overlap.

2. RAKE (Rapid Automatic Keyword Extraction):

   Rake algorithm removes all stop words from the text and then creates an array of possible keywords. Then we find the frequency of the words and the degree associated with each word (the number of times a word is used by other keywords). Then the degree/frequency of each keyword is calculated and based on this score the best keywords are found. We find the words found using this method to be good too and we will create a set of features using RAKE too.

Since both the methods are just different ways of getting top keywords, we chose bigrams mainly because it is more standardized way of getting common keywords. We chose the top 100 keywords in both restaurants with bad rating and top 100 keywords with restaurants with good ratings.

The biggest problem is that many of these keywords are common, however it is possible that restaurants can have reviews with similar words and we want to capture this too.

Hence we created three sets of ngram features. The first one included top 200 bigrams that included bigrams that were common across both the hygenic and the non hygienic restaurants. For the second nigram feature set, we included only the non hygienic bigrams. In other words, these bigrams were not present in hygienic restaurants. For the third and the final set, we included unique bigrams from both the hygienic and the non hygienic set.

4.4 Machine Learning Model

We built a baseline machine learning model in the first place before including text-based features to filter the power of non-text-based features. Our suggested algorithm is expected to be better than the baseline. Put in the plain words, we want to know how well can predict the inspection violation without Yelp reviews, and how NLP model of Yelp review can facilitate the prediction power of machine learning model.

Features

Non-text-based features include average historical inspection penalty score, last inspection penalty scores, cuisines types, district (based on zip code), review count of the restaurant and average review rating. Keywords generated from N Gram model and topics generated from topic modelling models are our social media review based features.

We created different feature set combinations understand different factors that affect our predictive analysis the most.

Data Labeling

Determine the right label (hygiene and non-hygiene) is vital and disputable due to several reasons. First and most importantly, the inspection penalty score is asymmetrically clustered at low penalty level [Figure 11], which brings the problem of imbalanced data. Setting the threshold at high penalty score leads to much less sample labeled as “non-hygiene” compared to “hygiene” ones. It is also ambiguous to simply cut the threshold of hygiene and non-hygiene by inspection penalty score because penalty score is cumulative: a relative high penalty score may not necessarily indicate indicate severe health issues. For example, a restaurant may violate “proper labeling” or “broken windows” for several times and get a total penalty score of 40. However, these violations seem minorly correlated to health issues and thus are very unlikely to be mentioned by the reviewers.

We bring data from Seattle restaurant inspection report to focus on restaurant with critical violations as they are exactly the set of inspector and customer need to pay attention to. Red critical violations are food handling practices that, when not done properly, are most likely to lead to food-borne illnesses. Blue critical violations are primarily maintenance and sanitation issues that are not likely to be the cause of a foodborne illness. The new labeling depends on the violation code. Therefore, restaurant with Red critical violation code are labeled as non-hygiene, while the rest are labeled as hygiene.

Models and Evaluation Method

We have built a machine learning pipeline that allows us to feed in different parameter tuning grids and multiple machine learning models including Naive Bayes and Logistics Regression, Decision Tree and Adaboost with 5-fold cross validation.

Since we have reduced the size of the dataset and labels by only looking at red violators, we assume there is sufficient inspection capacity to check all restaurants flagged red in our test set of around 2,000 instance. Therefore, we used accuracy as our evaluation metrics.

5. Result Analysis

5.1 Results across different feature sets

We divided our data set to test different hypothesis:

1. We considered only the baseline features (by not including any text features). This model performed well across all models, apart from Naive Bayes. Based on feature importance analysis [figure 12], we can see that previous inspection scores are the most important features.

2. We removed previous inspection scores in our next feature set. Performance dropped from 1% in Naive Bayes to 4% in Ada Boosting. Hence we kept previous inspection scores as a part of our analysis.

3. In analysis of just the words, we saw that certain words associated with cuisine types resulted in bad reviews. So we tested our base model by removing the cuisine type. model performance remained similar to our point 1 for all apart from NAive Bayes, where it improved by 4%. Hence, to avoid bias caused by people's sentiment associated with certain food type, we removed cuisine type from the rest of the testing.

4. So to our baseline (excluding cuisine type), we added topic modelling features.There was a slight improvement across different models. [Table 5]

5. To our baseline (excluding cuisine type), we tried the two types of ngram features: once with only the top unhygienic bigrams and once with all unique bigrams across hygienic and unhygienic features. across different models. [figure 14]

    a. On including the unique bigrams, our accuracy improved to 60.1% for ada boosting.

    b. On including just the bigrams that occur in unhygienic restaurants, the accuracy remained consistent across different models.

    This showed that adding just keywords improves the model performance.

    The most important keywords were food items like noodle soup, indian food, thai food, etc.

6. When we ran all the features (baseline + topic modelling + ngrams) we see a slight drop in performance. This is probably because it is an average of the best models and the worst models. [figure 15]. From the feature importance set we can

5.2 Results across different models

We ran the different feature sets across different models with different parameters. [Table 5]

1. Logistics outperforms Naive Bayes. Probably because we use Gaussian Naive Bayes while have a lot of binary features

2. Ada boosting performs the best across all combinations of the features. This is mostly due to the fact that our dataset is small and hence ensemble methods perform better.

6. Future works

Our best model had an accuracy of only 60%. Hence for future work, we want to explore text analysis using neural networks. Maybe the simple models were not able to capture the inherent characteristics of the reviews and using neural networks might help with that.

Bibliography

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing

Kang, Jun & Kuznetsova, Polina & Luca, Michael & Yejin, Choi. (2013). Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews.. EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference.

Nsoesie, E. O., Kluberg, S. A., and Brownstein, J. S. (2014). Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports. Preventive medicine, 67:264–269.

Schomberg, J. P., Haimson, O. L., Hayes, G. R., & Anton-Culver, H. (2016). Supplementing Public Health Inspection via Social Media. PLoS ONE, 11(3), e0152117.

Appendix

Figure 1. Inspection heatmap 2006-2012

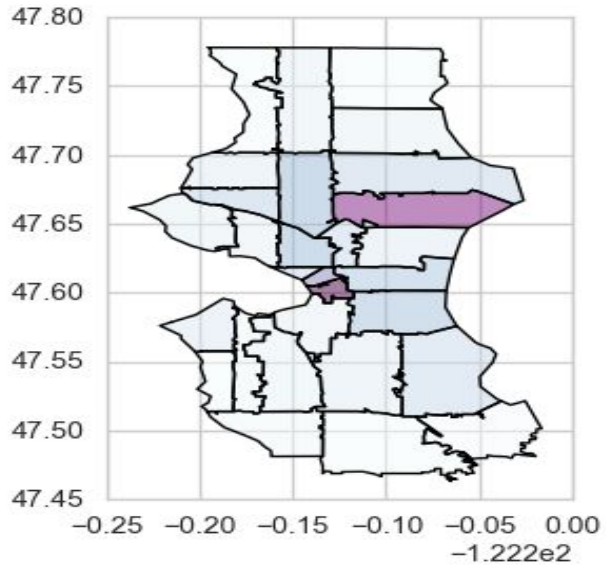

Figure 2. Annual total inspection instances

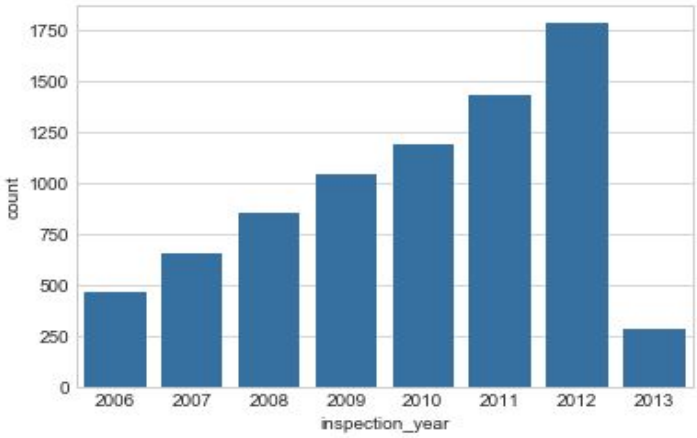

Figure 3. Monthly total inspection instances

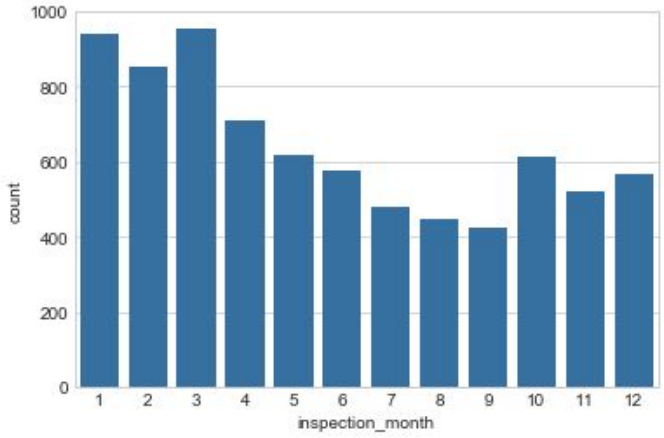

Figure 4. Inspected Restaurants with high code violation

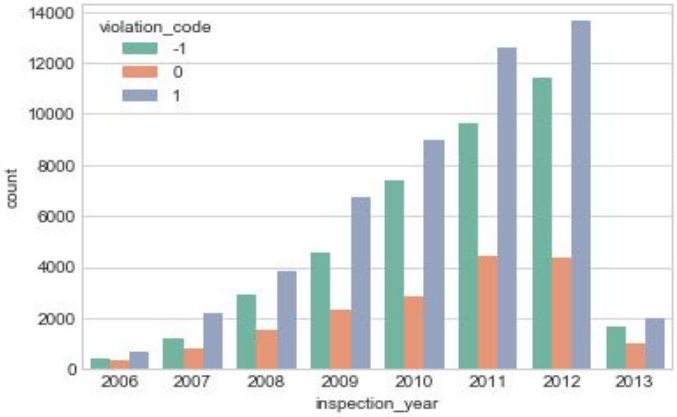

Figure 5. Comparison between review rating and high code violation

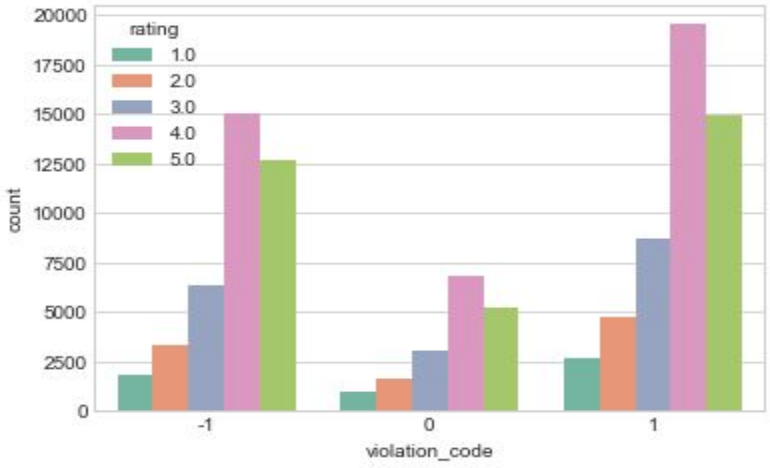

Figure 6: Spearman's coefficient for comparison between penalty scores and review counts

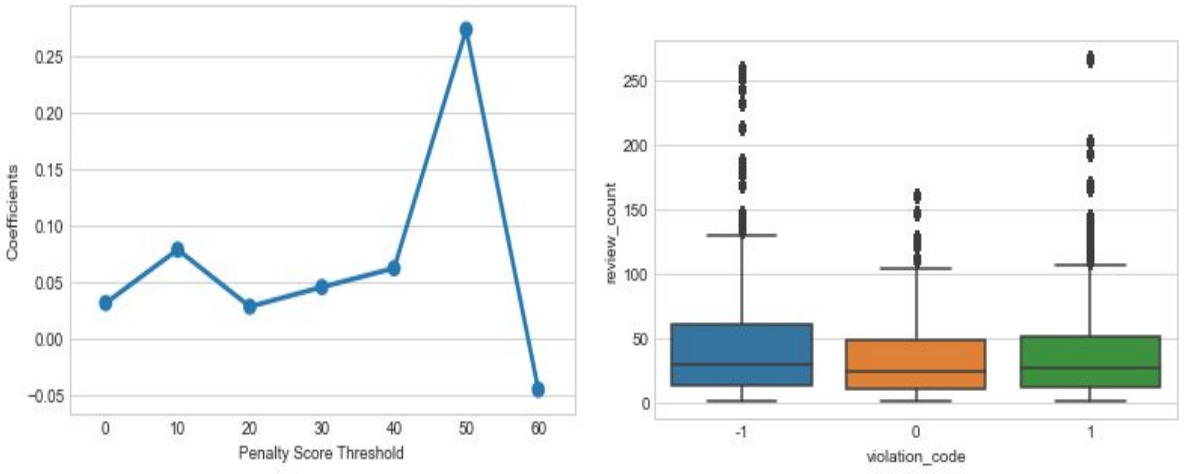

Figure 7. LDA model trained on all reviews together

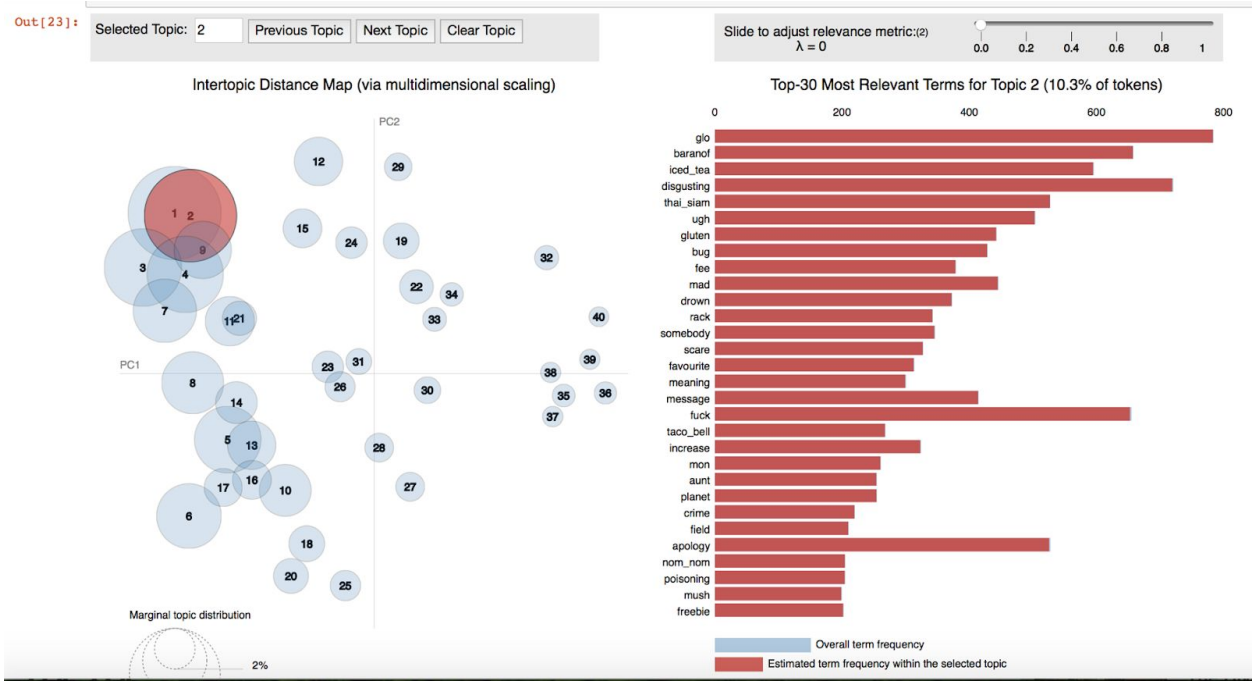

Figure 8 & 9. Number of topics vs Coherence score for LDA model trained on 1-star reviews (left) and 5-star reviews (right)

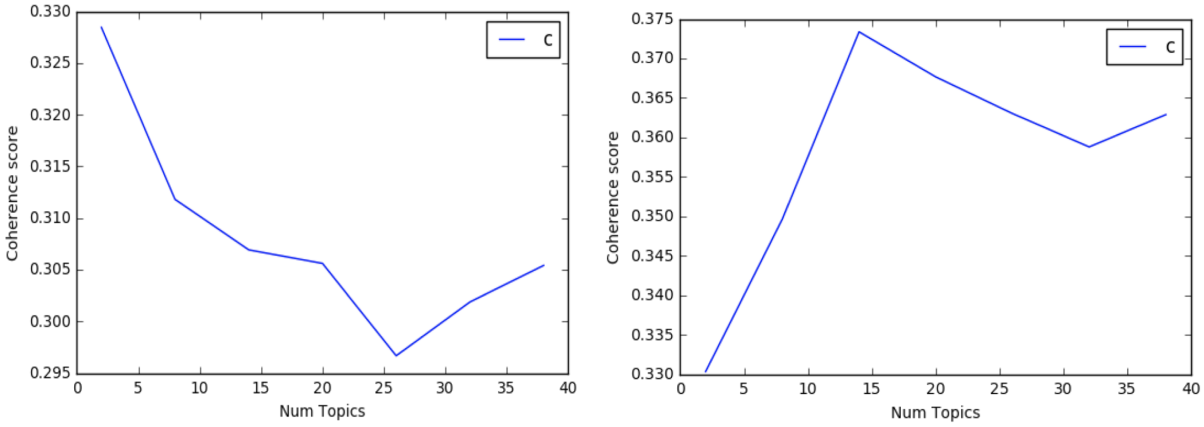

Figure 10. distribution of top words for hygiene topic from LDA model trained on negative reviews

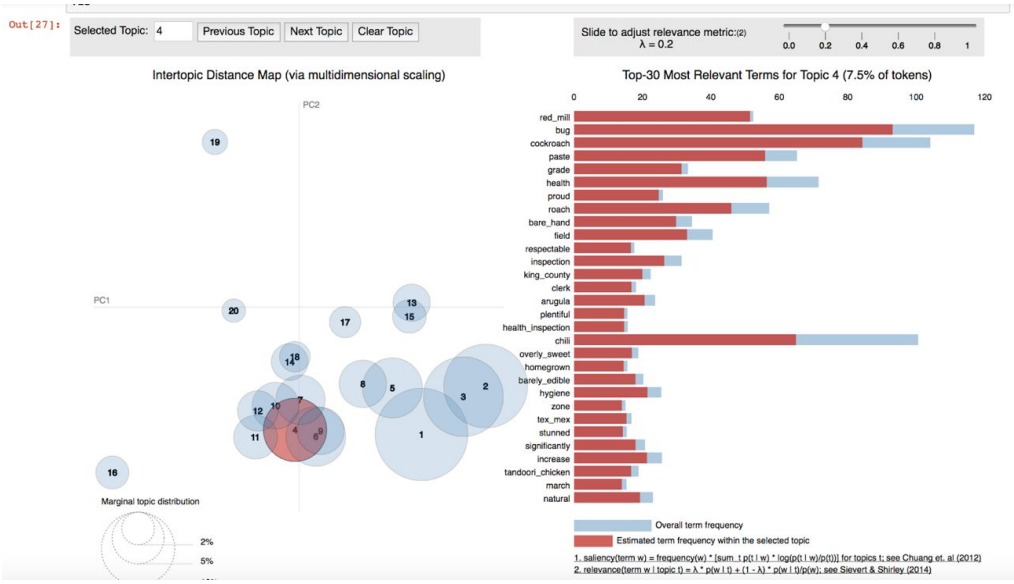

Figure 11. Distribution of inspection Penalty Score

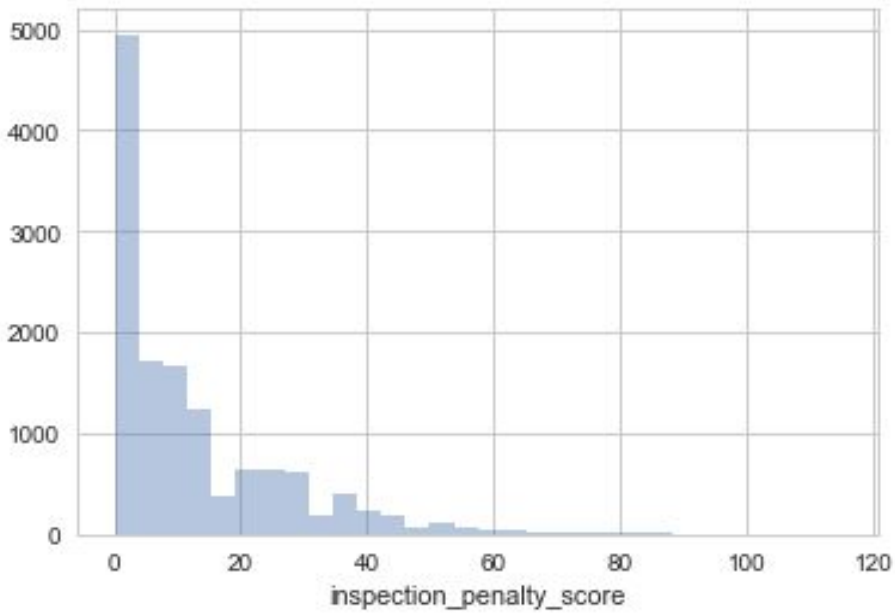

Figure 12. Baseline Feature Importance

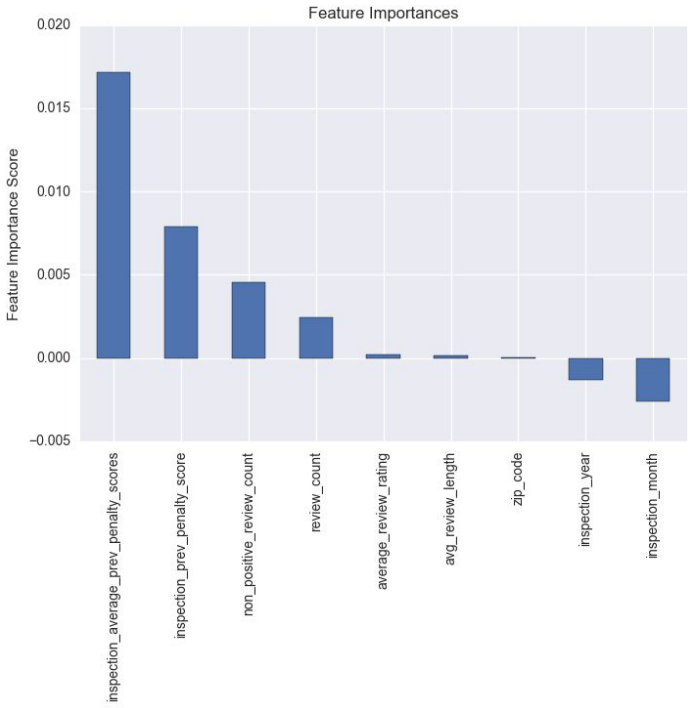

Figure 13: Feature Importance for Baseline + Topic Modelling

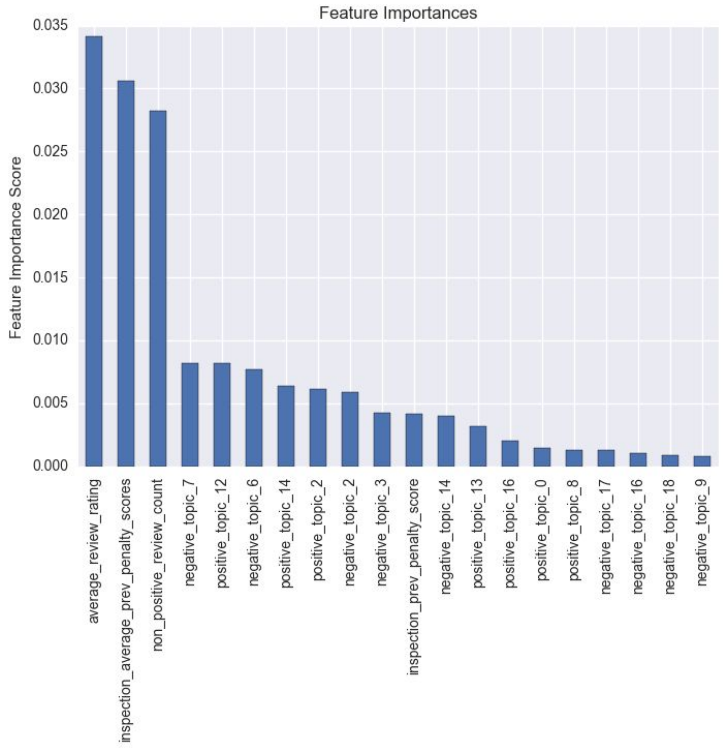

Figure 14: Feature Importance for Baseline + bigrams

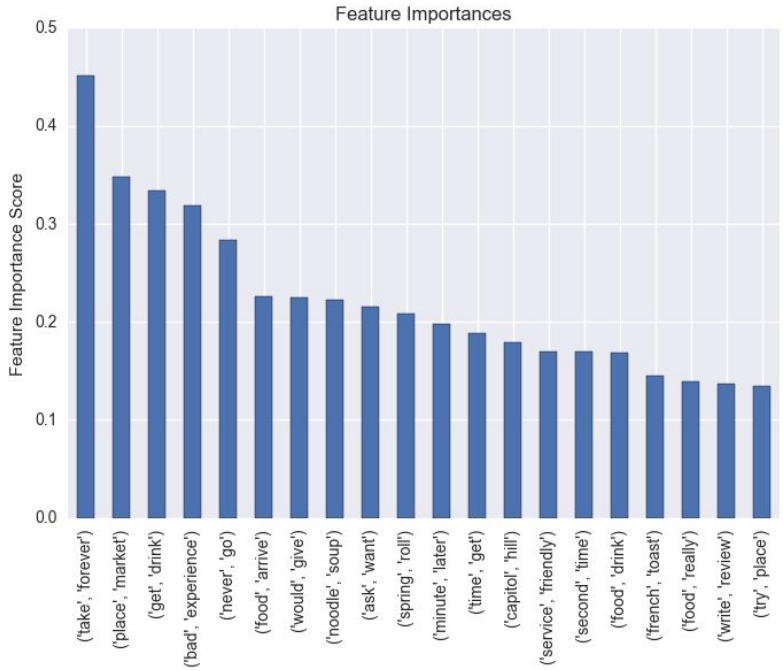

Figure 15: Feature Importance for all features

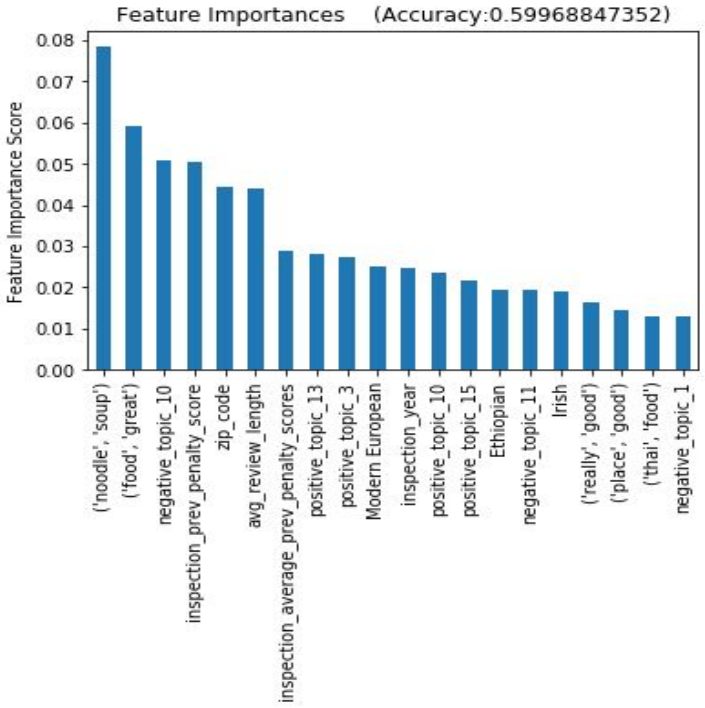

Table 1. Topic Modeling Models Comparison

Input	Customized Stop Words	Number of Topics	Coherence
1-star review	F	20	0.305
individual reviews	F	20	0.409
Merged reviews per inspection period	F	20	0.289
1-star review	T	20	0.324
Merged reviews per inspection period	T	20	0.414

Individual reviews	T	20	0.447

Table 2. Selected topics and top words (λ = 0.2) from LDA model trained on positive reviews

Topic	**Words**
Service	Seat, wait, minute, arrive, waiting, busy, line, seating, pricing…
General Praises	Great, friendly, excellent, awesome, amazing, reasonable, delicious…
Deli Food	Sandwich, burger, deli, turkey, lettuce, vessel, meat
Pub	Beer, pub, hipster, bagel, bar, pretzel, beer_selection, bloody_mary…
Asian Food	Noodle, Chinese, dumpling, raman, rice, katsu, bubble_tea, hot_pot…
Pizza	Pizza, crust, topping, thin, mozzarella, Hawaiian, calzone, slice, pesto
Coffee Shop	Coffee_shop, sweet_tooth, study, espresso, peanut_butter, fried, foam, barista, hospitable…

Table 3. Selected topics and top words (λ = 0.2) from LDA model trained on negative reviews

Hygiene	Bug, cockroach, roach, health, bare_hand, inspection, barely_eadible,…
General Complaint	Mislead, unaware, paper_thin, endless, depressing, grossly, scrub,…
Mexican Food	Mexican, taco, rice, burrito, taco_bell, roti, corn_tortilla,…
Pizza	Pizza, crust, gelato, topping, gluten_free, English_muffin, pie, oven,…

Table 4. Representative words for LabeledLDA results

Label	Representative Words
5-star	amazing sushi wine perfect favorite excellent dessert wonderful coffee worth...
4-star	sushi wine breakfast tasty amazing salmon bread dessert atmosphere perfect...
3-star	sushi stars wine bad breakfast server bread decent tasty price...
2-star	minutes bad server waitress sushi bland waiter left price mediocre ...
1-star	minutes bad waitress server left worst finally waited walked rude terrible horrible ...
no_violation	sandwich burger beer pork tasty rice meat hot thai fries day ...
minor_violation	rice thai sandwich beer hot pork burger tasty meat fried beef pho ...
severe_violation	thai pork rice noodles hot spicy pho cheap vietnamese soup fried curry broth tofu ...

Table 5: Comparison between different feature set and their performance on different models

Feature Set	**NB**	**LR**	**DT**	**AB**
base	0.516095535	0.586708204	0.588785047	0.599688474
base - cuisine	0.557632399	0.585150571	0.576323988	0.595534787
base - prev scores	0.504153686	0.549325026	0.570612669	0.566978193
base - cuisine + uniqe ngrams	0.531152648	0.587746625	0.585150571	0.600726895
base - cuisine + unhygience ngrams	0.534787124	0.586708204	0.591900312	0.595534787
base - cuisine + topic	0.550363448	0.589304258	0.59709242	0.586708204
base - cuisine + topic + uniqe-ngrams	0.528037383	0.587746625	0.585150571	0.598130841

All (base - cuisine + topic + unhygience ngrams)	0.529595016	0.589304258	0.576843198	0.591740721

Table 6: Top occuring topics from topic modelling

Important Topic Model Features	**Representative Words**
negative_topic_10	Mexican, taco, Chinese, bean, burrito, fish, sauce...
negative_topic_11	Pizza, cheese, crust, soggy, slice, bad, thin, dough...
negative_topic_1	Bar, drink, falafel, wait, bad, venue...
negative_topic_19 & positive_topic_10	Breakfast, waffle, egg, cafe, pancake, french toast …
positive_topic_13	Beer, bar, bartender, pub...
positive_topic_3	Drink, , cocktail, bartender, bar, wine, beer, whiskey, student, vodka, tequila...
negative_topic_7	Chicken, burger, fry, pho, dish, sauce, meat, experience...
negative_topic_6	Bad, taste, service, Indian, dry, price, quality, never, lamb, chicken, bland, bread...
positive_topic_12	Always, thai, vegan, vegetarian, friendly, time, family...
positive_topic_14	Burger, onion, delicious, salad, beef, grill, veggie, combo...