Where *Not* to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews

Jun Seok Kang† **Polina Kuznetsova**† **Michael Luca**‡ **Yejin Choi**†
†Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
{junkang,pkuznetsova,ychoi}
@cs.stonybrook.edu

‡Harvard Business School
Soldiers Field Road
Boston, MA 02163
mluca@hbs.edu

Abstract

This paper offers an approach for governments to harness the information contained in social media in order to make public inspections and disclosure more efficient. As a case study, we turn to restaurant hygiene inspections – which are done for restaurants throughout the United States and in most of the world and are a frequently cited example of public inspections and disclosure. We present the first empirical study that shows the viability of statistical models that learn the mapping between textual signals in restaurant reviews and the hygiene inspection records from the Department of Public Health. The learned model achieves over 82% accuracy in discriminating severe offenders from places with no violation, and provides insights into salient cues in reviews that are indicative of the restaurant's sanitary conditions. Our study suggests that public disclosure policy can be improved by mining public opinions from social media to target inspections and to provide alternative forms of disclosure to customers.

1 Introduction

Public health inspection records help customers to be wary of restaurants that have violated health codes. In some counties and cities, e.g., LA, NYC, it is required for restaurants to post their inspection grades at their premises, which have shown to affect the revenue of the business substantially (e.g., Jin and Leslie (2005), Henson et al. (2006)), thereby motivating restaurants to improve their sanitary practice. Other studies have reported correlation between the frequency of unannounced inspections per year, and the average violation scores, confirming the regulatory role of inspections in improving the hygiene quality of the restaurants and decreasing food-borne illness risks (e.g., Jin and Leslie (2003), Jin and Leslie (2009), Filion and Powell (2009), NYC-DoHMH (2012)).

However, one practical challenge in the current inspection system is that the department of health has only limited resources to dispatch inspectors, leaving out a large number of restaurants with unknown hygiene grades. We postulate that online reviews written by the very citizens who have visited those restaurants can serve as a proxy for predicting the likely outcome of the health inspection of any given restaurant. Such a prediction model can complement the current inspection system by enlightening the department of health to make a more informed decision when allocating inspectors, and by guiding customers when choosing restaurants.

Our work shares the spirit of recently emerging studies that explores social media analysis for public health surveillance, in particular, monitoring influenza or food-poisoning outbreaks from microblogs (e.g., Aramaki et al. (2011), Sadilek et al. (2012b), Sadilek et al. (2012a), Sadilek et al. (2013), Lamb et al. (2013), Dredze et al. (2013), von Etter et al. (2010)). However, no prior work has examined the utility of review analysis as a predictive tool for accessing hygiene of restaurants, perhaps because the connection is not entirely conspicuous: after all, customers are neither familiar with inspection codes, nor have the full access to the kitchen, nor have been asked to report on the hygiene aspects of their expe-

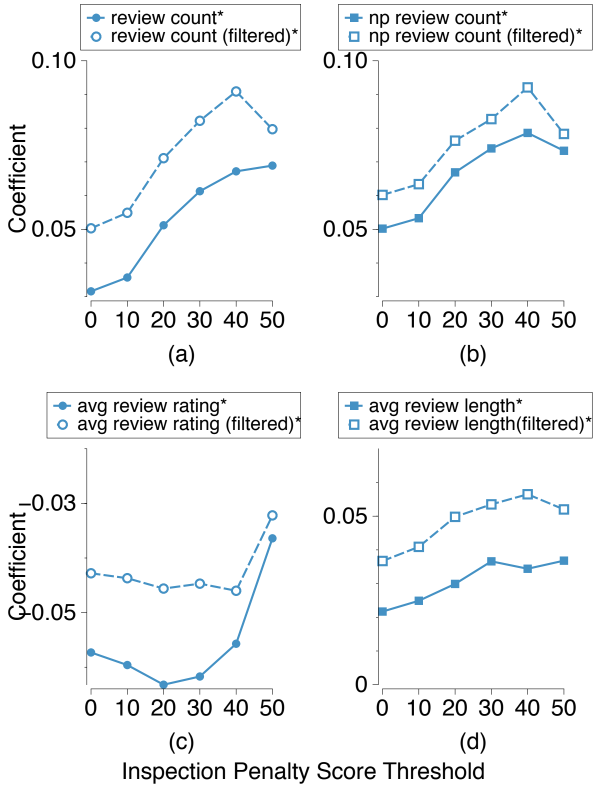

Figure 1: Spearman's coefficients of factors & inspection penalty scores. '*': statistically significant ($p \leq 0.05$)

rience.

In this work, we report the first empirical study demonstrating the utility of review analysis for predicting health inspections, achieving over 82% accuracy in discriminating severe offenders from places with no violation, and find predictive cues in reviews that correlate with the inspection results.

2 Data

We scraped entire reviews written for restaurants in Seattle from Yelp over the period of 2006 to 2013.[1] The inspection records of Seattle is publicly available at `www.datakc.org`. More than 50% of the restaurants listed under Yelp did not have inspection records, implying the limited coverage of inspections. We converted street addresses into canonical forms when matching restaurants between Yelp and inspection database. After integrating reviews with inspection records, we obtained about 13k inspections over 1,756 restaurants with 152k reviews. For each restaurant, there are typically several inspection records. We defined 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. Each inspection period corresponds to an instance in the training or test set. We merge all reviews within an inspection period into one document when creating the feature vector.

[1] Available at `http://www.cs.stonybrook.edu/~junkang/hygiene/`

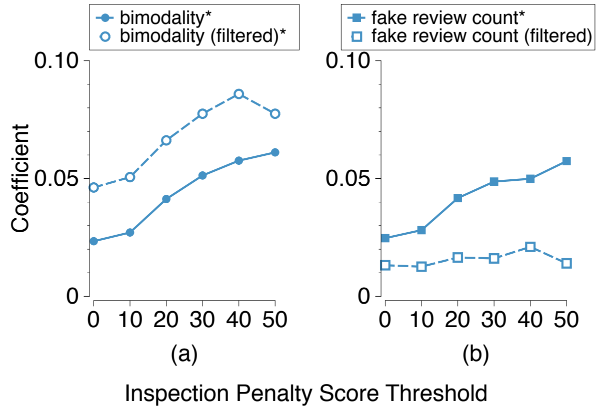

Figure 2: Spearman's coefficients of factors & inspection penalty scores. '*': statistically significant ($p \leq 0.05$)

Note that non-zero penalty scores may not necessarily indicate alarming hygiene issues. For example, violating codes such as *"proper labeling"* or *"proper consumer advisory posted for raw or undercooked foods"* seem relatively minor, and unlikely to be noted and mentioned by reviewers. Therefore, we focus on restaurants with severe violations, as they are exactly the set of restaurants that inspectors and customers need to pay the most attention to. To define restaurants with "severe violations" we experiment with a varying threshold t, such that restaurants with score $\geq t$ are labeled as *"unhygienic"*.[2]

3 Correlates of Inspection Penalty Scores

We examine correlation between penalty scores and several statistics of reviews:

I. Volume of Reviews:

[2] For restaurants with *"hygienic"* labels, we only consider those without violation, as there are enough number of such restaurants to keep balanced distribution between two classes.

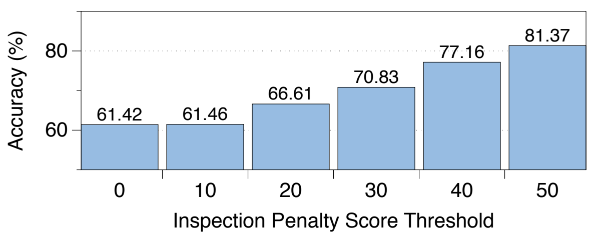

Figure 3: Trend of penalty score thresholds & accuracies.

- *count of all reviews*
- *average length of all reviews*

II. Sentiment of Reviews: We examine whether the overall sentiment of the customers correlates with the hygiene of the restaurants based on following measures:

- *average review rating*
- *count of negative (≤ 3) reviews*

III. Deceptiveness of Reviews: Restaurants with bad hygiene status are more likely to attract negative reviews, which would then motivate the restaurants to solicit fake reviews. But it is also possible that some of the most assiduous restaurants that abide by health codes strictly are also diligent in soliciting fake positive reviews. We therefore examine the correlation between hygiene violations and the degree of deception as follows.

- *bimodal distribution of review ratings*

  The work of Feng et al. (2012) has shown that the shape of the distribution of opinions, overtly skewed bimodal distributions in particular, can be a telltale sign of deceptive reviewing activities. We approximately measure this by computing the variance of review ratings.
- *volume of deceptive reviews based on linguistic patterns*

  We also explore the use of deception classifiers based on linguistic patterns (Ott et al., 2011) to measure the degree of deception. Since no deception corpus is available in the restaurant domain, we collected a set of fake reviews and truthful reviews (250 reviews for each class), following Ott et al. (2011).[3]

[3] 10 fold cross validation on this dataset yields 79.2% accuracy based on unigram and bigram features.

Features	Acc.	MSE	SCC
-	*50.00	0.500	-
review count	*50.00	0.489	0.0005
np review count	*52.94	0.522	0.0017
cuisine	*66.18	0.227	0.1530
zip code	*67.32	0.209	0.1669
avrg. rating	*57.52	0.248	0.0091
inspection history	*72.22	0.202	0.1961
unigram	78.43	0.461	0.1027
bigram	*76.63	0.476	0.0523
unigram + bigram	82.68	0.442	0.0979
all	81.37	0.190	0.2642

Table 1: Feature Compositions & Respective Accuracies, Respective Mean Squared Errors(MSE) & Squared Correlation Coefficients (SCC), np=non-positive

Filtering Reviews: When computing above statistics over the set of reviews corresponding to each restaurant, we also consider removing a subset of reviews that might be dubious or just noise. In particular, we remove reviews that are too far away (delta ≥ 2) from the average review rating. Another filtering rule can be removing all reviews that are classified as deceptive by the deception classifier explained above. For brevity, we only show results based on the first filtering rule, as we did not find notable differences in different filtering strategies.

Results: Fig 1 and 2 show Spearman's rank correlation coefficient with respect to the statistics listed above, with and without filtering, computed at different threshold cutoffs $\in \{0, 10, 20, 30, 40, 50\}$ of inspection scores. Although coefficients are not strong,[4] they are mostly statistically significant with $p \leq 0.05$ (marked with '*'), and show interesting contrastive trends as highlighted below.

In Fig 1, as expected, average review rating is negatively correlated with the inspection penalty scores. Interestingly, all three statistics corresponding to the volume of customer reviews are positively correlated with inspection penalty. What is more interesting is that if potentially deceptive reviews are filtered, then the correlation gets stronger, which suggests the existence of deceptive reviews covering up unhappy customers. Also notice that correlation is

[4] Spearman's coefficient assumes monotonic correlation. We suspect that the actual correlation of these factors and inspection scores are not entirely monotonic.

Hygienic gross, mess, sticky, smell, restroom, dirty
Basic Ingredients: beef, pork, noodle, egg, soy, ramen, pho,
Cuisines Vietnamese, Dim Sum, Thai, Mexican, Japanese, Chinese, American, Pizza, Sushi, Indian, Italian, Asian
Sentiment: cheap, never,
Service & Atmosphere cash, worth, district, delivery, think, really, thing, parking, always, usually, definitely - door: "The wait is always out the *door* when I actually want to go there", - sticker: "I had *sticker* shock when I saw the prices.", - student: "heap, large portions and tasty = the perfect *student* food!", - the size: "i was pretty astonished at *the size* of all the plates for the money.", - was dry: "The beef *was dry*, the sweet soy and anise-like sauce was TOO salty (almost inedible).", - pool: "There are *pool* tables, TV airing soccer games from around the globe and of course - great drinks!"

Table 2: Lexical Cues & Examples - Unhygienic (dirty)

generally stronger when higher cutoffs are used (x-axis), as expected. Fig 2 looks at the relation between the deception level and the inspection scores more directly. As suspected, restaurants with high penalty scores show increased level of deceptive reviews.

Although various correlates of hygiene scores examined so far are insightful, these alone are not informative enough to be used as a predictive tool, hence we explore content-based classification next.

4 Content-based Prediction

We examine the utility of the following features:

Features based on customers' opinion:

1. Aggregated opinion: average review rating
2. Content of the reviews: unigram, bigram

Features based on restaurant's metadata:

3. Cuisine: e.g., Thai, Italian, as listed under Yelp
4. Location: first 5 digits of zip code
5. Inspection History: a boolean feature ("hygienic" or "unhygienic"), a numerical feature (previous penalty score rescaled $\in [0, 1]$), a numeric feature (average penalty score over all previous inspections)

Hygienic:
Cooking Method & Garnish: brew, frosting, grill, crush, crust, taco, burrito, toast **Healthy or Fancier Ingredients**: celery, calamity, wine, broccoli, salad, flatbread, olive, pesto
Cuisines : Breakfast, Fish & Chips, Fast Food, German, Diner, Belgian, European, Sandwiches, Vegetarian
Whom & When: date, weekend, our, husband, evening, night
Sentiment: lovely, yummy, generous, friendly, great, nice
Service & Atmosphere: selection, attitude, atmosphere, ambiance, pretentious

Table 3: Lexical Cues & Examples - Hygienic (clean)

6. Review Count
7. Non-positive Review Count

Classification Results We use liblinear's SVM (Fan et al., 2008) with L1 regularization and 10 fold cross validation. We filter reviews that are farther than 2 from the average rating. We also run Support Vector Regression (SVR) using liblinear. Fig 3 shows the results. As we increase the threshold, the accuracy also goes up in most cases. Table 1 shows feature ablation at threshold $t = 50$, and '*' denotes statistically significant (p$\leq$0.05) difference over the performance with all features based on student t-test.

We find that metadata information of restaurants such as location and cuisine alone show good predictive power, both above 66%, which are significantly higher than the expected accuracy of random guessing (50%).

Somewhat unexpected outcome is aggregated opinion, which is the average review rating during the corresponding inspection period, as it performs not much better than chance (57.52%). This result suggest that the task of hygiene prediction from reviews differs from the task of sentiment classification of reviews.

Interestingly, the inspection history feature alone is highly informative, reaching accuracy upto 72%, suggesting that the past performance is a good predictor of the future performance.

Textual content of the reviews (unigram+bigram) turns out to be the most effective features, reaching upto 82.68% accuracy. Lastly, when all the features are combined together, the performance decreases slightly to 81.37%, perhaps because n-gram features perform drastically better than all others.

4.1 Insightful Cues

Table 2 and 3 shows representative lexical cues for each class with example sentences excerpted from actual reviews when context can be helpful.

Hygiene: Interestingly, hygiene related words are overwhelmingly negative, e.g., *"gross", "mess", "sticky"*. What this suggests is that reviewers do complain when the restaurants are noticeably dirty, but do not seem to feel the need to complement on cleanliness as often. Instead, they seem to focus on other positive aspects of their experience, e.g., details of food, atmosphere, and their social occasions.

Service and Atmosphere: Discriminative features reveal that it is not just the hygiene related words that are predictive of the inspection results of restaurants. It turns out that there are other qualities of restaurants, such as service and atmosphere, that also correlate with the likely outcome of inspections. For example, when reviewers feel the need to talk about *"door", "student", "sticker"*, or *"the size"* (see Table 2 and 3), one can extrapolate that the overall experience probably was not glorious. In contrast, words such as *"selection", "atmosphere", "ambiance"* are predictive of hygienic restaurants, even including those with slightly negative connotation such as *"attitude"* or *"pretentious"*.

Whom and When: If reviewers talk about details of their social occasions such as *"date", "husband"*, it seems to be a good sign.

The way food items are described: Another interesting aspect of discriminative words are the way food items are described by reviewers. In general, mentions of basic ingredients of dishes, e.g., *"noodle", "egg", "soy"* do not seem like a good sign. In contrast, words that help describing the way dish is prepared or decorated, e.g., *"grill", "toast", "frosting", "bento box" "sugar"* (as in "sugar coated") are good signs of satisfied customers.

Cuisines: Finally, cuisines have clear correlations with inspection outcome, as shown in Table 2 and 3.

5 Related Work

There have been several recent studies that probe the viability of public health surveillance by measuring relevant textual signals in social media, in particular, micro-blogs (e.g., Aramaki et al. (2011), Sadilek et al. (2012b), Sadilek et al. (2012a), Sadilek et al. (2013), Lamb et al. (2013), Dredze et al. (2013), von Etter et al. (2010)). Our work joins this line of research but differs in two distinct ways. First, most prior work aims to monitor a specific illness, e.g., influenza or food-poisoning by paying attention to a relatively small set of keywords that are directly relevant to the corresponding sickness. In contrast, we examine all words people use in online reviews, and draw insights on correlating terms and concepts that may not seem immediately relevant to the hygiene status of restaurants, but nonetheless are predictive of the outcome of the inspections. Second, our work is the first to examine online reviews in the context of improving public policy, suggesting additional source of information for public policy makers to pay attention to.

Our work draws from the rich body of research that studies online reviews for sentiment analysis (e.g., Pang and Lee (2008)) and deception detection (e.g., Mihalcea and Strapparava (2009), Ott et al. (2011), Feng et al. (2012)), while introducing the new task of public hygiene prediction. We expect that previous studies for aspect-based sentiment analysis (e.g., Titov and McDonald (2008), Brody and Elhadad (2010), Wang et al. (2010)) would be a fruitful venue for further investigation.

6 Conclusion

We have reported the first empirical study demonstrating the promise of review analysis for predicting health inspections, introducing a task that has potentially significant societal benefits, while being relevant to much research in NLP for opinion analysis based on customer reviews.

Acknowledgments

This research was supported in part by the Stony Brook University Office of the Vice President for Research, and in part by gift from Google. We thank anonymous reviewers and Adam Sadilek for helpful comments and suggestions.

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 804–812, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Song Feng, Longfei Xing, Anupam Gogar, and Yejin Choi. 2012. Distributional footprints of deceptive product reviews. In *ICWSM*.

Katie Filion and Douglas A Powell. 2009. The use of restaurant inspection disclosure systems as a means of communicating food safety information. *Journal of Foodservice*, 20(6):287–297.

Spencer Henson, Shannon Majowicz, Oliver Masakure, Paul Sockett, Anria Johnes, Robert Hart, Debora Carr, and Lewinda Knowles. 2006. Consumer assessment of the safety of restaurants: The role of inspection notices and other information cues. *Journal of Food Safety*, 26(4):275–301.

Ginger Zhe Jin and Phillip Leslie. 2003. The effect of information on product quality: Evidence from restaurant hygiene grade cards. *The Quarterly Journal of Economics*, 118(2):409–451.

Ginger Zhe Jin and Phillip Leslie. 2005. The case in support of restaurant hygiene grade cards.

Ginger Zhe Jin and Phillip Leslie. 2009. Reputational incentives for restaurant hygiene. *American Economic Journal: Microeconomics*, pages 237–267.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on twitter. In *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312, Suntec, Singapore, August. Association for Computational Linguistics.

NYC-DoHMH. 2012. Restaurant grading in new york city at 18 months. *New York City Department of Health and Mental Hygiene*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA, June. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Adam Sadilek, Henry Kautz, and Vincent Silenzio. 2012a. Predicting disease transmission from geotagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Adam Sadilek, Henry A. Kautz, and Vincent Silenzio. 2012b. Modeling spread of disease from social interactions. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM*. The AAAI Press.

Adam Sadilek, Sean Brennan, Henry Kautz, and Vincent Silenzio. 2013. nemesis: Which restaurants should you avoid today? *First AAAI Conference on Human Computation and Crowdsourcing.*

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio, June. Association for Computational Linguistics.

Peter von Etter, Silja Huttunen, Arto Vihavainen, Matti Vuorinen, and Roman Yangarber. 2010. Assessment of utility in web mining for the domain of public health. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 29–37, Los Angeles, California, USA, June. Association for Computational Linguistics.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792. ACM.