Is Yelp Actually Cleaning Up the Restaurant Industry? A Re-Analysis on the Relative Usefulness of Consumer Reviews

Kristen M. Altenburger
Stanford University
kaltenb@stanford.edu

Daniel E. Ho
Stanford University
dho@law.stanford.edu

ABSTRACT

Social media provides the government with novel methods to improve regulation. One leading case has been the use of Yelp reviews to target food safety inspections. While previous research on data from Seattle finds that Yelp reviews can predict unhygienic establishments, we provide a more cautionary perspective. First, we show that prior results are sensitive to what we call "Extreme Imbalanced Sampling": *extreme* because the dataset was restricted from roughly 13k inspections to a sample of only 612 inspections with only extremely high or low inspection scores, and *imbalanced* by not accounting for class imbalance in the population. We show that extreme imbalanced sampling is responsible for claims about the power of Yelp information in the original classification setup. Second, a re-analysis that utilizes the full dataset of 13k inspections and models the full inspection score (regression instead of classification) shows that (a) Yelp information has lower predictive power than prior inspection history and (b) Yelp reviews do not significantly improve predictions, given existing information about restaurants and inspection history. Contrary to prior claims, Yelp reviews do not appear to aid regulatory targeting. Third, this case study highlights critical issues when using social media for predictive models in governance and corroborates recent calls for greater transparency and reproducibility in machine learning.

KEYWORDS

consumer reviews; Yelp; food safety; regulation; replication

ACM Reference Format:
Kristen M. Altenburger and Daniel E. Ho. 2019. Is Yelp Actually Cleaning Up the Restaurant Industry? A Re-Analysis on the Relative Usefulness of Consumer Reviews. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308558.3313683

1 INTRODUCTION

Governments are increasingly exploring predictive approaches with large-scale datasets to improve regulation. Data mining approaches have witnessed great success across domains such as inferring wealth in a country using mobile phone records [5] or satellite images [31] or inferring migration stocks using both social media data like Facebook and survey data [53]. Social media datasets like Twitter have the potential to provide an additional signal of future, offline outcomes of interest as illustrated by predicting electoral outcomes [16], health outcomes [12], and many others [40].

Increasingly, government usage has expanded to leverage social media data for predictive governance [34]. Police departments have mined Facebook and Twitter for investigation and predictive policing [48]. Municipal service departments have used social media to forecast the need for municipal services, such as filling potholes [1]. Governments have used traffic complaints on Twitter to shape transportation planning [51]. One of the leading examples has been the case of food safety [14, 19, 23, 33, 44, 46, 47]. Health departments from major cities, including New York, Las Vegas, Boston, and Chicago, have, for instance, used Yelp and Twitter terms, such as "sick" and "vomit," to allocate enforcement and investigatory resources.

While governments have experimented with such novel data uses, researchers have begun to examine the reliability of such approaches. A prominent example is that of Google Flu Trends. Google Flu Trends was initially advertised as using search terms to forecast influenza epidemics, beating early detection of the Centers for Disease Control and Prevention (CDC) [18]. But researchers showed that the algorithm was prone to substantial overfitting [36]. As put in *Science*, big data "does not mean that one can ignore measurement and construct validity and reliability" [36].

Our article makes three contributions. First, we revisit the seminal case of food safety. Earlier work showed that the use of social media may import private biases into public enforcement [2, 11]. We show here instead that the leading case of social media data use for food safety [33] has substantial methodological shortcomings: (a) changing the prediction problem to predict only extremely hygienic and extremely unhygienic establishments and (b) failing to assess the impact of class imbalance in the population. We call this a case of "extreme imbalanced sampling." We show that a single, seemingly minor correction – drawing a random sample of hygienic establishments (as opposed to using the nonrandom order in the original dataset, which overrepresents the same establishments) – causes the originally reported results to disappear. This analysis follows other social media studies documenting classification results that are overoptimistic [9] or limited in their data sampling approaches [52].

Second, we conduct a comprehensive analysis, by using the full dataset of about 13k inspections, rather than the subset of approximately 612 inspections originally analyzed. We find that (a) Yelp reviews provide substantially less predictive power than inspection history already possessed by the health department, and (b) Yelp reviews do not significantly improve predictions, given existing information about restaurants and inspection history.

Third, we discuss some of the implications both for machine learning and government use of social media. Most importantly,

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313683

our paper points to emerging recognition of a replication crisis in machine learning [27, 49]. Methodological flubs can easily occur inadvertently with complex methods, and we highlight that the implications are severe when it comes to public policy. News outlets, including *NPR*, the *New York Times*, *Forbes*, and *Newsweek*, quickly touted the idea of food safety targeting based on social media. And private industry may have an incentive to promote use cases before thorough vetting. Yelp's CEO, for instance, proclaimed that Yelp could "beat gold standard healthcare measures" [50]. Our evidence shows that much more thought needs to be given for ethical and reliable use by government of social media [21].

Our article proceeds as follows. In Section 2, we describe the Seattle restaurant dataset that led to provocative claims that Yelp could "clean up the restaurant industry" [3]. In Section 3, we assess the original analysis that purported to show that Yelp reviews could be used to predict unhygienic establishments. We highlight how a single, minor modification – random rather than non-random sampling of the majority class – invalidates the claim that Yelp terms are more informative than inspection history. In Section 4, we conduct a more comprehensive analysis that (a) makes use of all of the data (13k as opposed to 612 inspections), (b) uses all of the information in inspection outcomes (regression vs. classification), and (c) uses word embeddings to exploit review terms [42]. In Section 5, we discuss the more general implications for replication in machine learning and government use of social media data.

2 THE SEATTLE RESTAURANT DATASET

The prediction problem considered throughout this work is to infer the inspection penalty status of a restaurant (either in terms of a restaurant's raw inspection score or classifying restaurants as unhygienic if above a certain score threshold [33]) given features created based on the restaurant, online Yelp reviews, or past inspection history. This prediction problem can be framed as either a classification problem if one sets a threshold for defining "unhygienic" restaurants or a regression problem if one aims to predict the actual prediction score. We will use the terms penalty score and inspection score interchangeably throughout this work.

Data: We utilize the Yelp and inspection dataset compiled in earlier work [33], which we summarize here. In Seattle, food safety inspectors conduct routine inspections one to two times a year per establishment, resulting in an inspection score based on health code violations. The data contain about 13k inspections covering 1,756 Seattle restaurants with 152k Yelp reviews from 2006 to 2013. Yelp reviews are collected for the preceding period of time between inspections (or six months earlier when no prior inspection existed). Review data is concatenated as a text string.[1]

Table 1: Features for prediction. Inspection history includes an establishment's previous penalty score and an average measure across previous inspections. Cuisine indicates type of food offered, and ZIP code denotes restaurant's location. Yelp features we consider include ratings, review counts, and then the review text itself using a TF-IDF or word embedding representation. The negative review count feature is referred to as non-positive review count in [33].

Feature Name	Feature Category
inspection history	Inspection
cuisine	Restaurant
zip code	Restaurant
avrg. rating	Yelp
negative review count	Yelp
review count	Yelp
review text	Yelp

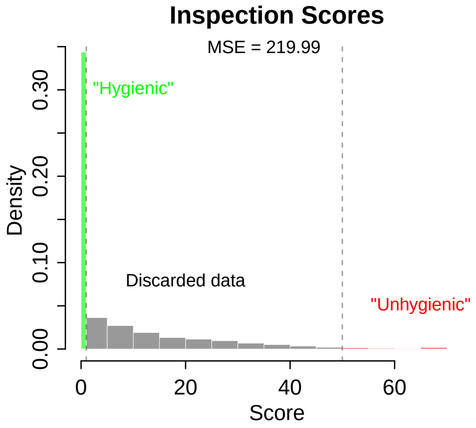

Figure 1: Distribution of inspection scores, with vertical lines indicating cutoffs for "unhygienic" restaurants (red) with scores above 50 and "hygienic" restaurants (green) with scores of 0. This figure shows how the prior analysis sampled from the extremes of the distribution.

Features: We evaluate three categories of features for predicting a restaurant's sanitation level: restaurant-fixed features (cuisine and location), inspection history features, and features derived from Yelp for the preceding inspection period (e.g., average rating, review text). We list these features in Table 1. Inspection history includes information derived from an establishment's past inspection record (e.g., average past inspection history or normalized metrics).[2] The principal question is whether Yelp reviews help predict sanitation.

Prior Analysis: Prior analysis of this data attempted to classify "hygienic" and "unhygienic" restaurants. First, all features were calculated on the full dataset of approximately 13k observations and outcomes were labeled as "unhygienic" if the inspection score was greater than 50.[3] Figure 1 displays the outcome distribution, showing highly skewed scores. Only 306 restaurants, 2.3% of the sample, are labeled as "unhygienic," resulting in an extremely imbalanced setting for classification. Second, because of this class imbalance, 306 "hygienic" restaurants with no violation history

[1] See https://www3.cs.stonybrook.edu/~junkang/hygiene/. We are not able to observe different reviews or reviewers. While there are other data quality issues, such as duplicated review texts, we do not address these issues here for comparability.

[2] While cuisine information is technically derived from Yelp data, we denote this as a restaurant category because this information is not derived from user-input. It is common for health agencies to retain information about cuisine.

[3] While the original paper describes the labels as a non-strict inequality, the implementation code uses a strict inequality (see line 720 of train_model.py). The original analysis also assessed performance at other thresholds, but mainly focused on the 50 point threshold as that with highest accuracy (see Figure 3 and Table 1 in [33]).

were (non-randomly) selected from the majority class.[4] We discuss below the effects of this non-random selection. Figure 1 shows that 34% of restaurants receive scores of 0 and the prior analysis discards 63% of restaurants receiving scores between 0 and 50.

Third, the prior analysis trained ℓ_1-regularized support vector machines (SVMs), with a constant C parameter, for each feature separately, evaluating results via 10-fold cross-validation on the dataset of 612 observations. By only evaluating predictive performance on this dataset of 612 observations, predictive performance may be overstated compared to performance on the full population of 13k inspections. The top left panel of Figure 3 presents accuracy statistics as reported in the original paper.[5] Because unigrams and bigrams from Yelp reviews outperformed inspection history, the study concluded that Yelp reviews can aid in targeting food safety inspections. We obtained the original code and successfully replicate the results in the top right panel of Figure 3. Minor differences likely stem from different train/test splits, as random seeds were not originally set.

3 EXTREME IMBALANCED SAMPLING

We now explain the methodological challenges in the prior analysis. First, rather than predicting hygienic and non-hygienic restaurants with a particular cutoff for classification, the prior analysis dramatically simplified the prediction problem. Hygienic restaurants were those with inspection scores of 0 violations[6] (green in Figure 1) and unhygienic restaurants were those with over 50 violation points (red in Figure 1). In other words, the analysis sampled extreme end points, making the prediction substantially easier. Consider an educational analogy: instead of predicting the full grade range for students, it is akin to classifying straight-A students versus drop-out students, discarding the majority of students in the mid-range.

Moreover, hygienic restaurants were not randomly selected from the 4,571 hygienic observations. Instead, they were selected based on (non-random) order in the dataset, which was sorted by restaurant and cuisine type. As a result, hygienic establishments in the training data overrepresent the *same* restaurants and cuisines, as illustrated in Figure 2. As can be seen from that Figure, 72% of training observations represent multiple inspections for the same establishment (left panel), which would be expected for only 12% of training observations under random sampling[7] (right panel illustrates one random sample instance). This is problematic because terms unique to a restaurant (e.g., “café”, “Chipotle”) may result in overfitting. To continue the educational analogy, it would be akin to predicting grades for each school year, but rather than randomly sampling training data, repeatedly selecting the same students observed over multiple school years.

Second, instead of implementing methods to account for class imbalance using undersampling [15] or oversampling [8], the study reported accuracy only on the extreme balanced sample of 612 observations. By not reporting results on the full held-out dataset, results were likely overly optimistic, particularly in light of non-random sampling of hygienic restaurants noted above. Jointly, we refer to this problem as one of “extreme imbalanced sampling,” and consider two minor modifications to illustrate the consequences of these methodological choices.

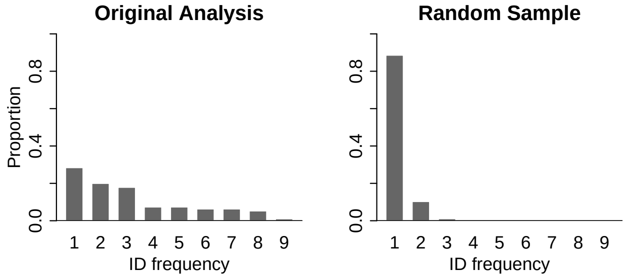

Figure 2: Histogram of counts of occurrences for a unique hygienic restaurant (i.e. defined by inspection scores of 0) in the original analysis (left panel) and under one instance of a random sample of 306 hygienic restaurants (right panel). This panel shows that by selecting observations in the non-random order of the dataset (sorted by restaurant ID), the hygienic sample overrepresents specific establishments compared to a sample one would expect under random sampling.

Random Sample of Hygienic Establishments: First, we consider the consequence of over-representing the same restaurants in the hygienic training sample. We randomly sample 306 hygienic restaurants (i.e. restaurants with scores of 0) and fit the same SVM models. We report mean accuracy results across 5 different datasets created with a random sample of hygienic restaurants. Contrary to the prior analysis, we find that inspection history has *higher* predictive performance than Yelp review features as shown in the bottom left panel of Figure 3. Indeed, this simple modification reduces the predictive power of Yelp review substantially: unigrams and bigrams drop from 83% reported accuracy to 69% reported accuracy, compared to 71% accuracy of inspection history, suggesting that the original results were due to overfitting.

Full Imbalanced Dataset: Next, we consider the consequence of not taking into consideration the imbalanced nature of this dataset by only reporting results on the sample of 612 inspections. Instead, we evaluate predictive performance on the full population of 13k inspections (via 10-fold cross-validation), accounting for class imbalance [30]. We implement undersampling on the training data, and report class averaged accuracy results on the test data.[8] The lower right panel of Figure 3 illustrates results. Inspection history alone achieves higher relative performance (67% class-averaged accuracy) than all other features. The (class-averaged) accuracy for unigrams and bigrams drops to 62%. Moreover, the overall predictive performance is substantially lower than what previous results suggested.

[4] As described in the original paper [33]: “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”.

[5] See Table 1 in [33].

[6] There is one observation with a score of -1 that is treated as an inspection score of 0.

[7] We simulate random draws of hygienic restaurants and compute the mean number of duplicates across 10k draws.

[8] The class averaged accuracy represents a class-specific measure of accuracy calculated as $:= \frac{1}{2} \cdot \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$, where TP denotes true-positives, FN denotes false-negatives, TN denotes true-negatives, and FP denotes false-positives. Note that in settings with exact (or approximately) class balance that accuracy is exactly (or approximately) the same as class averaged accuracy.

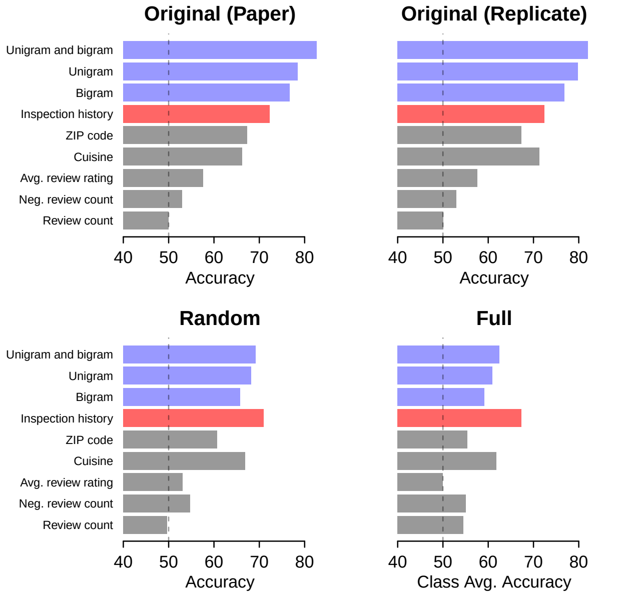

Figure 3: Predictive performance of Yelp review words (blue) and inspection history (red) in prior analysis (top left), replication (top right), random sample of hygienic restaurants (bottom left), and the full imbalanced dataset (bottom right). The figure shows that the relative performance of Yelp reviews depends entirely on a non-random and unrepresentative sample. The vertical dashed line represents baseline accuracy of 50%.

The fact that these two minor modifications reverse the core claim of the original paper is sobering. Higher levels of accuracy with predictive targeting could be achieved exclusively with data already possessed by the health department. The bottom panels of Figure 3 also show that predictive performance degrades when the hygienic class is chosen in random order and/or the full dataset is evaluated.

4 A RE-ANALYSIS

We now re-analyze this dataset from a regression perspective, to more closely align with the actual prediction problem of modeling inspection scores (without discarding data), and offer an alternative interpretation of the relative usefulness of inspection features versus online Yelp features. We view this re-analysis as a step forward in understanding the ability to predict an establishment's future inspection performance.

4.1 Data, Features, and Model

First, we reduce the dataset to routine (unannounced) inspections, by matching the Seattle Restaurant Dataset to data from the public health department. This removes (a) return inspections, which are triggered by poor performance in routine inspections, and typically occur a month after the routine inspection, and (b) educational inspections, which are unscored.[9] This leaves us with 12,613 observations.

Second, for comparability, we use largely the same feature set. Minor modifications are that we (a) construct cuisine features that indicate if a restaurant is denoted by a particular cuisine category, (b) do not normalize inspection scores,[10] (c) use the raw average rating score as opposed to further processing this score, and (d) use the raw previous penalty score instead of discretizing scores. The more substantial change is that we address sparseness of Yelp review text by leveraging word embedding `doc2vec` [37], instead of raw unigrams and bigrams. We apply a standard text cleaning process[11] that converts text to lower case and tokenizes all words. Results for embeddings created using the `doc2vec` parameters `vector_size=100`, `window=5`, and `min_count=3` are reported for the remainder of this paper.[12]

Third, rather than discarding information in inspection scores by discretization at an arbitrary cutoff, we evaluate each feature's prediction performance in a regression set-up to predict the full inspection score. We evaluate a Random Forest (RF) model [7], which is suitable for capturing higher-order interactions.[13] We report results across a 10-fold cross-validation set-up, where we select hyperparameters via 3-fold cross-validation on each training data portion.[14] Model parameters are selected via a randomized search for quicker run-time with textual features,[15] but we note comparable model performances between random versus grid search implementation on all the other features.[16]

4.2 Evaluation of Inspection, Restaurant, and Yelp-derived Features

Consistent with the earlier study, Figure 4 presents the predictive performance for each of the features individually. The baseline prediction is the mean inspection score from training observations used to predict on testing observations. The dashed vertical line represents the mean performance across the 10-folds for this baseline classifier. For each of the features, we plot densities of mean-squared-error (MSE) across the 10 folds in gray, with the vertical dash representing mean MSE. Figure 4 shows that inspection history achieves the lowest MSE (i.e., the highest predictive performance). Yelp features outside of the review text (review count, non-positive review count, average review reading) perform poorly.

[9] There are 2 observations in the original dataset that do not have corresponding inspection IDs in our dataset, which we drop. We also set one observation with an inspection score of -1 to 0.

[10] Inspection history features here will include the following fields provided in the original dataset: `inspection_average_prev_penalty_scores` and `inspection_prev_penalty_score`.

[11] In the `gensim` library in Python, we use `gensim.utils.simple_preprocess()`.

[12] We vary the hyperparameter settings when training a doc2vec model and observe comparable prediction performance across the range of hyperparameters considered. See [35] for more technical details on hyperparameter parameters in doc2vec.

[13] Technically inspection penalty scores are integers but RF computes averages in subsamples of the model resulting in fractional numbers. We evaluate the raw predictions from RF and avoid rounding.

[14] We consider the following search range for each hyperparameter parameter: `max_features = [None, 0.25, 0.5, 0.75]`, `max_depth = [None, 5, 10]`, `min_samples_leaf = [0.0005, 0.01, 0.05, 0.1]`, `min_samples_split = [2, 5, 10]`, and `n_estimators = [100, 200, 500]`

[15] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

[16] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

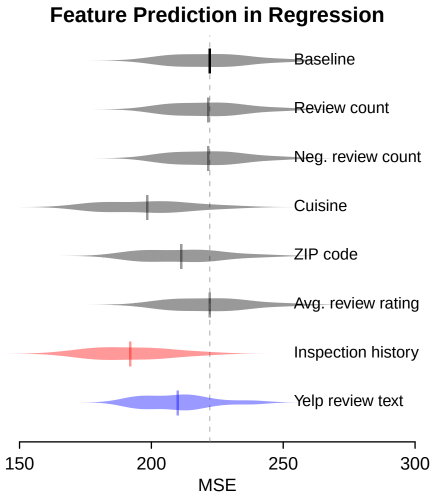

Figure 4: Predictive performance of restaurant-, inspection- or Yelp-derived features as measured by mean squared error (MSE) in regression problem. Results are based on 10-fold cross-validation set-up (presented as densities, with means in vertical lines), fitting a RF regression model with hyperparameter tuning. Yelp review (blue) indicates word embedding representation of review content.

We observe only modest performance of Yelp reviews, providing less predictive power than inspection history and about the same as ZIP code.

4.3 Marginal Value of Yelp Features

While Figure 4 considered features individually, we now compare the performance of a "full" model consisting of inspection-, restaurant-, and Yelp-derived features, versus a "non-Yelp" model to evaluate the marginal value of Yelp. A full model with all features considered in Figure 4 yields a mean MSE of 185.70, while a non-Yelp model (excluding Yelp review, ratings, and review counts) yields a comparable mean MSE of 183.61, with a non-significant p-value of 0.984 from a paired Wilcoxon signed rank one-sided test for whether the additional Yelp features improve the fit. These results indicate that Yelp features do not significantly improve predictions, given existing information available about restaurants and inspection history. Table 2 illustrates the marginal value of Yelp reviews given different feature sets. Yelp adds predictive value only in instances where the features are extremely limited (e.g., only ZIP code). The bottom four rows show that with any realistic feature set (e.g., inspection history and cuisine), Yelp reviews do not statistically (and substantively) improve predictive performance.

4.4 Limitations

We note several limitations to our analysis. First, as in the prior analysis, our review information is aggregated. It may be important to treat each user's review as a separate block of text instead of a concatenated text field across the inspection period. That said, review information is already exceedingly sparse, so accounting for specific reviews may be challenging. Second, it is possible that more refined feature extraction from text (e.g., sentence encoding) would improve the predictive performance of Yelp reviews or that some alternative feature representation of the review text would be useful.[17] The prior analysis used raw unigrams and bigrams, which we also found lacking in predictive power. Third, while we have followed the original setup of evaluating performance by cross-validation, a more appropriate evaluation should take into account that the use case involves forecasting *future* sanitation [32]. Understanding how algorithms generalize temporally has been raised in other contexts such as inferring economic conditions in a country [6]. In our setting, a more appropriate model may involve early detection prediction to trigger an early inspection [55]. Fourth, alternative datasets or features, such as the length of time a restaurant has been open, might provide a better signal of a restaurant's future hygienic conditions.

Table 2: This table evaluates the additional predictive value of the Yelp review text given different subsets of other features such as ZIP code, cuisine, or inspection history. We observe that Yelp's additional predictive value is minimal. We report p-values from a paired Wilcoxon signed rank test, testing whether the difference in MSE without versus with the Yelp text features across the same test folds is greater than 0. $^{*}/^{**}$ indicate statistical significance at 0.05 and 0.01.

Features included			Mean Squared Error		
ZIP	Cuisine	Inspection history	Without Yelp review text	With Yelp review text	p-value
			222.13	209.93	0.003^{**}
✓			211.30	203.27	0.003^{**}
	✓		198.50	204.13	0.996
		✓	192.00	189.12	0.004^{**}
✓	✓		188.44	199.02	0.998
✓		✓	188.65	187.49	0.042^{*}
	✓	✓	187.91	188.65	0.762
✓	✓	✓	183.61	185.61	0.988

5 DISCUSSION

While web and social media data offer tantalizing possibilities for governance, we have shown that one of the most celebrated cases for "reinventing food safety" overstates the utility of Yelp reviews. While this may be no fault of the authors – indeed the original authors generously offered replication code, data, and comments to help assess the methodology – the popular susceptibility to "big data hubris" has considerable ethical and policy implications. As mentioned above, health departments in Boston, Las Vegas, Chicago, and New York City each adopted techniques for targeting based on Yelp or Twitter, and much more serious work needs to be conducted to validate and assess these methods for policy adoption. We here highlight three particular implications.

5.1 Replication and Transparency

Our study corroborates that the replication crisis that has affected the social sciences and biomedicine [10, 29] may also affect the machine learning community [15, 39]. Only 6% of papers at top artificial intelligence conferences made code available [22]. As data and methods become more complex, unarticulated researcher "degrees of freedom" expand. Here, such discretionary choices included: the

[17] For observations on the consequences of pre-processing for topic modeling, see [13].

feature set, feature representation (e.g., TF-IDF weighting of word terms), data cleaning, software version, hyperparameter tuning (the kernel, C, and γ in the original SVM setup), the penalty threshold, and selection of majority class from an ordered dataset. The original authors deserve much credit to providing the replication data. But even with the exact data in hand, we were unable to replicate the findings after three months of efforts. Only after the code was generously shared did it become clear that classification was limited to ≈600 restaurants with extreme imbalanced and non-random sampling. Moreover, while we were able to perform a "direct replication," this uncovered substantial methodological problems.

Our evidence hence underscores recent calls for more general reproducibility (or "conceptual replication") in machine learning [15, 20, 38]. More generally, our findings highlight the importance of recent projects in machine learning to promote transparency and replication , such as including code with conference submissions and details on the trained model [43], automating hyperparameter searching [4], developing open source software and computational infrastructure for reproducibility [41, 49], and providing datasheets for detailed dataset information [17].

We also note that another form of transparency is relevant in the law and policy context, namely surrounding the role of private industry sponsorship. When private data is used for public purposes, a greater obligation of transparency may attach. Yelp, for instance, sponsored a predictive tournament for health scores in Boston [19], posing a potential conflict of interest. Recall Yelp's CEO statement that Yelp could "beat gold standard healthcare measures." Disclosure both of sponsorship, data, and methodology would help ensure that these findings can be externally validated when private data enters public enforcement.

5.2 Machine and Policy Translation

Our results also show that much work remains to be done for machine learning to be translated into policy. First, notwithstanding wide media coverage of how Yelp could solve food safety, we show that the prediction is much more difficult than previously appreciated. As Table 2 shows, predictive power of *any* features is not very high. By definition, health authorities already retain rich administrative data on inspection history and restaurant meta-information, and Table 2 confirms that with any reasonable baseline set of features, Yelp reviews add no statistically significant predictive power.

Second, machine learning scholars need to understand the substantive policy setting of a prediction problem. For instance, when the prediction problem is defined as a classification problem, cutoff points should be policy-relevant. The 50-point cutoff, predicting only the most extreme 2.3% of the sample, does not accord to any meaningful quantity in Seattle. As a policy matter, Seattle (and King County) mandate (i) a return inspection when more than 35 points for critical violations are assessed and (ii) a shutdown of the restaurant when more than 90 critical violation points are assessed. The prior analysis neither distinguished between critical and non-critical violations nor paid attention to these substantive thresholds. Converting a regression to classification problem must be done with care and attention to the substantive setting [28], particularly when discretization can discard meaningful information.

Third, applying this technique in policy has to grapple with the fact that Yelp reviews reflect the digital divide. In Seattle and King County, for instance, Yelp has a much higher penetration in areas with lower minority presence, higher educational attainment, and lower unemployment rates [26]. When reviews themselves are of an unrepresentative sample of restaurants, there are considerable ethical concerns about the distributive implications for adopting food safety enforcement based on such data. For a similar example, Boston's StreetBump application, which used smartphone accelerometer and GPS data to trigger service requests for potholes, initially deployed more public resources to wealthier areas where smartphone penetration was higher [11]. None of these concerns are explored in work proposing to use social media to target inspections and thus, these concerns warrant much greater attention.

Last, scholars may also need to contemplate what "predictive accuracy" means in a given context. One of the predominant substantive concerns in food safety enforcement has been that inspection stringency may be largely driven by the differences across inspectors [24, 25]. If so, targeting based on predictive scores may contravene rather than support policy goals: more inspection resources would simply be deployed to where inspections are already the most stringent [26].

5.3 Class Imbalance

For classification problems, class imbalance is a common concern [45, 54]. The prior study exacerbated this concern by engaging in extreme imbalanced sampling. But the fact that accuracy statistics were only presented on the sample of ≈600 restaurants may illustrate a more pervasive challenge in machine learning: while it may be desirable to train a classifier on a balanced sample (e.g., via undersampling or oversampling), class imbalance in the population must be accounted for when assessing model performance. This can clearly be seen in the bottom two panels of Figure 3, where class-averaged accuracy from the population appears much worse than the originally reported 82% accuracy statistic. This concern is particularly acute in the policy setting. Food safety predictions, for instance, necessarily must be conducted on the population of restaurants.

6 CONCLUSION

We have shown in this work that one of the most widely cited examples of machine learning with social media for public policy suffers from considerable concerns about replicability. We close by noting that the original authors are not to be blamed: they advanced a novel hypothesis and amassed new data to test it. Given the large number of researcher degrees of freedom, any researcher may inadvertently introduce methodological problems in one of hundreds of lines of code. Yet the case also illustrates the profound challenges of the replication crisis given researcher degrees of freedom in machine learning. We believe that the ethical concerns with such applications are compounded when applied to law and public policy, heightening the importance of recent calls for improving the replication, transparency, and validation of methods in the machine learning community.

Code Availability Statement: For access to the original code for replicating the original results in [33], please contact Jun Kang at junkang@cs.stonybrook.edu. All other code for the analysis presented in this paper is available at: `https://github.com/kaltenburger/YelpReplicate`

Acknowledgements: Altenburger was supported in part by the National Defense Science and Engineering Graduate Fellowship program. We thank Kang et al. for sharing their original code, and Yejin Choi, Alex Chin, Tristan Cossio, Dirk Hovy, Eun Seo Jo, Philip Kegelmeyer, Erin LeDell, Michael Luca, Anne McDonough, Oluchi Mbonu, Stephen Ragain, Arjun Seshadri, Sarah Shugars, George Pakapol Supaniratisai, Angela Teuscher, Johan Ugander, and anonymous reviewers for helpful comments.

REFERENCES

[1] Laura Adler. 2016. Learning from Location. *Data-Smart City Solutions* (2016).

[2] Kristen M Altenburger and Daniel E Ho. 2018. When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions. *Journal of Institutional and Theoretical Economics* (2018).

[3] Emily Badger. 2013. How Yelp Might Clean Up the Restaurant Industry. *Atlantic* (2013).

[4] James Bergstra, Daniel Yamins, and David Daniel Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. (2013).

[5] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350, 6264 (2015), 1073–1076.

[6] Joshua E Blumenstock. 2018. Estimating Economic Characteristics with Phone Data. In *AEA Papers and Proceedings*, Vol. 108. 72–76.

[7] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[9] Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on Twitter: It's not easy!. In *ICWSM*.

[10] Open Science Collaboration et al. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[11] Kate Crawford. 2013. The hidden biases in big data. *Harvard Business Review* 1 (2013).

[12] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 13 (2013), 1–10.

[13] Matthew J Denny and Arthur Spirling. 2018. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis* 26, 2 (2018), 168–189.

[14] Katelynn Devinney, Adile Bekbay, Thomas Effland, Luis Gravano, David Howell, Daniel Hsu, Daniel O'Hallorhan, Vasudha Reddy, Faina Stavinsky, HaeNa Waechter, et al. 2018. Evaluating Twitter for Foodborne Illness Outbreak Detection in New York City. *Online Journal of Public Health Informatics* 10, 1 (2018).

[15] Chris Drummond, Robert C Holte, et al. 2003. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on Learning from Imbalanced Datasets II*, Vol. 11. Citeseer, 1–8.

[16] Daniel Gayo-Avello. 2013. A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review* 31, 6 (2013), 649–679.

[17] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for Datasets. *arXiv preprint arXiv:1803.09010* (2018).

[18] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012.

[19] Edward L. Glaeser, Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2016. Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy. *American Economic Review* 106, 5 (May 2016), 114–18.

[20] Steven N Goodman, Daniele Fanelli, and John PA Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine* 8, 341 (2016), 341ps12–341ps12.

[21] Verena Grubmüller, Katharina Götsch, and Bernhard Krieger. 2013. Social media analytics for future oriented policy making. *European Journal of Futures Research* 1, 1 (26 Sep 2013), 1–20.

[22] Odd Erik Gundersen and Sigbjørn Kjensmo. 2017. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence and the Twenty-Eighth Innovative Applications of Artificial Intelligence Conference*.

[23] Cassandra Harrison, Mohip Jorder, Henri Stern, Faina Stavinsky, Vasudha Reddy, Heather Hanson, H Waechter, Luther Lowe, Luis Gravano, Sharon Balter, et al. 2014. Using online reviews by restaurant patrons to identify unreported cases of foodborne illness – New York City, 2012–2013. *Morbidity and Mortality Weekly Report* 63, 20 (2014), 441–445.

[24] Daniel E. Ho. 2012. Fudging the Nudge: Information Disclosure and Restaurant Grading. *Yale Law Journal* 122, 3 (2012), 574–688.

[25] Daniel E. Ho. 2017. Does Peer Review Work: An Experiment of Experimentalism. *Stanford Law Review* 69 (2017), 1–119.

[26] Daniel E Ho. 2017. Equity in the Bureaucracy. *Irvine Law Review* 7 (2017), 401–451.

[27] Matthew Hutson. 2018. Artificial intelligence faces reproducibility crisis.

[28] Nitin Indurkhya and Sholom M Weiss. 2001. Solving regression problems with rule-based ensemble classifiers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 287–292.

[29] John PA Ioannidis. 2005. Why most published research findings are false. *PLoS Medicine* 2, 8 (2005), e124.

[30] Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent Data Analysis* 6, 5 (2002), 429–449.

[31] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794.

[32] Sham Kakade, Percy Liang, Vatsal Sharan, and Gregory Valiant. 2016. Prediction with a short memory. *arXiv preprint arXiv:1612.02526* (2016).

[33] Jun Seok Kang, Polina Kuznetsova, Michael Luca, and Yejin Choi. 2013. Where not to eat? improving public policy by predicting hygiene inspections using online reviews. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1443–1448.

[34] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105, 5 (2015), 491–95.

[35] Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368* (2016).

[36] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343, 6176 (2014), 1203–1205.

[37] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*. 1188–1196.

[38] Jeffrey T Leek and Roger D Peng. 2015. Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences* 112, 6 (2015), 1645–1646.

[39] Zachary C Lipton and Jacob Steinhardt. 2018. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341* (2018).

[40] Yelena Mejova, Ingmar Weber, and Michael W Macy. 2015. *Twitter: a digital socioscope*. Cambridge University Press.

[41] Jill P Mesirov. 2010. Accessible reproducible research. *Science* 327, 5964 (2010), 415–416.

[42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[43] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 220–229.

[44] Elaine O Nsoesie, Sheryl A Kluberg, and John S Brownstein. 2014. Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports. *Preventive Medicine* 67 (2014), 264–269.

[45] Rachel A Oldroyd, Michelle A Morris, and Mark Birkin. 2018. Identifying Methods for Monitoring Foodborne Illness: Review of Existing Public Health Surveillance Techniques. *JMIR Public Health and Surveillance* 4, 2 (2018), e57.

[46] Adam Sadilek, Sean Brennan, Henry Kautz, and Vincent Silenzio. 2013. nEmesis: which restaurants should you avoid today?. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[47] John P. Schomberg, Oliver L. Haimson, Gillian R. Hayes, and Hoda Anton-Culver. 2016. Supplementing Public Health Inspection via Social Media. *PLOS ONE* 11, 3 (03 2016), 1–21.

[48] Somini Sengupta. 2013. In Hot Pursuit of Numbers to Ward Off Crime. *New York Times* (2013).

[49] Sören Sonnenburg, Mikio L Braun, Cheng Soon Ong, Samy Bengio, Leon Bottou, Geoffrey Holmes, Yann LeCun, Fernando Pereira, and Carl Edward Rasmussen. 2007. The Need for Open Source Software in Machine Learning. *Journal of Machine Learning Research* 8 (2007), 2443–2466.

[50] Jeremy Stoppelman. 2016. Yelp CEO says online reviews could beat "gold standard" healthcare measures. *Modern Healthcare* (2016).

[51] Tim Thompson. 2015. How Our Cities Are Using Social Data. *IBM Big Data & Analytics Hub* (2015).

[52] Zeynep Tufekci. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM* 14 (2014), 505–514.

[53] Emilio Zagheni, Kivan Polimis, Monica Alexander, Ingmar Weber, and Francesco C Billari. 2018. Combining Social Media Data and Traditional Surveys to Nowcast Migration Stocks. (2018).

[54] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 3 (2017), 689–722.

[55] Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. *arXiv preprint arXiv:1805.05345* (2018).