Predicting Food Safety Violations via Social Media to Improve Public Health Surveillance

Carmen Kar Hang Lee
School of Business, Singapore University of Social Sciences, Singapore
carmenleekh@suss.edu.sg

Abstract: Foodborne illness outbreaks originating in food service establishments (FSEs) are a serious public health concern. There are multiple ways that FSEs can contribute to the propagation of foodborne illnesses. However, not all these risks can be detected by health authorities during official inspections for two reasons. First, health authorities have limited resources, meaning not every FSE can be inspected as often and expediently as necessary. Second, because the food safety and hygiene standards of FSEs are dynamic, the inspection results might represent a snapshot that fails to resemble the current FSE situation. Consequently, it would be helpful to have an early warning system that can allow health authorities to predict the risk levels of FSEs and proactively intervene by demanding that high-risk FSEs take appropriate countermeasures. Considering that customers who feel sick after visiting an FSE may not report their experience to health authorities but instead disclose this information by posting online reviews on social media platforms, this study leverages social media data to predict food safety violations by FSEs. The empirical data for this study derive from two sources: the FSE inspection results from the Department of Health and Mental Hygiene in New York City, which labeled each FSE according to whether it violated food safety regulations, and textual reviews posted by FSE customers on Yelp. Machine learning algorithms were employed to extract textual features from the reviews and construct predictive models. Our results demonstrate that customer reviews posted on social media platforms are valuable for predicting FSE food safety violations. Health authorities can leverage social media data to construct predictive models and deploy them to optimize resource allocation and prioritize interventions to prevent foodborne illness outbreaks originating in FSEs, improving public health. FSE managers can also deploy the models to improve their hygiene standards in response to alerts raised. By taking appropriate countermeasures earlier, they can reduce their risk of financial loss due to negative inspection results.

Keywords: Public Health Policy, Foodborne Illness, Social Media, Customer Reviews, Predictive Modeling

1. Introduction

Although eating in food service establishments (FSEs) is an essential part of the lifestyle of many individuals, it is also a risk factor for foodborne illnesses (Choi et al., 2019; Green & Selman, 2005). To prevent FSE visitors contracting a foodborne illness, the food safety and hygiene standards of FSEs should be closely monitored. In many countries, health authorities conduct regular inspections of FSEs to investigate various aspects of their conduct, including kitchen cleanliness and food handling. If FSEs fail to comply with food safety regulations, they can be fined or forced to close immediately. Nonetheless, the effectiveness of the inspections and surveillance conducted by health authorities remains uncertain, with foodborne illness outbreaks often originating in FSEs. For instance, a foodborne illness outbreak originating at a catering organization in Singapore in 2018 affected more than 60 people, 45 of whom were hospitalized and 1 of whom died (Food Safety News, 2020). If FSEs that fail to comply with food safety regulations are identified and forced to take appropriate countermeasures earlier, foodborne illness outbreaks can be prevented, saving lives. Unfortunately, because health authorities have limited resources, FSEs are not inspected as often and expediently as necessary (Kang et al., 2013; Siering, 2021).

This study aims to develop predictive models that can identify FSEs with a high probability of violating food safety regulations. Previous studies have shown that social media data, such as online reviews, are a good source of information for public health surveillance (Harrison et al., 2014; Schomberg et al., 2016). This study leverages online reviews of FSEs posted on social media platforms (namely, Yelp), extracting textual cues that predict FSE inspection outcomes. The predictive models developed can be deployed as an early warning system that supports public health policy by enabling health authorities to prioritize their resources in favor of investigating high-risk FSEs. This approach not only prioritizes interventions but also prevents the public from suffering negative health consequences.

This paper is structured as follows. Section 2 summarizes the relevant literature. Section 3 introduces the study methodology. Section 4 presents the results. Section 5 discusses the results and concludes the study.

2. Literature Review

Foodborne illnesses are a serious public health concern that can be prevented by inspection and surveillance conducted by health authorities. However, although FSEs can contribute to the propagation of foodborne

illnesses in various ways, not all these risks can be detected by health authorities in a timely manner (Schomberg et al., 2016). Furthermore, the hygiene of FSEs can be very dynamic, changing over time. Consequently, the inspection results obtained during each inspection cycle might represent a snapshot that fails to represent the current hygiene standards of FSEs (Siering, 2021). This highlights the need for an early warning system that can allow health authorities to predict the risk levels of individual FSEs and proactively intervene by demanding that high-risk FSEs take appropriate countermeasures.

Researchers have recently started using social media platforms to conduct various types of public health investigations. Harrison et al. (2014) used online reviews to identify unreported foodborne illness cases. They discovered that a series of reviews described an incident consistent with the development of foodborne illness that was not reported to health authorities. Their results confirmed that there was publicly reported information about foodborne illness available on social media platforms. Therefore, health authorities can perform social media analytics for disease detection, identifying unreported outbreaks of foodborne illness and FSEs with deficient food handling protocols (i.e., FSEs that are more likely to violate food safety regulations). According to Siering and Janze (2019), severely negative health inspection results of FSEs impacted consumer perceptions of FSE service quality, leading to changes in their online review star ratings. This shows that, to a certain extent, online FSE reviews posted on social media platforms reflect the outcome of food safety inspections. This suggests that review features, such as star ratings, might be valuable for predicting FSE inspection results. Nonetheless, solely using star ratings is insufficient because that approach oversimplifies quality measures by assuming them to be unidimensional (Archak et al., 2011). In contrast, textual comments contain personal narratives of experiences with a specific service (von Helversen et al., 2018), providing an alternative that offers a more fine-grained understanding of the customer experience, better helping to predict FSE hygiene levels. Specific terminology appears in textual reviews that see customers focus on hygiene aspects. Siering (2021) selected particular words (e.g., “sick,” “dirt,” and ‘poisoning”) as features to represent online reviews and used them to predict FSE health violations. The results showed that classifiers that consider these textual features are valuable for detecting future health violations.

In addition, an academic consensus suggests that service quality affects customer satisfaction. In general, if customers experience an illness after visiting an FSE, their satisfaction will be lessened. Previous studies have used customer sentiments detected from textual reviews as a proxy for customer satisfaction (Lee, 2022; Wang et al., 2020). Siering (2021) used the sentiment polarity of reviews to predict the health violations of FSEs, demonstrating that sentiment polarity represents a useful predictor. Nonetheless, the sentiment polarity of the reviews only indicates whether the review is associated more with positive, negative, or neutral feelings. Emotions such as joy, fear, and anger were excluded from the study. In contrast, Peng et al. (2017) discovered that FSE stimuli, such as food quality, influenced FSE visitors’ emotions, including joy, excitement, peacefulness, anger, distress, and disgust. Given the relationship between FSE food handling performance and customer emotions, it is worth inquiring about the predictive power of customer emotions, in addition to sentiment polarity, for detecting FSE food safety regulations.

To summarize, this study leverages social media data by extracting a set of features from online reviews for prediction. Specifically, we aim to use (i) hygiene-related words, (ii) FSE visitor sentiments, and (iii) FSE visitor emotions detected from social media reviews to predict FSE food safety violations.

3. Methodology

The research methodology, as Figure 1 shows, comprised three phases: Data Collection, Feature Extraction, and Model Construction and Evaluation. The first phase involved collecting historical FSE inspection results and extracting FSE customer reviews from the social media platform Yelp. Based on the FSE name and the address, each customer review was matched with the corresponding FSE. The second phase involved aggregating reviews at the FSE level. For each FSE, the total count of hygiene-related words appearing in the reviews was computed, followed by the detection of the sentiments and emotions contained in the reviews. The outputs were used as the textual features for the proposed predictive models, with the average rating given by the customers of each FSE computed and used to construct baseline models. Using the performance of these baseline models, we could determine whether models leveraging textual data can give better prediction results. This study used k-fold cross-validation to construct and evaluate models. Details of each phase appear in the following sections.

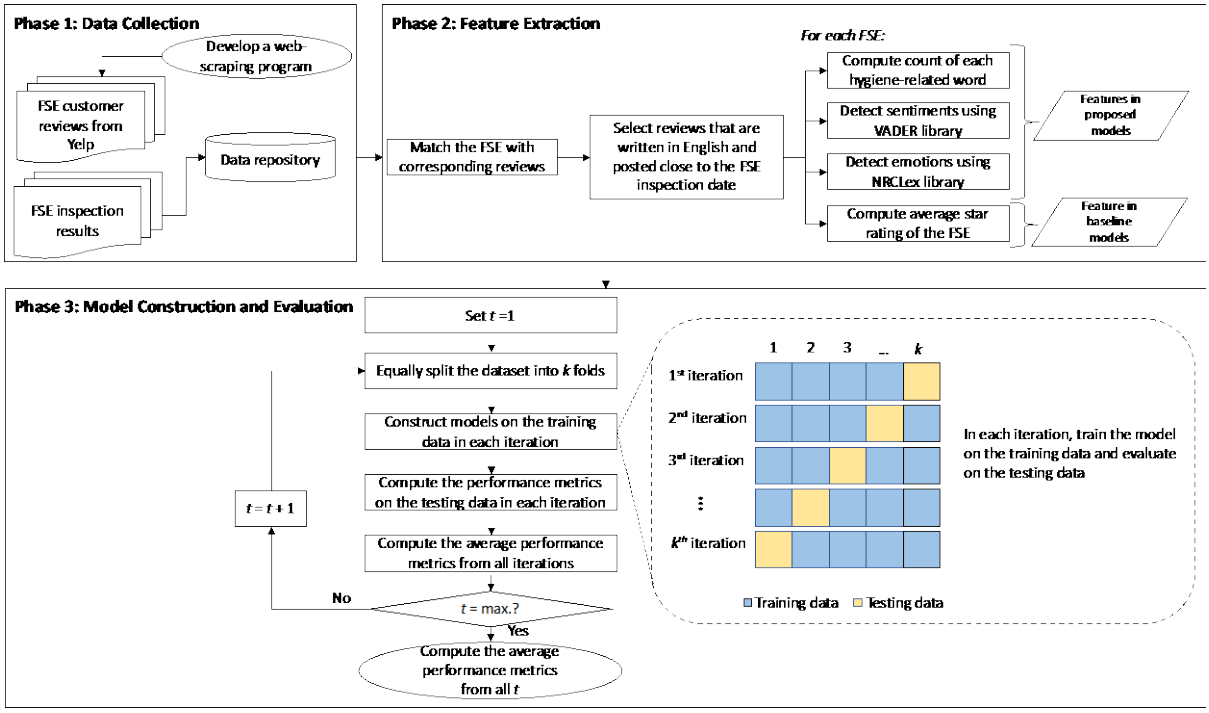

Figure 1: Study Methodology

3.1 Data Collection

The empirical data used for this study were open-source data from two sources. The first data source was the restaurant inspection results from the Department of Health and Mental Hygiene (DOHMH) in New York City. New York City Open Data allows access to such data via API. The dataset contains FSE inspection results for up to three years prior to the most recent inspection. For each FSE, there are fields for FSE name, inspection date, violation description, zip code, cuisine description, and critical flag. The critical flag indicates whether an FSE severely violated regulations. We used this as the target variable of the predictive models, that is, to observe whether it can classify whether an FSE is likely to severely violate regulations (i.e., critical flag = "Critical") or not (i.e., critical flag = "Not Critical"). This study only considered FSEs that had been inspected by the DOHMH, as indicated in the dataset (because only these FSEs have inspection results). Some FSEs might have been inspected more than once. In those cases, only the latest inspection was considered. The second data source was FSE reviews on Yelp. A web-scraping program was developed to extract reviews posted on Yelp. Each review contained fields for FSE name, reviewer name, review date, textual review, and star rating.

3.2 Feature Extraction

After data collection, data preparation was performed to extract features from the textual reviews. Based on the FSE name and the address, we matched the FSE with the corresponding Yelp reviews. Two selection criteria enabled review selection. First, we selected reviews posted a maximum of 90 days before the inspection date to ensure that the reviews reflected the situation of the FSEs at the time of inspection. Second, we selected reviews written in English. For each FSE, we aggregated all the reviews and counted the frequency with which the following hygiene-related words appeared: sick, dirty, poisoning, sticky, disgusting, filthy, ill, diarrhea, nausea, cramp, stomachache, vomit, rancid, and fever. Then, we conducted sentiment analysis. The VADER library on Python was chosen because it is sensitive to sentiment expression in social media contexts (Hutto & Gilbert, 2014). For a body of text, VADER generates ratios indicating the proportions of the text that falls into the three categories positive, neutral, and negative, as well as including a compound value (a commonly used measure of sentiment polarity). These three ratios and the compound value were used as the features of customer sentiments. Because VADER works best on short documents rather than large documents, we split each review into individual sentences to apply VADER at the sentence level. For each review, sentiment scores (positive, neutral, negative, compound) were the mean of the sentiment scores of all sentences, following the method of Stratigi et al. (2019). The overall sentiment scores associated with an FSE are the mean of the sentiment scores of all its reviews. Finally, we used the NRCLex library on Python to measure the emotional

affect contained in the reviews, detecting eight emotions: fear, anger, anticipation, trust, surprise, sadness, disgust, and joy. Based on the words in a review, it generates a score for each emotion. For each FSE, the emotion score is the average of all its reviews. Table 1 summarizes the features extracted from the textual reviews.

Table 1: Textual features used for predictions

Feature category	Features used	Description
Hygiene-related words	TF_p	Total word count of each hygiene-related word *p* in all reviews of an FSE
Sentiment	Avg_compound	The average compound value (sentiment polarity) of all reviews of an FSE
	Avg_pos	The average proportion of text appearing in all reviews of an FSE that is positive
	Avg_neg	The average proportion of text appearing in all reviews of an FSE that is negative
	Avg_neu	The average proportion of text appearing in all reviews of an FSE that is neutral
Emotion	Avg_fear	The average fear-related affect in all reviews of an FSE
	Avg_anger	The average anger-related affect in all reviews of an FSE
	Avg_anticip	The average anticipation-related affect in all reviews of an FSE
	Avg_trust	The average trust-related affect in all reviews of an FSE
	Avg_surprise	The average surprise-related affect in all reviews of an FSE
	Avg_sadness	The average sadness-related affect in all reviews of an FSE
	Avg_disgust	The average disgust-related affect in all reviews of an FSE
	Avg_joy	The average joy-related affect in all reviews of an FSE

3.3 Model Construction and Evaluation

To avoid issues with bias, three different machine learning algorithms were employed to predict whether an FSE was violating food safety regulations: decision tree, random forest, and support vector machine. To avoid overfitting, *k*-fold cross-validation was used to enable training and testing to be performed on different parts of the dataset. The cross-validation was repeated *t* times to remove any resampling bias (Van Nguyen et al., 2020). The overall performance of each model is based on the average of the metrics of the testing sets from all iterations. This study uses four performance metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

where *TP*, *TN*, *FP,* and *FN* represent true positive, true negative, false positive, and false negative, respectively. In this study, the positive case represents an FSE that violates food safety regulations, and the negative case represents an FSE that does not violate food safety regulations. Accuracy is defined as the total number of correct predictions divided by the total number of predictions. Precision concerns the number of correct positive predictions. It shows, of those predicted to be positive, how many are actual positive cases. Recall concerns the

number of correctly predicted positive cases and provides insight into how many actual positives are predicted as positive. The F1 score is considered the harmonic mean of precision and recall.

4. Results

After randomly extracting 180 New York City FSEs inspected by the DOHMH in 2021, we excluded FSEs with an undefined inspection outcome (i.e., critical flag = "not applicable"), leaving 178 FSEs. The inspections were conducted in the period between May 13, 2021, and December 27, 2021, and saw 52.2% of the FSEs severely violate regulations. Next, we checked whether the FSEs had a valid Yelp review page. FSEs not associated with any Yelp review were omitted. In our sample, all FSEs had associated Yelp reviews, so none were omitted. A total of 1,789 Yelp reviews were extracted.

To extract features from the textual reviews, we first computed the frequency of hygiene-related words. Figure 2 shows the frequency with which the chosen hygiene-related words appeared in the reviews. The most popular word is "sticky," followed by "disgusting" and "dirty," with 15.7% of the FSEs receiving a review containing at least one hygiene-related word. Among these FSEs, 64.3% had severely violated regulations.

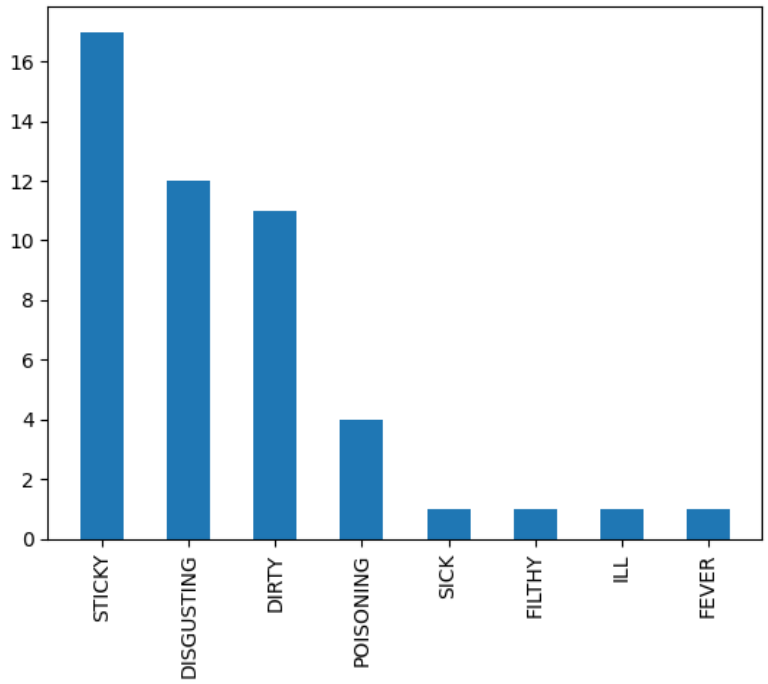

Figure 2: Frequency of hygiene-related words in the reviews

Using the VADER library, the sentiment of each review was detected. To verify the reliability of our sentiment analysis, we compared the compound values to the star ratings of the reviews. In Yelp, each review includes a 1–5 star rating. Higher scores imply greater customer satisfaction and, in turn, greater sentiment polarity. Figure 3 shows that reviews with higher ratings have higher compound values, and reviews with lower ratings have lower compound values, confirming the reliability of the sentiment analysis.

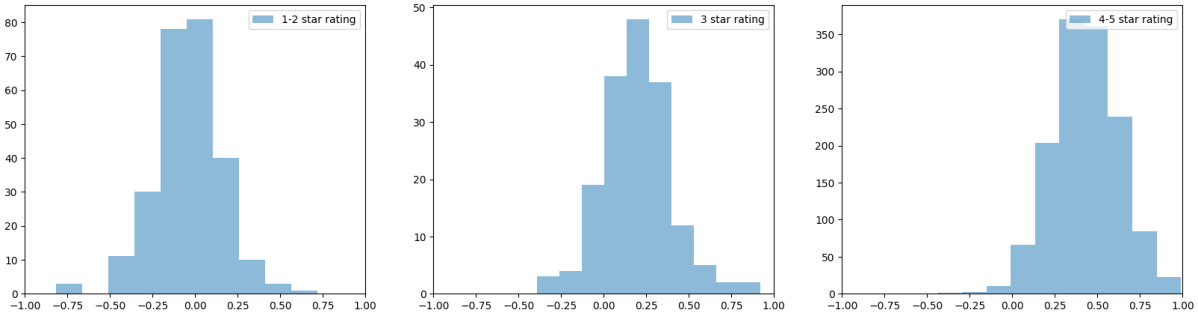

Figure 3: Sentiment polarity and overall rating of the reviews

After extracting the textual features, predictive models were constructed. To compare the performance of multiple models, we ran all models with stratified 3-fold cross-validation (i.e., k=3) and repeated the stratified 3-fold cross-validation a further three times (i.e., t=3). Table 2 summarizes the results.

Table 2. Prediction results based on the three-fold, three-time repeated stratified cross-validation

Algorithm	Model	Accuracy	Precision	Recall	F1 score
Decision tree	Baseline	47.2	49.4	41.9	45.3
	Proposed model	51.1	53.1	54.8	54.0
Random Forest	Baseline	48.1	50.4	47.7	49.0
	Proposed model	45.0	47.6	52.3	49.8
Support vector machine	Baseline	43.8	47.6	75.3	58.3
	Proposed model	52.8	53.0	86.0	65.6

Most of our proposed models leveraging online reviews from Yelp performed better than the baseline models in terms of accuracy, precision, recall, and F1 score, with the exception of the accuracy and precision of the random forest models, which were slightly lower than those of the baseline model.

Precision is a useful measure when a high cost is associated with *FP*. In our research context, *FP* indicates that an FSE that is not violating regulations is predicted to be violating regulations. In contrast, recall is useful when the cost of *FN* is high. *FN* indicates that an FSE that is violating regulations is predicted to be not violating regulations. If we deploy our model when recall is low, the consequences for the public are negative because customers may visit an unsafe FSE that the model determines to be safe. Therefore, from the perspective of public safety, recall is an important model evaluation metric. Lastly, the F1 score is a function of precision and recall that can be used to summarize the evaluation of a model using a single number.

Because this study aims to support public health policy by enabling health authorities to identify FSEs that are likely to be at risk of violating regulations and to prevent the public from contracting foodborne illnesses, the cost of *FN* is very high, meaning that recall is of considerable importance. We primarily evaluated model performance based on recall. The best-performing algorithm is support vector machines, which demonstrated recall of 86%, meaning that 86% of the FSEs with violations were predicted correctly. Deploying our model with a high recall can help health authorities identify FSEs that are likely to be at risk of violating regulations (i.e., predicted positive cases). Because health authorities may not have sufficient resources to inspect every FSE, they can use our model as an early warning system, allowing them to detect high-risk FSEs and plan inspections. This is valuable not only for optimizing resource allocation but also for preventing outbreaks of foodborne illnesses by forcing FSEs that violate food safety violations to close or adopt countermeasures to improve food safety. Our results illustrate that social media data (leveraged by the appropriate machine learning algorithms) can support public policy, substantially improving public health.

5. Discussion and Conclusion

This study leverages social media data to identify FSEs that are more likely to be at risk of violating food safety regulations. Our findings show that textual comments posted by FSE visitors on social media platforms are valuable for predicting violations. However, although we recognize the value of using social media data in the context of public health surveillance, we do not imply that actual inspections should be fully substituted with social media analytics. Instead, we argue that social media data, when appropriately analyzed, can complement current inspection policy by enabling health authorities to make more informed decisions when allocating resources and prioritizing interventions.

These findings have important implications for health authorities and FSE managers. First, health authorities can deploy our models to prioritize high-risk FSEs when scheduling official inspections. Although extracting textual cues from social media for analytics might be an expensive investment, the costs are justified by the benefits, which include reducing search costs and protecting the public against foodborne illnesses.

Our results also suggest that customers can act as hygiene inspectors who disclose inspection results on social media. This can provide FSE managers with access to almost-real-time feedback on food safety, precluding the need to passively await official inspection results from health authorities. If customers report food safety risks

on social media, FSE managers can review their food handling practices without delay. In this context, following up on customer reviews and taking appropriate countermeasures prior to official inspections is vital, not only for improving public health but also for minimizing the risk of financial losses due to a negative inspection outcome.

This study combines machine learning and public policy, responding to the call from Leist et al. (2021) for more research effort into applying machine learning to social and health research questions. We have presented a sophisticated research framework that city authorities can apply to support public policy, and we have confirmed how social media analytics can be used as a diagnostic tool to enhance FSE food safety, incorporating customer feedback into the FSE inspection process. Although this study concerns FSE inspection schemes, the results can be reasonably generalized to authorities in other fields that have access to large volumes of social media data. This represents a paradigm shift, with customer feedback available on social media platforms now having the potential to produce more thorough evaluations of policy effectiveness.

Nonetheless, this study has certain limitations, opening future research avenues. First, this study focused on FSEs from New York City, with evidence affirming the generalizability of the findings limited. Future research that includes FSEs from other cities and countries could resolve the generalizability issue. Second, although this study has shown that keywords related to hygiene aspects of FSEs in textual reviews can usefully predict the inspection results of FSEs, other aspects of FSEs, such as service and atmosphere, might relate to the outcome of inspections. Future research could use different text-mining approaches, such as topic modeling, to extract different service aspects of FSEs from social media data and use these as features in the application of our methodology for the purpose of comparing results. Third, this study did not consider the influence of the operating environment: FSEs operating in less competitive environments might pay less attention to food quality, putting them at a higher risk of violating food safety regulations. Future research could investigate how the performance of our models differs when the FSEs face different levels of competition in their operating environment.

Acknowledgements

This project is funded by Internal Seed Funding from Singapore University of Social Sciences (Project Code: RF00015).

References

Archak, N., Ghose, A. and Ipeirotis, P. G. (2011) “Deriving the pricing power of product features by mining consumer reviews”, *Management Science*, Vol 57, No. 8, pp. 1485-1509.

Choi, J., Nelson, D. and Almanza, B. (2019) “Food safety risk for restaurant management: use of restaurant health inspection report to predict consumers’ behavioral intention”, *Journal of Risk Research*, Vol 22, No. 11, pp. 1443-1457.

Food Safety News (2020) “Caterer in Singapore fined for 2018 food poisoning outbreak”, [online], https://www.foodsafetynews.com/2020/12/caterer-in-singapore-fined-for-2018-food-poisoning-outbreak/

Green, L. R. and Selman, C. (2005) “Factors impacting food workers’ and managers’ safe food preparation practices: A qualitative study”, *Food Protection Trends*, Vol 25, No. 12, pp. 981-990.

Harrison, C., Jorder, M., Stern, H., Stavinsky, F., Reddy, V., Hanson, H., Waechter, H., Lowe, L., Gravano, L. and Balter, S. (2014) “Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York City, 2012–2013”, *MMWR*, Vol 63, No. 20, pp. 441–445.

Hutto, C. J. and Gilbert, E. (2014) “VADER: A parsimonious rule-based model for sentiment analysis of social media text”, in Proceedings of *the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14),* Ann Arbor, MI, USA, pp. 216–225.

Kang, J. S., Kuznetsova, P., Luca, M. and Choi, Y. (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*, Seattle, WA, USA, pp. 1443-1448.

Lee, C. K. H. (2022) “How guest-host interactions affect consumer experiences in the sharing economy: New evidence from a configurational analysis based on consumer reviews”, *Decision Support Systems*, Vol 152, Article 113634.

Leist, A. K., Klee, M., Kim, J. H., Rehkopf, D. H., Bordas, S., Muniz-Terrera, G. and Wade, S. (2021) “Machine learning in the social and health sciences”, *arXiv preprint* arXiv:2106.10716.

Peng, N., Chen, A. and Hung, K.-P. (2017) “The effects of teppanyaki restaurant stimuli on diners’ emotions and loyalty”, *International Journal of Hospitality Management*, Vol 60, pp. 1-12.

Schomberg, J. P., Haimson, O. L., Hayes, G. R. and Anton-Culver, H. (2016) “Supplementing public health inspection via social media”, *PLoS ONE*, Vol 11, No. 3, Article e0152117.

Siering, M. (2021) “Leveraging online review platforms to support public policy: Predicting restaurant health violations based on online reviews”, *Decision Support Systems*, Vol 143, Article 113474.

Siering, M. and Janze, C. (2019) “Information processing on online review platforms”, *Journal of Management Information Systems*, Vol 36, No. 4*,* pp. 1347-1377.

Stratigi, M., Li, X., Stefanidis, K. and Zhang, Z. (2019) “Ratings vs. reviews in recommender systems: A case study on the Amazon Movies Dataset”, In: Welzer, T. et al. (eds.) *New Trends in Databases and Information Systems. ADBIS 2019. Communications in Computer and Information Science*, Vol 1064, Springer, Cham.

Van Nguyen, T., Zhou, L., Chong, A. Y. L., Li, B. and Pu, X. (2020) “Predicting customer demand for remanufactured products: a data mining approach”, *European Journal of Operational Research*, Vol 281, No. 3, pp. 543-558.

von Helversen, B., Abramczuk, K., Kopeć, W. and Nielek, R. (2018) “Influence of consumer reviews on online purchasing decisions in older and younger adults”, *Decision Support Systems*, Vol 113, pp. 1-10.

Wang, Y., Zhang, M., Tse, Y. K. and Chan, H. K. (2020) “Unpacking the impact of social media analytics on customer satisfaction: do external stakeholder characteristics matter?”, *International Journal of Operations & Production Management*, Vol 40, No. 5, pp. 647-669.