Predictive Analytics Using Text Classification for Restaurant Inspections

Zhu Wang, Booma Sowkarthiga Balasubramani, Isabel F. Cruz
University of Illinois at Chicago
{zwang260,bbalas3,ifcruz}@uic.edu

ABSTRACT

According to the Center for Disease Control (CDC), there are almost 48 million people affected by foodborne diseases in the U.S. every year, including 3,000 deaths. The most effective way of avoiding food poisoning would be its prevention. However, complete prevention is not possible, therefore Public Health departments perform routine restaurant inspections, combined with the practice of inspecting specific restaurants once a disease outbreak is identified. Following other health applications (e.g., prediction of a flu outbreak using Twitter), we use social media and a predictive analytics approach to identify the need for targeted visits by city inspectors.

CCS CONCEPTS

• **Information systems → Data mining**; *Web mining*; • **Computing methodologies → Machine learning**; *Natural language processing*; • **Applied computing → Life and medical sciences**;

KEYWORDS

Text mining, supervised learning, prediction models, online reviews, public health

1 INTRODUCTION

People are severely affected by foodborne illnesses in their daily life. According to the Center for Disease Control (CDC), there are almost 48 million people affected by foodborne diseases in the U.S. every year, including 3,000 deaths [4]. Also, every year in the U.S. there are over 128,000 people hospitalized due to food-related disease infections. This issue poses a serious threat to public health, especially in big cities with a large number of residents and visitors, such as New York City, Las Vegas, or Chicago. Therefore, their Public Health departments inspect restaurants, food markets, and other foodservice establishments at least once a year. However, this method is not without problems, for instance, a restaurant can prepare to pass their regular annual inspection but not meet the legal requirements during the rest of the year.

The best approach for a foodservice establishment to avoid foodborne diseases includes following all food codes, training the staff, and conducting self inspections regularly [21]. However, complete prevention of the conditions that lead to foodborne diseases is a challenge [5].

UrbanGIS'17, November 7–10, 2017, Redondo Beach, CA, USA

ACM ISBN 978-1-4503-5495-0/17/11.
https://doi.org/10.1145/3152178.3152192

Following other health applications (e.g., prediction of flu outbreaks using Twitter), we investigate the use of Yelp to predict foodborne illnesses. The Yelp datasets contain a large number of reviews from many active users in each metropolitan city. For example, the Yelp academic dataset for Boston has more than 5,800 users who have published about 235,000 reviews. Yet, the lack of credibility of these reviews can decrease the accuracy of the systems that use those datasets. For example, some restaurants use paid services that write fake negative reviews about their competitors [12]. For accuracy, those fake reviews should not be considered. While Yelp uses algorithms to detect and filter out such reviews, additional mechanisms can be used, for example to detect users who post few but very negative reviews.

Previous studies have examined the helpfulness of consumers' online reviews and ratings in the Health Care and Emergency domains, even though reviewers may lack medical knowledge [10, 19]. One such example review from Yelp is *"DON'T EAT HERE. I got really bad food poisoning after eating the chicken pad thai."* Considering these, we retrieve and analyze Yelp reviews to predict foodborne illnesses in restaurants, with the goal to prevent more people from being affected by such illnesses. This framework could be used to alert other users and report the restaurants to the Public Health department of the city along with the time window and the kind of foodborne illness every time there is a user review indicating food poisoning. However, there are hundreds of millions of reviews on Yelp and therefore purely manual methods would be very inefficient. In our approach, we use machine learning algorithms on the existing reviews to generate features that allow us to develop prediction models.

This paper is organized as follows. In Section 2, we present briefly some of the existing techniques that analyze social media (such as Yelp or Twitter) with the aim to prevent foodborne diseases. In Section 3, we introduce our system and its components, including extraction of language features, feature selection, and prediction models. Section 4 presents the evaluation of our prediction models for the Yelp academic dataset for Boston. Finally, we summarize the paper and discuss future research in Section 5.

2 RELATED WORK

A previous paper [9] involves the identification of unreported cases of foodborne illness in New York City. Their framework uses keywords to narrow down 294,000 Yelp reviews to 893 reviews containing information about possible unreported illness cases. The reviews were sifted using manual labeling, which led to the identification and inspection of three restaurants with related issues.

There is also a previous study [12] on recognizing fake reviews on Yelp by analyzing the reasons for such fake reviews, including reputation and competition for business incentives. The authors

found that fake reviews tended to be extreme, positive fake reviews occurred primarily for restaurants with weak reputations, restaurants with increasing competitors received a higher number of unfavorable fake reviews, and chain restaurants were less likely to create fake reviews about other businesses.

There are works [2, 15] that consider online reviews evaluating hospital care quality. An adaptive and real-time system called "nEmesis" [16, 17] aimed at the prevention of foodborne illnesses, using weighted SVM language model on data collected from Twitter. The results show that *nemesis* has helped in preventing over 9,000 cases of foodborne illness and 557 hospitalizations annually.

To the best of our knowledge, this paper persents the first study that envisions the use of prediction models to recognize possible foodborne illness through online consumer reviews.

3 SYSTEM ARCHITECTURE

The proposed system for predictive analysis consists of language features and embedded classification models. The dataset is split into test and training data. The first step is to generate the ground truth, and is performed based on the training reviews and ratings which point to users suffering from foodborne illness. A binary classifier is then used to predict whether the test reviews are indicators or not. In the process of building models, we extract: (a) statistical features of users and reviews, and (b) language features extracted from the review corpus. The architecture of the system is shown in Figure 1.

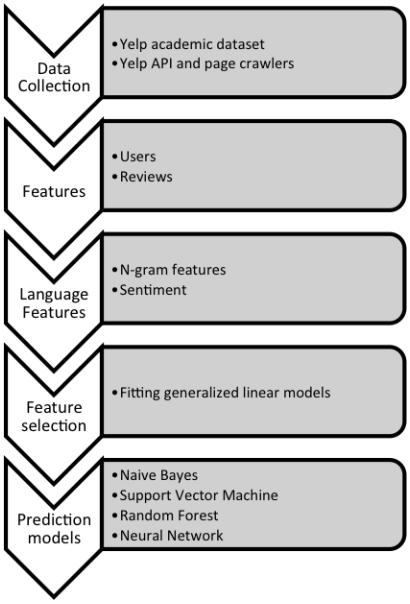

Figure 1: Proposed system.

Typical features extracted based on *users* and *reviews* include user location, user with profile image, number of reviews per user, average rating per user, reviews per restaurant, review date, and overall rating per restaurant. The average ratings of restaurants is 3.78, and the average number of reviews per user is 42.53. User features are mainly used to verify the users' credibility and assign weights to their reviews.

3.1 Language Features

For text classification tasks, each word can be treated as a feature, but it is expensive to directly use words as features. Therefore, we convert the textual representation of information into a Vector Space Model using TF-IDF [20]. Also, the language features are extracted from emotional words, and N-gram [7] is used to improve accuracy based on the context.

Following previous studies, we explore reviews with keyword sets that denote foodborne illness. For example, {sick, vomit, diarrhea, poison, stomach} form a keyword set. In the sampled Yelp academic dataset, there are 189 reviews that contain these keywords. The average ratings of these reviews are much lower than the overall ratings. We apply reviews with/without keywords as a binary predicator variable. The summary is shown in Table 1.

Table 1: Summary of reviews with keywords.

Keywords	Counts	Avg. ratings	Avg. length (N of words)
sick*	106	2.8	184
vomit*	12	1.7	181
diarrhea*	4	2.3	193
poison*	29	1.9	190
stomach*	38	2.1	183

3.1.1 Frequency. Sentence counts and word counts are significant predictors for food poisoning and fake reviews. Typically, fake reviews are posted by payable services, and are charged based on word counts. Therefore, we consider reviews that are longer as providing more evidence. Also, reviews with a single sentence or less than ten words were not considered useful for predictions, hence we ignored them to improve accuracy.

3.1.2 Sentiment with lexicons. LIWC [18] provides comprehensive dictionaries that can be used to extract language features related to the foodborne illness domain, including health, anger, anxiety, negative, ingest, and swear terms. In addition, we add our dictionary terms to LIWC features variables, as listed in Table 2, where N represents the number of words in each category. For example, if the user reviews contained "never" or "pill", then the reviews are more likely to predict the restaurants that require inspections. Once we detect such terms from each review from different categories, we count the terms and obtain ratios as language features.

For negative features, we calculate a sentiment score for each review r in the corpus:

$$negative(r) = \frac{freq(negmo)}{freq(negmo) + freq(posmo)} \quad (1)$$

where *freq(negmo)* is the count of negative terms, and *freq(posmo)* is the count of positive terms both detected by LIWC in each review.

For other features, we calculate their scores for each review r in the corpus:

$$feature(r) = \frac{freq(feature)}{freq(word)} \quad (2)$$

where *freq(feature)* is the count of each categorical terms detected by LIWC and *freq(word)* is the total word count in each review.

Table 2: Language feature terms - Examples.

Category	Abbrev	Examples	N
Health	health	ache*, allerg*, dyspeps*	238
Biology process	bio	anal*, appeti*, breath*	572
Body	body	anus*, belly, crotch*	183
Anger	anger	afraid*, alarm*, doubt*	186
Anxiety	anx	discomfort*, enemie*, fake*	92
Negative emotion	negmo	abandon*, fear*, harm*	501
Positive emotion	posmo	admir*, happy*, enjoy*	410
Negations	negate	must'nt, neither, never	98
Ingestion	ingest	ate, boil*, chew*	150
Swear	swear	arse, crap, damn*	55
Sadness	sad	cry, depress*, disappoint*	102
Inhibition	inhib	avoid*, ban*, block*	114
Death	death	bereave*, dying*, epidemic*	64

3.2 Feature Selection

Once language features are extracted from the reviews, we need to verify whether all features are significant for prediction. We also add the features related to ratings, because the ratings ranging from 1 to 5 implies the differences in the attitude of the users. Therefore, we transfer the ratings feature from numerical to categorical.

We fit generalized linear models to examine different combinations of predictors and drop insignificant predictors based on *p-value*. Then we compare the performance of each linear model based on *Accuracy*, *Precision*, *Recall*, and *F-score* to select the most efficient features [8].

3.3 Predictions Models

As we have reviews with keywords set as ground truth, we utilize supervised learning techniques for classifying high-dimensional data in order to predict whether a review indicates foodborne illness. During the training process, we also re-weight each feature with *k-fold cross-validation* to improve the performance. The reviews related to foodborne illness have sparse features, so we need to evaluate a range of γ parameter in the models. The models which we have considered are as follows:

(1) *Naïve Bayes (NB)* [13] is a probabilistic model with maximum likelihood to classify categories. We use NB to compare the results of different distributions.
(2) *Support vector machines (SVM)* [6] represents examples as points in space for non-linear classification. Different kernels lead to differences in the performance based on the dataset.
(3) *Random forest (RF)* [3] is an ensemble technique that fits a number of decision tree classifiers on various sub-samples of the dataset. We have experimented with different number of decision trees as estimators.
(4) A *recurrent neural network (RNN)* [1] is a class of neural networks with a directed cycle. After converting indexes of words into an embedding matrix with other features, it maps word indexes as a sequence into the matrix. Then, we create a gated recurrent unit cell with hidden size of embedding size and to pass word as input for each unit.

4 EXPERIMENTS AND EVALUATION

4.1 Dataset and Parameter Settings

Due to the limitations of the Yelp API, we first apply the Yelp academic dataset for restaurants in Boston to build and evaluate models. An N-gram feature is a high dimension matrix, but the reviews containing keywords are sparse and imbalanced. Also, text classification is a time consuming task. Therefore, we selected 15,213 reviews randomly to extract reviews with keywords, then applied to the original dataset to predict.

Raw data from the reviews are noisy, because the users have posted them with many words that are not in the English corpus, special characters, punctuations and so on, which leads to a sparse vector space and an increase in the runtime and storage. Thus, the data pre-processing tasks involved removing words which were not in English Corpus, URLs, special characters and stopwords, and standardizing the corpus to lower cases. The NLTK toolkit[1] is used to perform data pre-processing.

Using selected features, we construct prediction models with NB, SVM, RF, and RNN. We run 10-fold cross validation to split the dataset into training sets and test sets. RNN is implemented using the Scikit-learn[2] library. The parameter settings for each model are as follows:

- *NB distributions:* Multinomial and Gaussian.
- *SVM kernel:* Radial, sigmoid and polynomial, which we operated to fine-tune the model with different kernel and parameter settings. The γ parameter ranges from 0.000001 to 0.1 and *cost* parameter ranges from 0.1 to 10.
- *RF trees:* Number of trees - 10, 30, and 50.
- *RNN:* The size of the hidden layer is 15, and the learning rate is constant with 0.001.

4.2 Results

After performing text classification, we used the models to predict whether the review indicates foodborne illness with high probabilities. We set the best performing generalized linear model as the baseline, and compare the results of our prediction models.

First, we examine the effects of parameter settings on each classification model. The results show that the multinomial NB model has better results, indicating that the dataset tends to have multinomial distribution. Similarly, SVM with a sigmoid kernel outperforms other kernels. It is also to be noted that RF with 30 trees performs better than RF with 10 and 50 trees.

Finally, we compare the performance of each predicting model. Figure 3 shows the performance of different classification algorithms with their best parameters, including accuracy, precision, recall and F-score measures. It is evident that all the prediction models outperform the generalized linear model, and SVM and RNN have better performance than others with higher accuracy and F-scores. NB model predicts more correctly, which means that part of the features have higher probabilities to predict foodborne illness. However, all recall values are lower than 70%, because the dataset is imbalanced and illness related features are sparse.

[1] http://www.nltk.org/
[2] http://scikit-learn.org/stable/

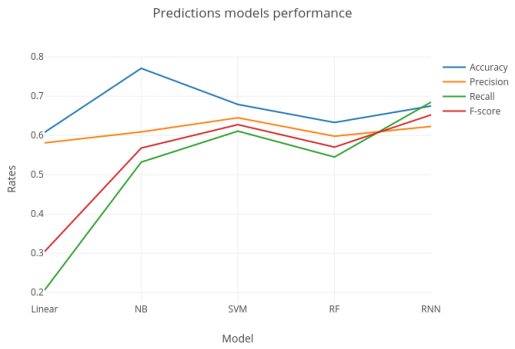

Figure 2: Performance of prediction models.

5 CONCLUSIONS AND FUTURE RESEARCH

People tend to write down their feelings online rather than reporting issues to government authorities. If they suffer from foodborne illness after visiting a restaurant, they post reviews to alert other users. We utilize these resources to automatically detect potential foodborne illness and to flag related business for inspection.

In this paper, we have introduced our vision of constructing a real-time inspection system combined with public health department routine inspections to prevent users from suffering foodborne illness. In particular, we have detected reviews that reliably predict foodborne illness by selected features and classification methods.

The Yelp academic dataset for Boston helped us understand the correlation between ratings and reviews of a business, for example, negative reviews are usually given lower ratings. All language features are significant in the linear model, which means that related keywords terms can be used to predict food poisoning in user reviews. However, determining whether a review indicates foodborne illness is complex, because real world datasets are noisy and people use different styles of writing, languages and sentences online, which further increases the difficulty to decide whether a review could be used as a predictor. Though there are explicit messages, for example, "it is sick", it is quite difficult to understand whether it represents the food quality or it indicates the service of the restaurant. Therefore, we plan to focus on semantic analysis based on aspect-term to label reviews in categories [11, 14] such as food quality, service, ambience, and so on.

Future work includes collecting real-time data from Yelp and training the data adaptively with embedded methods in order to improve prediction performance. An interactive view of the system for the network of users and restaurants is also among our future goals. Twitter is another social network that displays people's feelings, for example, after dining at a restaurant. Tweets with geo-information can be connected with nearby food venues and used to identify related users who have suffered foodborne illness [16]. This would enable combining Twitter results and Yelp results to refine a hybrid prediction model. This approach could also be extended to other applications, such as predicting the increase or decrease in the number of malaria cases in a region by extracting features indicating malaria from tweets.

Acknowledgments

We thank Tom Schenk, Chief Data Officer of the City of Chicago, and his team at the Department of Innovation & Technology. This work was partially supported by NSF awards CNS-1646395, III-1618126, CCF-1331800, and III-1213013, by a Bill & Melinda Gates Foundation Grand Challenges Explorations grant, and by the Bloomberg Philanthropies Mayor's Challenge Award "Chicago SmartData Platform."

REFERENCES

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. TensorFlow: Large-scale Machine Learning on Heterogeneous Distributed Systems. *CoRR abs/1603.04467* (2016).
[2] Bardach, N. S., Asteria-Peñaloza, R., Boscardin, W. J., and Dudley, R. A. The Relationship between Commercial Website Ratings and Traditional Hospital Performance Measures in the USA. *BMJ Quality & Safety* (2012), 194–202.
[3] Breiman, L. Random Forests. *Machine Learning 45*, 1 (2001), 5–32.
[4] CDC. Burden of Foodborne Illness: Overview, July 2016.
[5] Choucair, B. The Future of Public Health: Preventing Food Poisonings from Occurring, February 2014.
[6] Colas, F., and Brazdil, P. Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. *Artificial Intelligence in Theory and Practice* (2006), 169–178.
[7] Ghiassi, M., Skinner, J., and Zimbra, D. Twitter Brand Sentiment Analysis: A Hybrid System Using n-gram Analysis and Dynamic Artificial Neural Network. *Expert Systems with Applications 40*, 16 (2013), 6266–6282.
[8] Goutte, C., and Gaussier, E. A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation. In *European Conference on IR Research (ECIR)* (2005), vol. 5, Springer, pp. 345–359.
[9] Harrison, C., Jorder, M., Stern, H., Stavinsky, F., Reddy, V., Hanson, H., Waechter, H., Lowe, L., Gravano, L., Balter, S., et al. Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness-New York City, 2012–2013. 441–445.
[10] Kilaru, A. S., Meisel, Z. F., Paciotti, B., Ha, Y. P., Smith, R. J., Ranard, B. L., and Merchant, R. M. What Do Patients Say About Emergency Departments in Online Reviews? A Qualitative Study. *BMJ Quality & Safety 25*, 1 (2016), 14–24.
[11] Kiritchenko, S., Zhu, X., Cherry, C., and Mohammad, S. NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews. In *SemEval@COLING* (2014), pp. 437–442.
[12] Luca, M., and Zervas, G. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *Management Science 62*, 12 (2016), 3412–3427.
[13] Metsis, V., Androutsopoulos, I., and Paliouras, G. Spam Filtering with Naive Bayes – Which Naive Bayes? In *CEAS* (2006), vol. 17, pp. 28–69.
[14] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Workshop on Semantic Evaluation (SemEval–2016)* (2016), Association for Computational Linguistics, pp. 19–30.
[15] Ranard, B. L., Werner, R. M., Antanavicius, T., Schwartz, H. A., Smith, R. J., Meisel, Z. F., Asch, D. A., Ungar, L. H., and Merchant, R. M. Yelp Reviews of Hospital Care Can Supplement and Inform Traditional Surveys of the Patient Experience of Care. *Health Affairs 35*, 4 (2016), 697–705.
[16] Sadilek, A., Brennan, S., Kautz, H., and Silenzio, V. nEmesis: Which Restaurants Should You Avoid Today? In *AAAI Conference on Human Computation and Crowdsourcing* (2013).
[17] Sadilek, A., Kautz, H. A., DiPrete, L., Labus, B., Portman, E., Teitel, J., and Silenzio, V. Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media. In *AAAI* (2016), pp. 3982–3990.
[18] Tausczik, Y. R., and Pennebaker, J. W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology 29*, 1 (2010), 24–54.
[19] Tran, N. N., and Lee, J. Online Reviews as Health Data: Examining the Association between Availability of Health Care Services and Patient Star Ratings Exemplified by the Yelp Academic Dataset. *JMIR Public Health and Surveillance 3*, 3 (2017).
[20] Ullman, J. D. *Mining of Massive Datasets.* Cambridge University Press, 2011.
[21] webstaurantstore. Preparing for a Health Inspection, September 2017.