e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-1090

Monitoring Food Safety Violation Reports from Internet Forums

Kiran KATE[a], Sumit NEGI[b], Jayant KALAGNANAM[c]
[a]*IBM Research Collaboratory, Singapore*
[b]*IBM Research India, New Delhi*
[c]*IBM T.J. Watson Research Center, Yorktown Heights, New York, USA*

Abstract. Food-borne illness is a growing public health concern in the world. Government bodies, which regulate and monitor the state of food safety, solicit citizen feedback about food hygiene practices followed by food establishments. They use traditional channels like call center, e-mail for such feedback collection. With the growing popularity of Web 2.0 and social media, citizens often post such feedback on internet forums, message boards etc. The system proposed in this paper applies text mining techniques to identify and mine such food safety complaints posted by citizens on web data sources thereby enabling the government agencies to gather more information about the state of food safety. In this paper, we discuss the architecture of our system and the text mining methods used. We also present results which demonstrate the effectiveness of this system in a real-world deployment.

Keywords. Food-borne illness, Food safety, Food safety violation, Text mining

Introduction

Food-borne illness is a growing public health concern around the world. The US Centers for Disease Control and Prevention (CDC) estimated that roughly 1 in 6 Americans or 48 million people fell ill, 128,000 were hospitalized and 3,000 died of food-borne diseases each year [1]. Many countries have setup special government agencies that monitor and act on complaints related to food safety, which includes complaints received from citizens with regard to food hygiene, food poisoning incidents. These government agencies actively seek inputs/feedback from citizens using traditional channels such as call centers, e-mails etc. Once sufficient or severe violations are reported against an establishment (supermarket, public or private canteens, restaurants etc.), the government agency acts on these complaints by carrying out a physical inspection of the facility followed by punitive actions that include fine, closure or license revocation of the establishment. The current process of gathering such feedback from citizens has its own limitations. Due to the formal nature of the process not many citizens report such incidents. Also the incidents that are reported through the traditional channels are not current i.e. reports/complaints are filed days or weeks after the actual incident took place. These facts severely restrict the amount and recency of information available with the government agency with respect to food safety violations.

To address the above limitation we propose a solution that sifts through several internet forums (user forums, citizen journalism websites, blogs etc.) looking for "up-to-date'' information related to food safety violations as reported by citizens on such websites. Our work is similar in spirit to recent work on using citizen' updates on social media to characterize electoral debates [2], detect real time events such as natural calamities [3] etc. Considering the fact that these forums also contains discussion threads and posts on other topics such as politics, entertainment, technology etc automatically detecting posts that are relevant to food safety violations requires the use of text mining methods. Our proposed approach, which was deployed for a real world government agency, employs state of art machine learning/text mining techniques to automatically *detect* and *process* citizen posts that mention food safety violations.

1. Methods

The first task is to (periodically) crawl internet forums that are of interest. As mentioned earlier, an internet forum receives thousands of user posts daily of which only a few mention food safety violations (food safety violation posts could include posts that mention food hygiene lapses in food establishments, supermarkets or shops selling expired food, or food poisoning incidents after consuming food etc.). Identifying such user posts automatically is done using the *text classification* module as part of the *Text Analytics* layer. Once such a post has been identified the next step is to (automatically) process this post by extracting relevant nuggets of information from the post such as the name of the food establishment, the address/location etc. This is done using the *Entity detection and extraction* module which is part of the *Text Analytics* layer. Due to space constraints this paper provides details of only the *text classification* module.

2. Text Analytics Component Description

Any popular internet forum usually receives a few thousand new user posts every day. Not all of these postings mention food safety violations. A key challenge, which is addressed by the *text classification* system, is to automatically identify which user post mentions food safety violations. To be able to do this automatically requires domain specific feature engineering and deciding which classification model to use. We discuss these two topics in Section 2.1 and 2.2 respectively.

2.1. Feature Engineering

Feature engineering is the step of identifying textual features which will help a classification model differentiate food safety violation reports from posts on other topics. It is a critical step that affects performance of a classification model. For our setting, we experimented with a combination of features described below:

1. *Features derived from training data:* Lexical features such as unigrams (single words) and bigrams (pairs of words). We use TF-IDF scores on these after removal of stop-words.

2. *Features capturing sentiments:* Complaints often use words that express negative sentiments such as “bad", “disgusting" etc. Adding a manually crafted dictionary of negative words as features showed improvement in the classification accuracy. One could also use any of the state of the art sentiment analysis techniques to classify a post into one of (positive, negative, neutral) classes and then use this label as a feature. However, such a label would not capture the severity of occurrence of the negative words which we intend to capture using the dictionary. We also observed that complaint articles were relatively short compared to other types of discussions (personal experiences detailing a story, political discussions, discussions and stories about celebrities etc). Using the length of an article as a feature further improved the classification performance.
3. *Domain specific features:* We also observed that using domain specific annotations such as food names, food center names as features improved classification accuracy.

Through our experiments we observe that different feature types contribute differently to the overall classification performance. A weighted combination of these three feature types was used for our classification task. A grid search with 5-fold cross validation on the test data set as described in the results section was used to find the optimum weights (according to F1 score) for the three types of features. The final weights we use are (2, 5, 5) for the three feature types described above in that order.

2.2. Text Classification Models

Text classification is the process of building a model that distinguishes posts that report food safety violation from posts on other topics. This distinction is learnt from a training data set that contains posts reporting food safety violations as well as posts on other topics. This fits into a binary classification task where the class labels are *Food_Safety* and *Non-Food_Safety*. We experimented with different state of the art classification algorithms including Multinomial Naive Bayes, k-NN and Support Vector Machine (SVM) with different parameters. For our setting, linear SVM gave the best performance and hence we report results with that method. A practical problem encountered in learning a classifier for the food safety domain is the class distribution skew. Most of the posts are crawled from citizen web forums where articles on other topics (e.g. entertainment, tourism) outnumber food safety articles approximately by 1:40. The problem of learning from imbalanced data is a relatively new challenge that has attracted growing attention from both academia and industry [5]. To address this challenge, we experiment with sampling (over-sampling and under-sampling) for imbalanced learning and report our results.

3. Results

Our system was deployed for the environment health department of a real world government agency. In Section 3.1 we provide an overview of the data sets and methods used to obtain those data sets. We report the performance of the classification model with different feature sets in Section 3.2.

3.1. Dataset

The system was used to detect food safety related complaints from a popular citizen journalism website in the country. The website has news stories on a variety of topics including movie gossips, science news, local events, food safety complaints, experiences with public transport etc. A total of 14722 user posts were collected by crawling this website. Subset of the training data for the classification task was obtained by hand labeling 1000 web-pages, where 20 belong to the *Food_Safety* class and the rest to the *Non-Food_Safety* class. We obtained additional training data by following two approaches*: (i) With the help of Wikipedia category tree:* We used Wikipedia category tree as described in [4] to obtain 407 samples of *Food_Safety* class and 13335 samples of *Non-Food_Safety* class. The training data obtained in this manner is referred to as "*gen-data"* in the discussion of results (ii) *From the agency's call center data:* The government agency we worked with also has a dedicated call center which logs citizen's food safety related complaints (citizens can call this call center to inform the government body about any food safety related observations/complaints). These calls are transcribed by call center agents and serve as training data for our classifier after some. We added randomly selected 407 complaints as examples to the *Food_Safety* class. The training data obtained in this manner is referred to as "*call-center-data"* in the discussion of results. The test set contains 52 posts from the *Food_Safety* class and 824 posts from the *Non-Food_Safety* class. These posts were hand labeled for the experiments.

3.2. Evaluation

Table 1 summarizes the results of classification. We report the standard information retrieval metrics precision, recall and F1 score only for the positive i.e. *Food_Safety* class since this is the class of interest.

Effect of training data generation: The first set of results, as indicated by the "*gen-data"* columns of Table 1 is obtained by using the training data set generated using the approach described in 3.1. The F1 score (using plain TF-IDF scores on unigram and bigram as features) for the *"call-center-data"* setup is higher than "*gen-data"* as shown in the first row of Table 1. This is expected since the call center training data is more "on-topic" as compared to the generated training dataset. It should be observed that the recall is significantly affected because the class distribution is highly imbalanced in the *"gen-data"* data set. However, as we add more informative features, it can be seen that the performance of the classifier improves, demonstrating the effectiveness of feature engineering.

Performance with different feature sets: The different rows in table 1 show the effect of using different types of features as discussed. The baseline TF-IDF scores on unigrams and bigrams, TF-IDF scores along with domain features, TF-IDF scores in conjunction with negative-words and a weighted combination of all these features together. It can be observed that adding informative features improves the overall performance, especially for model built using the generated training data. The improvement in performance is not very significant when the training data is obtained from the call center, and the distribution of classes is not very skewed.

Performance of methods to handle class imbalance: Under-sampling was performed by choosing random samples from the *Non-Food_Safety* class for creating a training set with more balanced distribution of the two classes. We performed experiments with

different values for the number of *Non-Food_Safety* samples and observed that the precision increases as we increase the number of *Non-Food_Safety* samples, while the recall drops. This behavior is expected as with a more balanced training set, more articles in the test set tend to be labeled as *Food_Safety* complaints, resulting in higher recall and lower precision. We report a precision of 0.75, recall of 0.71 and F1 score of 0.73 for this setup. Over-sampling was performed by repeating samples of the *Food_Safety* class, we observed that different class distributions achieved by this method did not change the performance much and we report the best results. We report a precision of 0.76, recall of 0.69 and F1 score of 0.72 for this setup. We also experimented with the misclassification cost parameters of SVM, and all the results reported in the table are with optimum misclassification costs.

Table 1. Classification performance with different feature sets and training datasets

Feature Types	**gen-data**			**call-center-data**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
TF-IDF	0.875	0.2692	0.4117	0.7872	0.7115	0.7474
TF-IDF+ Domain	0.7441	0.6153	0.6736	0.7959	0.75	0.7722
TF-IDF + Neg	0.7777	0.6730	0.7216	0.78	0.75	0.7647
All	0.7608	0.6730	0.7142	0.7959	0.75	0.7722

4. Discussion

In this paper, we presented a platform that allows a real world government agency to detect and analyze food safety related complaints posted by citizens on internet forums. The output of this system serves as an additional source of information related to the state of food safety which is not captured through the traditional process. We believe that our system is a valuable tool to monitor such content and to use it for improving food safety practices and standards. We employ text mining techniques to identify user posts related to food safety violations. We also demonstrate that by using creative ways of generating training data and feature engineering we are able to obtain a satisfactory performance. Currently, our system focuses on internet forums as the source of such information. One challenge in applying the system to micro-blogs was the length of each post. The short posts often do not contain enough information about the location of violation and hence the government agency cannot take any action on such reports. We are working on other ways of using information from micro-blogs.

References

[1] Estimates of foodborne illness in the United States, US centers for disease control and prevention. 2011. Last accessed on 30 Jan 2014, Available at: http://www.cdc.gov/foodborneburden/index.html

[2] Diakopoulos NA, Shamma DA. Characterizing debate performance via aggregated twitter sentiment. Proceedings on Human factors in computing systems (CHI’ 10). 2010 April.

[3] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: real-time event detection by social sensors.WWW '10 Proceedings of the 19th international conference on World Wide Web. 2010 April.

[4] Chenthamarakshan V, Melville P, Sindhwani V, Lawrence R. Concept labeling: Building text classifiers with minimal supervision. IJCAI'11 Proceedings of the Twenty-Second international joint conference on Artificial Intelligence. 2011.

[5] He H. and Garcia E.A. Learning from imbalanced data. IEEE Transcations on Knowledge and Data Engineering, 21(9):1263–1284, Sept. 2009