Hidden warnings; using topic modeling on Yelp reviews to predict food safety violations

Michael Bostwick
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
mgb2188@live.unc.edu

Abstract

This paper offers an approach for improving the predictive power of topics models that can be applied to public policy decisions made from social media data. Prior work on Supervised Latent Dirichlet Allocation (sLDA) has been found to surpass unsupervised topic models in predictive tasks, but may still not find topics along the dimensions that a user desires. This work examines techniques to pre-train the vocabulary used in the topic model and then tests it on a dataset including Yelp text reviews and restaurant health inspection scores. Findings suggest...

1 Introduction

Local governments have limited resources to carry out monitoring and inspection of restaurants, yet the CDC estimates that 48 million people get sick from foodborne illnesses in the United States each year cdc. At the same time, customers are constantly providing restaurant feedback that could inform inspection operations via reviews on Yelp. In recent years there has been increased interest for local governments to not only release more of their own open data reference, but to repurpose other data to improve efficiency and effectiveness. A prime source of data relating to restaurants can be found on Yelp, a crowdsourcing platform for all types of businesses, but most popular for food and beverage reveiws reference. Here customers can leave numeric ratings (on a scale of 1-5) and text reviews of a particular establishment they have visited. In larger metroplitan areas it is common for a restaurant to have hundreds or thousands of reviews, providing lots of potential clues to hygiene of a restaurant.

While there is a large amount of metadata associated with yelp reviews (time of review, average rating of reviewer, etc.), this work will focus on just the text review itself. In order to peform regression or classification from text, it is common to transform the text into a bag-of-words, that is each unique word is a feature and cell values represent the absence or presence of the word or the frequency count of the word in the observation. From there standard machine learning algorithms can be applied to the dataset, but since there are many unique words in the english language the dimensionality it usually very high. This dimensionality challenge, both for predictive power and interpretability will be examined in further detail throughout the paper.

This paper is organized as follows. I first review related work pertaining to the application of social media to civic functions and to the techniques used in topic modeling. Next, I elaborate on the model and how inference and learning can be performed on it. From there I evaluate the procedure on synthetic data and a combined Yelp and health inspection dataset from Boston, MA, comparing several approaches. Lastly, I discuss tradeoffs to consider that make this approach more or less favorable.

2 Related Work

2.1 Applications

This paper builds off of similar work by Kang et al. Kang and Joaristi et al. Joaristi, who predicted health inspection scores from Yelp data for Seattle, and Seattle and Las Vegas, respectively. The work done by Kang et al. Kang included many aspects of the review and restaurant data, including zip code and inspection history, and used Support Vector Machine on unigram+bigram text features. They formulated their prediction task as a classification problem and in order to keep balanced classes only considered restaurants with zero violations, "hygienic", or many violations, "unhygienic". Joaristi et al. Joaristi differ in their approach, as they apply Latent Dirichlet Allocation (LDA) to extract topics separately from the positive and negative text reviews. They then use these topic features along with other generated features as inputs into SVM, logistic regression and Random Forest algorithms.

2.2 Topic Modeling

Topic modeling is well established blei, griffiths technique to reduce the dimensionality of a set of text documents, typically in an unsupervised fashion. sLDA modifies the common LDA approach when the goal is prediction,

More generally, an oft cited work Yang and Pedersen on feature selection for text classification examines document frequency, information gain, mutual information, chi-squared statistic and term strength.

3 Model

3.1 Model

Lasso

Random Forest

3.2 Inference and Learning

4 Evaluation

4.1 Synthetic data

4.2 Yelp and Health Inspection data from Boston

5 Discussion

6 Conclusion

References

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.

[2] https://www.cdc.gov/foodborneburden/