Explaining Restaurant Ratings by Topic Sentiment Analysis
Juanita Ordonez, Diego Aguirre, and Olac Fuentes
Department of Computer Science, University of Texas at El Paso
jordonez6@miners.utep.edu daquirre6@miners.utep.edu ofuentes@utep.edu

Abstract
Online reviews of restaurants contain insights on topics that ratings alone cannot express. These topics can give owners clues on specific issues of the restaurant. Extracting such insights from online text presents many challenges due to noise in the data and the presence of irrelevant information. For instance, online reviews usually have misspellings and informal language. Potential customers can review the text and find certain topics that are positive or negative, which can ultimately influence their decision. Such topics can alter the restaurant's overall ratings, which can help affect the restaurant's revenue. In this study we explore extracting topics and sentiments from restaurants’ online reviews in the Yelp’s Challenge Dataset and find the correlation between topics sentiment to restaurants’ ratings. We apply basic pre-processing techniques and extract sentences from each review. Sentiment analysis was applied on each sentence with Stanford’s Recursive Neural Tensor Network (RNTN) model. Then, online learning LDA from Vowpal Wabbit was used on each sentence for topic modeling, with the assumption that each sentence contains at least one topic. Finally, we found there was a correlation between the sentiment towards topics and restaurant’s ratings. In our preliminary results we found that topic’s sentiment had a positive correlation to star ratings.

Introduction
Research has shown that social media has an impact on restaurant revenues, with up to a 9% increase when there is a single star rating improvement [5]. In this study, we explored topic sentiment analysis and its relationship to star ratings.

Data
We extracted restaurants’ text reviews and star ratings from the Yelp Challenge Dataset [1]. We chose the 525 restaurants with the largest number of reviews across the whole dataset.
• Total reviews of ~180,000
• Total sentences of ~1.7M

Pre-processing
Leaps and bounds better than other Chinese food in the Madison area. Their lunch buffet is a GREAT!!! (4) value...

1) Leaps and bounds better than other Chinese food in the Madison area.
2) Their lunch buffet is a great value.

Figure 1: Flow chart of two step pre-processing with before and after examples. In step one all extra punctuation, numbers, and nonsensical words were removed, misspellings were corrected, all characters were converted into lower case, and reviews separated into sentences. In step two terms were represented as a bag of words, stopwords (e.g. “the” “a” ...) and terms with frequency less than 300 were removed, and words stemmed using WordNet.

Sentiment Analysis
We obtained the sentiment of each extracted sentence using Recursive Neural Tensor Network (RNTN) trained with the Sentiment Treebank dataset on each extracted sentence.

Figure 2: RNTN’s sentiment analysis classification on each sentence taken from the dataset. Output score is between -2 and 2, where -2 is most negative and 2 most positive.

Figure 3: Representative words of each topic selected. Each topic is associated to a different restaurant trait: topic 0 to atmosphere, 1 to ingredients, 2 to steaks, 3 to burgers, 4 to price, 5 to desserts, 6 to staff and appearance, 7 to atmosphere and appearance , 8 to service and experience, and 9 to sauce and seasoning.

Topic Modeling
We used the Vowpal Wabbit’s online learning Latent Dirichlet Allocation (LDA) [2] to extract the topics from the database. Ten out of the fifty topics we found to be the most interesting were selected.

Figure 4: Flow chart describing the overall process for a single restaurant.

Figure 5: Restaurant’s correlation coefficient of star ratings and topics sentiment. It shows a slight positive correlation, with topic 6 having the highest correlation.

Results
• Continue exploring the sentiment behavior of topics in other datasets, such as twitter and news feeds.
• Explore how the behavior changes based on location and time of year.
• Improve our topic representation by extending the associated word list using WordNet.
• Develop method for automated topic labeling.

Future Work

References

Acknowledgement
Presentation of this work has been supported in part by NSF Grant CNS-1042341.