The Yelp Review Scorer

Max Grossenbacher
Galvanize Data Science Immersive

Contact Information
Email: grossenbacher.max@gmail.com
Linkedin: /in/maxgrossenbacher
Github: maxgrossenbacher

Motivation

Why is Natural Language Processing Important?
Approximately "80% of business-relevant information originates in unstructured form, primarily text" (breakthroughanalysis.com). Obviously, if some company wants to utilize all this information, then they must be able to take this unstructured free text and turn it into something meaningful and actionable. Natural language processing (NLP) does exactly this!

Social media is a burgeoning field built on the premise of human-to-human interaction (mainly through free text) on the internet. In this field, the ability to wrangle unstructured data can provide key insights about specific users or businesses. These insights can be used to optimize marketing campaigns to target specific users' interests, build recommender systems, or improve overall user experience.

Data

Yelp's Challenge Dataset provides access to millions of user reviews. Latent topic modeling was used to discover key insights into what topics reviewers discuss about certain businesses. Additionally, machine learning techniques were used to classify rating, sentiment and usefulness of a review.

Target engineering:
Usefulness: total useful votes received by a review

Not Useful	Useful	Very Useful
useful = 0	0 < useful < 5	useful >= 5

Sentiment: group reviews by rating to determine overall sentiment. Ratings range from 1-5.

Negative	Neutral	Positive
rating < 3	rating = 3	rating > 3

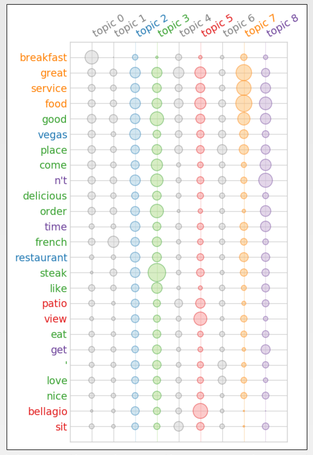
Methods

Yelp reviews were processed using the library Textacy. Textacy allows for multiprocessing of documents using SpaCy. During text processing, stop words are removed, words are tokenized and lemmatized, and a vocabulary of terms is generated.

Latent Topic Analysis

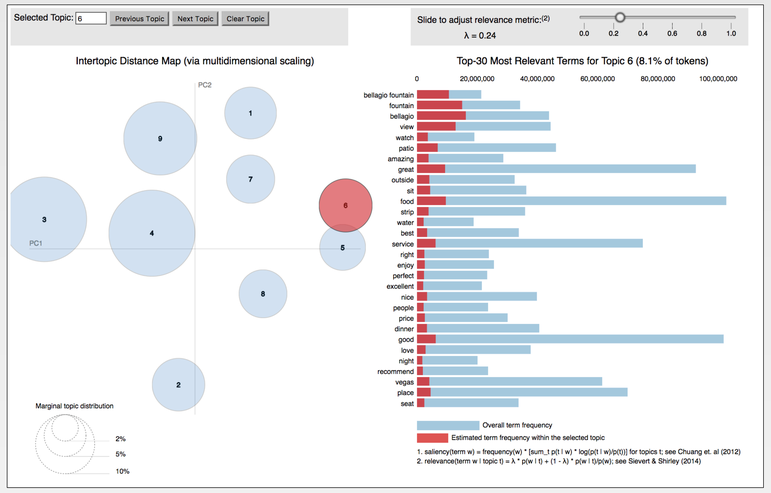
Vector representations of each review are created according to their term-frequency (TF). Latent Dirichlet Allocation (LDA) finds related terms among reviews, and clusters reviews into groups or topics based on their distribution of terms.

Termite Plot of Latent Topics for a Restaurant on Yelp



The bigger the circle, the more important the term is to the topic. The colored topics illustrate the 5 most important topics.

PyLDAvis Interactive Visualization of Latent Topics for a Restaurant on Yelp



Screen shot of pyLDAvis interactive visualization of latent topics. The terms to the right of the circle relate to the highlighted topic 6. **This corresponds to topic 5 in the termite plot to the left.**

The figures above demonstrate an example of latent topics found in reviews for an individual restaurant on Yelp. Topic 6, in the pyLDAvis plot, infers that this restaurant has a view of Bellagio Fountain in Las Vegas. Additionally, LDA can provide insight into which topics are most important to reviewers. For this particular restaurant, topic 4 (the topic containing the terms steak, french and delicious) is the most important topic. Furthermore, using the pyLDAvis tool, we can visualize the relationship between different topics. Circles that overlap i.e. topic 5 and 6 contain similar or overlapping terms. Whereas topics spread further apart i.e. topic 1 and 2 contain terms that are not so similar to each other.

Classification

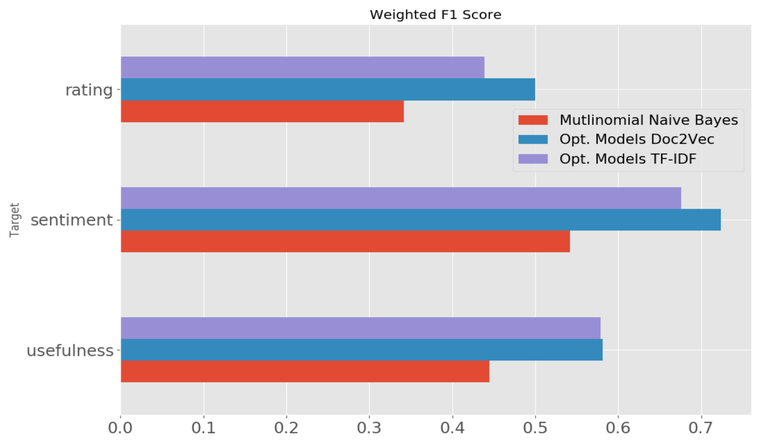
Machine learning techniques were used to predict the usefulness, sentiment and rating of a review. All models were compared to a baseline model (Multinomial Naïve Bayes trained on TF-IDF vectors) to demonstrate effectiveness. Grid Search CV was used to compare different model types according to their mean test weighted F1 score. We see an a 28%, 25%, and 30% increase in mean test weighted F1 score for predicting rating, sentiment and usefulness of a review when using the optimized models and parameters in the table below.

Training models on doc2vec representations of reviews increases the mean test weighted F1 score of each model (rating, sentiment, and usefulness) by 46%, 34%, and 31% respectively, as compared to the baseline model.

Optimized Models and Parameters

Target	Model	Parameters
Usefulness	Random Forest	Max Features = sqrt(# features) # Estimators = 1000
Sentiment	Gradient Boosted Trees	Learning Rate = 0.1 Max Features = sqrt # Estimators = 500
Rating	Gradient Boosted Trees	Learning Rate = 0.1 Max Features = sqrt(# features) # Estimators = 500

Models and parameters with the highest mean test weighted F1 scores after grid search for each target



Mean test weighted F1 scores of three models. All models were trained on 10,000 randomly chosen reviews with 4-Fold cross validation.

Conclusion

Final Models

Target	Model	Parameters	Accuracy	F1 Score
Usefulness	Random Forest	Max Features = sqrt # Estimators = 1000	62.5%	0.625
Sentiment	Gradient Boosted Trees	Learning Rate = 0.1 Max Features = sqrt # Estimators = 500	68.0%	0.680
Rating	Gradient Boosted Trees	Learning Rate = 0.1 Max Features = sqrt # Estimators = 500	50.7%	0.504

*Final models were trained on 300,000 doc2vec vectors of reviews

Final models were trained on 300,000 reviews. These models were incorporated into The Yelp Review Scorer. The Yelp Review Scorer scores reviews for probability of usefulness, sentiment, and suggested rating. More useful reviews will improve the user experience by providing users with more helpful and relevant reviews.

Future Directions:
Trying different classification techniques, for example a multinomial neural network may improve F1 scores.
Additionally, targets can be modified to better capture the essence of what we are trying to predict. Usefulness was based solely on the number of useful votes a review received at the time Yelp complied the challenge dataset. It would be helpful to collect data on the date a review was posted or the number of views a review received. By applying a time or number of views penalty, one may be able to improve upon these usefulness scores.
Lastly, expanding sentiment analysis to include feelings such as anger, joy, sadness, disgust and fear may allow further insights into which sentiment(s) are most common among reviews of different ratings.

Tech Stack

python pandas matplotlib NumPy TensorFlow Flask spaCy jupyter scikit learn