Using Yelp to Predict Restaurant Closings

Harsh Jha, John Maloney and Dominic Smith*

December 12, 2014

1 Introduction

It is commonly reported that between 60 and 90 percent of businesses fail within three years of opening.[1] This paper investigates whether we can predict restaurant closings using reviews on Yelp, a website that tracks reviews and other descriptive information for businesses, in addition to local economic and demographic variables. Predicting business closings is an important topic for two distinct reasons. From a business owner's perspective, the earlier an owner knows the business is going to fail the quicker she can cut her losses instead of investing more time and money into a business that is almost surely going to fail. From the government's perspective, spending at restaurants is highly correlated with GDP which is one of the first things to decline as the economy enters a recession.Current measurement methods for GDP rely on surveys that take time to collect and introduce a lag between when the decline occurs and when the measurement showing that decline is available. Data from Yelp is accessible in real time and could be used to construct a real time measure that serves as a leading economic indicator of GDP decline.[2]

This paper tries to account for two major reasons businesses close. First, a business may close because they provide poor service and/or bad food. Second, a business may close because poor economic conditions have caused the people who would normally frequent the business to decrease the frequency with which they eat out. The Yelp data provides a unique opportunity to measure business quality and the attention a business is getting. It also provides information on the competition each business faces.

We evaluate our performance along two fronts. The first is whether we can predict whether an individual business will close. The second is whether we can predict the number of businesses that will close in some area. The first task is much more difficult and stretches the limits of our data. In the end, we are unable to find a classifier that performs well on the first task but we are able to do reasonably well at the second task.

This paper builds on work by Hood et al. [2013] who acknowledge that it would be interesting to use Yelp data to understand whether a business is successful, but instead try to predict the number of reviews a business will receive in the next 6 months. We use many of the same data preprocessing steps as them. For example, both of us use sentiment analysis on review text to extract a richer set of features from the data. One step that Hood et al. take that we did not because of time constraints is to segment users based on review status. The idea is that reviews from different types of people should be weighted differently. For example, a rave review from a user with no other reviews may simply be the business owner trying to prop up the rating of her restaurant so it should receive less weight than a really negative review from a user with

*The code for this project is hosted at https://github.com/theharshest/machine_learning_5525. Figures are placed after references so they can be larger.

[1] See for example http://www.businessweek.com/stories/2007-04-16/the-restaurant-failure-mythbusinessweek-business-news-stock-market-and-financial-advice

[2] Doing this within Yelp's API Terms of Use would be difficult at best.

hundreds of reviews which range from positive to negative. We believe this would be a nice extension to this project, but do not use it at this time.

Real time economic forecasting/nowcasting has become a popular topic. Antenucci et al. [2014] found that they can create a real time measure of the unemployment rate in the U.S. using postings on Twitter. These predictions are particularly useful in accessing the impact of unforseen events such as Hurricane Sandy and the 2013 government shutdown. This paper aims to explore whether Yelp data could be used to establish a similar measure for sales at restaurants.

Having good real time data on restaurant sales is particularly interesting because restaurant sales are highly correlated with GDP. Figure 1 shows time series of sales at restaurants measured monthly by the U.S. Census Bureau and U.S. gross domestic product (GDP) measured at a quarterly level. These two series have been indexed to be 100 in January/First Quarter of 2006. The correlation coefficient between the two series is .97.

2 Data

Our data comes from three sources. The first source is the Yelp dataset challenge.[3] We have restricted this sample to 13,010 restaurants operating in either Phoenix, Las Vegas, or Madison between January 2006 and July 2014. These data contain descriptive attributes and user reviews for each restaurant.[4] The user reviews include textual comments and star ratings provided by Yelp users while the descriptive attributes include categorical and ordinal attributes such as type of food served, ambiance provided and price range. Our second source is the U.S. Decennial Census.[5] These data contain data on the income level, ethnic composition, and population of an area. The data is summary level information provided for Census Tracts which are geographical divisions of the U.S. that contain between 1,200 and 8,000 people.[6] The final data are from the Local Area Unemployment Statistics published by the Bureau of Labor Statistics (BLS). They contain monthly unemployment rates at the city level.[7] The Yelp user reviews allow us to assess restaurant quality through review stars and sentiment expressed in textual comments as well as measure business attention through the number of reviews provided for each restaurant. The Yelp descriptive attributes allow us to measure the degree of similarity between restaurants. Data from the Census allow us to control for whether a restaurant is in an area where it should do well. For example, high priced restaurants should do better in high income areas. Data from the Bureau of Labor Statistics (BLS) allow us to control for aggregate economic conditions which make all restaurants more or less likely to close.

2.1 Measuring Business Closings

One concern with this project is how accurately we can measure business closings. Yelp provides a way for users to flag a business as closed, but it is not immediately obvious how to mark a business as closed or whether it is even possible to do so. It took the authors a few misclicks before they figured out how to mark a business as closed.[8] Once a user marks a business as closed, a Yelp employee will physically review the information available for a business before it is officially marked as closed. This implies that businesses marked as closed forms a lower bound for the number of restaurants which are actually closed. In further work we would build a web crawler to search web archives to help determine whether a business is closed. Figure 2 shows the survival rate for one cohort of restaurants, those that entered in December of

[3] Downloaded from http://www.yelp.com/dataset_challenge

[4] See **Hood et al.** for additional description of the data.

[5] Downloaded from American Fact Finder at the Census Tract level.

[6] See https://www.census.gov/geo/reference/gtc/gtc_ct.html for more information.

[7] Downloaded from http://www.bls.gov/lau/#data

[8] The authors also do not know how long this feature has been available.

2008, measured two ways. In the first series we mark a business as closed only if the closed flag is checked. We assign a closed date as the date of the last review. In the second series we mark a business as closed on its last review date regardless of the state of the closed flag. Of the 153 restaurants that entered the sample in December 2008, 80% have their last review before January 1, 2013. The last review included in our data set was submittedon July 30, 2014 which means that these restaurants went at least 16 months without a single review. This would not be very indicative of closing if most restaurants only receive one review, but we find that most restaurants receive at least a few reviews. The distribution of reviews per business over our sample is presented in Figure 3.

Another concern is that Yelp has become much more popular over time. So it could be that businesses with a lot of reviews are simply those opened in 2013. Figure _ presents the number of reviews per business for the cohort that entered in 2008. Figures 4, 5, and 6 show the number of restaurants entering and exiting our sample over time as well as the number of reviews in the data per month. At this time we only treat a business as closed if the closed flag is actually checked. Future work would try to improve on our measure of business closings.

3 Methods

3.1 Evaluation Methods

In our baseline model we treat this as a 5-class classification problem with the following classes: “closed 0-3 months after”, “closed 4-6 months after”, “closed 7-9 months after”, “closed 10-12 months after”, “still open after 12 months”. We also considered this problem as a binary classification problem with the classes split on whether the restaurant was still open after 3, 6, 9, or 12 months. The classifiers we tried performed roughly the same across all of these problems so we only present results from the 5-class problem in this paper, but provide some results from the binary classifier in the appendix. We also considered framing this problem as a regression problem with “days until closing” as the dependent variable. This ends up not working well because many restaurants do not close in our sample. We tried assigning these restaurants a high value for “days until closing”, but the results varied a lot depending on what value we chose.

We evaluate each classifier using a walk forward or shifting window approach[9]. In this approach, we select a date d to serve as the date on which predictions will be made and use all data available before this date to create a training set. We train the classifier with the training set using grid search and 3-fold cross validation in order to obtain optimal values for both hyperparameters and model parameters. To test the classifier, we move the prediction date ahead by a specified amont t, use all data available before this new date to create a test set and then test the trained classifer using the test set. We repeat this process in an iterative process, increasing the date by t on each iteration, until the date exceeds the available data. For most runs we set the initial date to Jan 1, 2009 and set the step size t equal to 12 months.

The results we present are the sums of the results for each iteration of this algorithm. Initially we suffered from the problem that few of the restuarants that are open at time d are marked as closed by time $d+t$ because of the fact that our measure of businesses closings does not pick up all closings. Since the closed classes were very small compared to the open class, many classifiers would just predict that all businesses will stay open. We fixed this problem by sampling the “open” class at each date so that it is the same size as the largest of the “closed” classes. We present misclassification rates using this selected sample.

[9] https://riskcalc.moodysrms.com/us/research/crm/validation_tech_report_020305.pdf

3.2 Sentiment Analysis

Previous work such as Hood et al. has used sentiment analysis with reviews on Yelp data. We had hoped to include sentiment measured on a number of dimensions, but due to time and computer processing constraints we calculate sentiment from review text using a training dataset available from metashare.[10] After preprocessing the data to remove spelling mistakes, repetitions and contractions we do stop word removal. Then we extract features as unigrams, bigrams and trigrams. After that, we vectorize the corpus and calculate the sentiment of each review as an integer between (and including) -2 and 2 using the linear SVM classifier. Further details are in the appendix.

3.3 Other preprocessing steps

All features are standardized to have mean 0 and variance 1 because different features have wildly different ranges. We include a constant when using evaluation methods for which a constant is appropriate.

4 Results

We focus on SVM and Decision Trees as our classifiers in the text of this paper. Additional results are saved for the appendix.

4.1 SVM

The problems with SVM are indicative of the problems with many of the classifiers we tried across a wide range of feature combinations as well as attempts with using recursive feature elimination and principle component analysis for dimensionality reduction. SVM predicts too many businesses will stay open and does no better than random guessing on the rest of the observations.

Table 1: SVM - 5-class with all features

Actual\Predicted	0-3	4-6	7-9	10-12	12+	Total
0-3	24%	20%	22%	2%	32%	443
4-6	24%	16%	22%	3%	35%	449
7-9	25%	19%	22%	3%	31%	486
10-12	19%	16%	20%	3%	42%	402
12+	21%	11%	18%	2%	48%	480
Total	514	369	472	59	846	2260

4.2 Decision Tree

Decision trees do not do better than SVM when is comes to classifying whether an individual business will close. They do perform better when it comes to the number of businesses that close.

[10] http://metashare.elda.org/repository/browse/semeval-2014-absa-restaurant-reviews-trial-data/

Table 2: Decision Tree - 5-class with all features

Actual\Predicted	0-3	4-6	7-9	10-12	12+	Total
0-3	23%	17%	23%	17%	18%	443
4-6	20%	18%	21%	22%	19%	449
7-9	25%	17%	22%	16%	19%	486
10-12	22%	17%	20%	18%	22%	402
12+	18%	18%	20%	17%	27%	512
Total	497	403	491	415	488	2294

5 Conclusions

We are unable to find a classifier that is able to predict whether a particular restaurant will close in the next year, but we do see some promise that an economic indicator predicting the aggregate number of restaurant closings could be developed using this data. The main issue we face is difficulty measuring whether a restaurant is actually closed. The focus of future work would be to address this problem by using information available elsewhere on the web. Another possible solution would be to use data from the I.R.S. and the U.S. Census on which businesses file tax returns in a given year. This would certainly let us train and test our classifier in a setting where we could be much more confident that we were measuring business closings accurately. However, these data have a number of barriers that must be overcome to get access to them making it impossible to use Census data for this project.

References

Dolan Antenucci, Michael Cafarella, Margaret Levenstein, Christopher Re, and Matthew Shapiro. Using social media to measure labor market flows. 2014.

Bryan Hood, Victor Hwang, and Jennifer King. Inferring future business attention. 2013.

J. Huang, S. Rogers, and E. Joo. Improving restaurants by extracting subtopics from yelp reviews. In *Presented at IConference 2014 Berlin*, 2014.

J. McAuley and J. Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *RecSys '13: Proceedings of the 7th ACM Conference on Recommender Systems*, 2013.

S. Wang and C. Manning. Baselines and bigrams: simple, good sentiment and topic classification. In *ACL '12: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, 2012.

Figure 1: Comparison of Key Economic Indicators

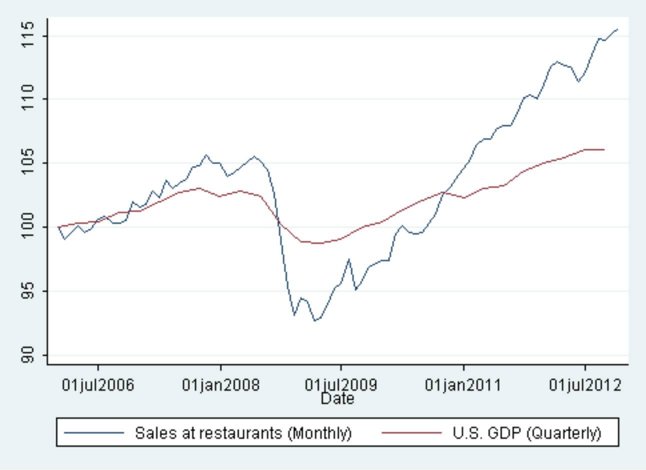


Figure 2: Survival Rate of December 2008 Cohort

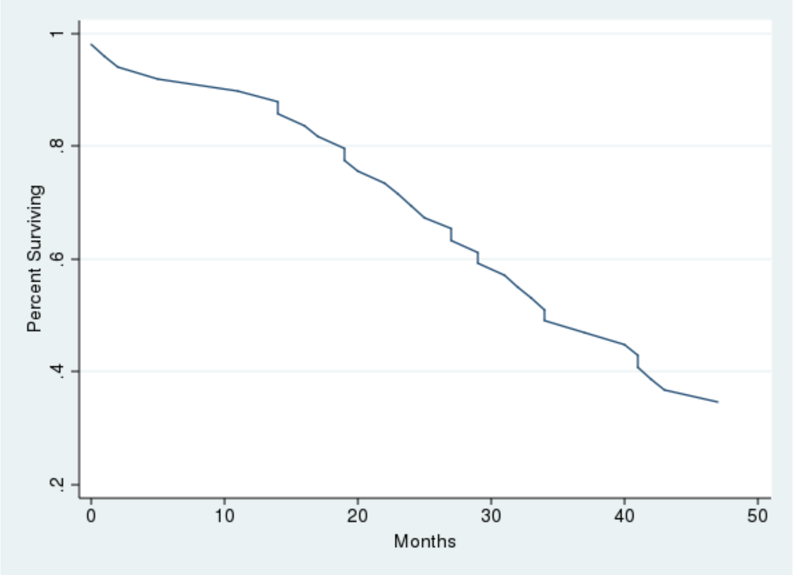


Figure 3: Distribution of Reviews per Restaurant

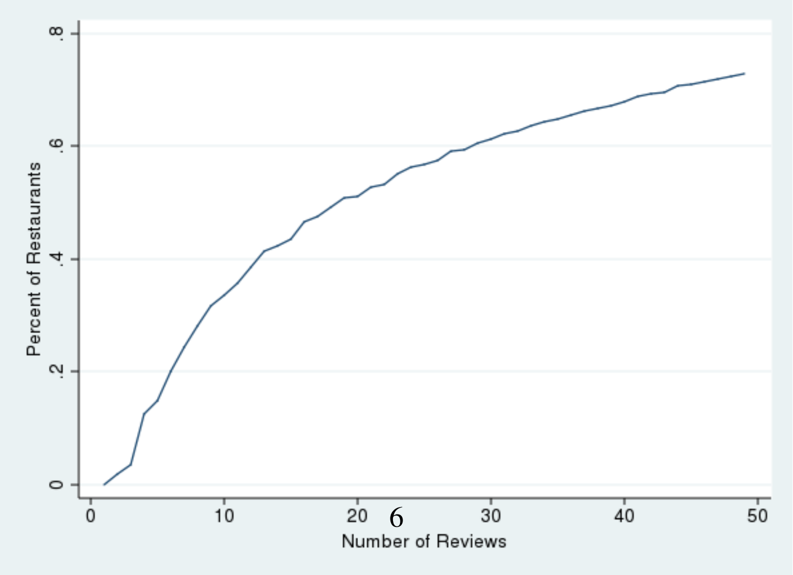


Figure 4: Number of Restaurants Entering the Sample

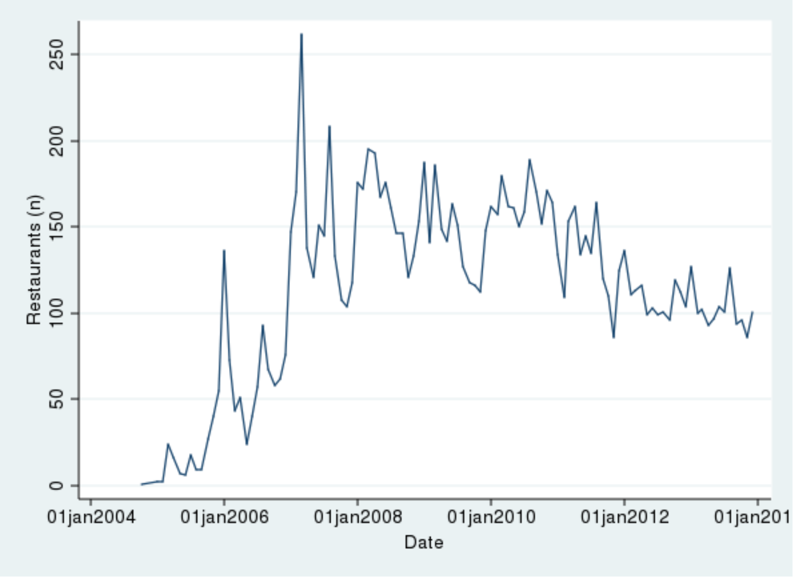


Figure 5: Number of Restaurants Exiting the Sample (Defined as not receiving another review by July 2014 i.e. increase is artificial)

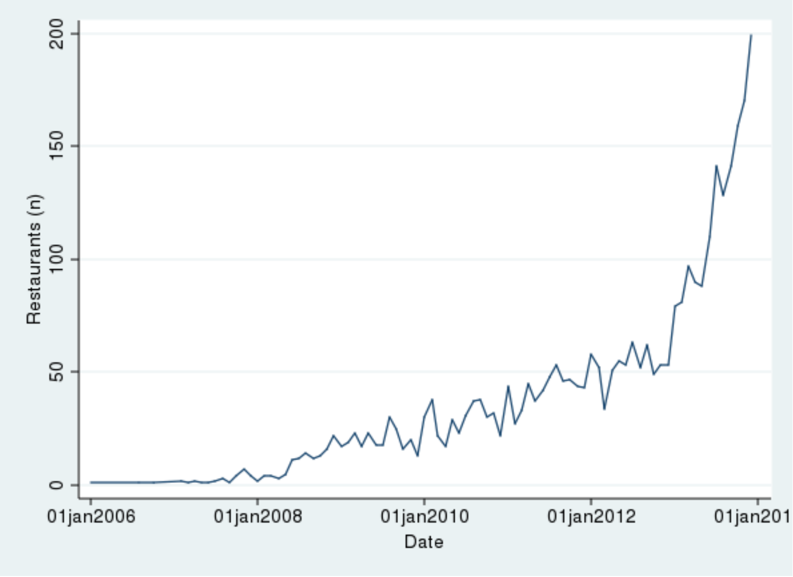
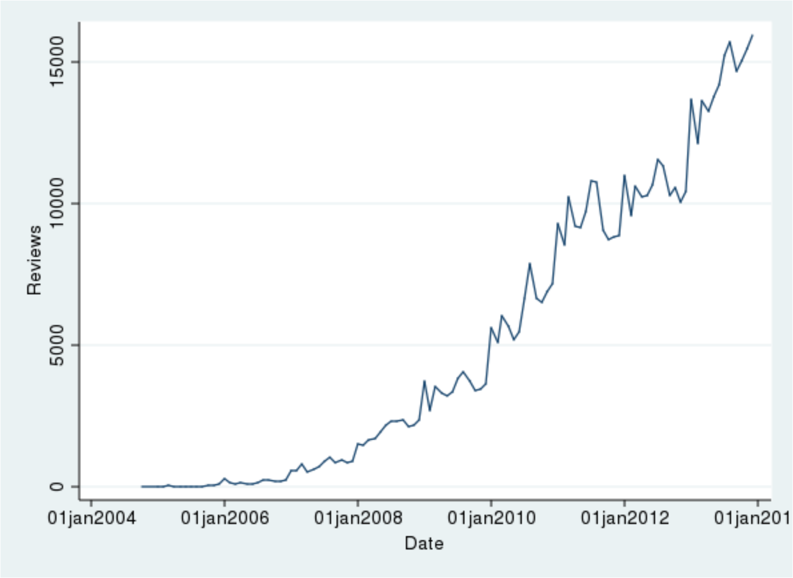


Figure 6: Number of Reviews per Month



A List of Features

- Mean Sentiment Rating
- Mean Star Rating
- Percent Change Star Rating between quarters
- Number of African Americans (Tract Level)
- Percent of population African American (Tract Level)
- Number of Restaurants in Census Tract
- Percent change in Number of Restaurants in Census Tract
- Days since business' first review
- Quintile of Income per Capita (Tract Level)
- Total Income of residents
- Income per Capita
- Number of people over the age of 65
- Percent of population of the age of 65
- Total Population
- Number of Reviews for Restaurant
- Percent Change in Number of Reviews for Restaurant

- Total Sentiment Rating
- Total Number of Stars
- Total Number of Stars Percent Change
- Total Number of Tips
- Total Number of Tips Percent Change
- Number of people under the age of 18 (Tract Level)
- Percent of population under the age of 18 (Tract Level)

B Additional Results

B.1 Random Forest

Random forests actually perform worse than decision trees when it comes to predicting the number of businesses that will close. Additionally, we find that random forests have a higher variance over the time windows in our sample than most predictors.

Table 3: Random Forest

Actual\Predicted	0	1	2	3	4	Total
0	40%	14%	24%	5%	17%	443
1	39%	13%	23%	5%	20%	449
2	44%	14%	19%	4%	20%	486
3	37%	14%	20%	5%	25%	402
4	35%	9%	16%	5%	36%	483
Total	879	283	459	108	534	2263

B.2 Adaboost

Adaboost performs well at hitting the number of restaurants that will close, but are basically guessing on which restaurants will close. It is curious that many of the classifiers have difficulty differentiating the businesses that will close in 9-12 months from those that won't close. We were tempted to say this means we should be using "Open at 9 months" as the last class, but if you look into where classes 0, 1, and 2 are misclassified to they end up getting send to 4 fairly often.

Table 4: Random Forest

Actual\Predicted	0	1	2	3	4	Total
0	20%	21%	29%	10%	19%	443
1	19%	22%	27%	11%	20%	449
2	24%	17%	27%	12%	19%	486
3	18%	23%	28%	9%	21%	402
4	12%	23%	26%	9%	30%	511
Total	425	488	626	242	510	2291

B.3 KNN

KNN performs similarly to random forest. The results are not much different from guessing and don't hit aggregate numbers well.

Table 5: KNN

Actual\Predicted	0	1	2	3	4	Total
0	35%	22%	21%	11%	11%	443
1	33%	24%	20%	10%	14%	449
2	33%	24%	16%	12%	15%	486
3	32%	21%	19%	13%	14%	402
4	30%	26%	15%	10%	20%	486
Total	733	535	408	252	338	2266

B.4 Limited Features - SVM

We tried SVM with only number of days since the restaurant's first review, average star rating (last quarter, second to last quarter), total number of reviews before *d*, review count last 3 months and found the results were basically the same as will 44 feature SVM.

Table 6: Limited Feature - SVM

Actual\Predicted	0	1	2	3	4	Total
0	24%	20%	22%	2%	32	443
1	24%	16%	22%	3%	35%	449
2	25%	19%	22%	3%	31%	486
3	19%	16%	20%	3%	42%	402
4	21%	11%	18%	2%	48%	480
Total	514	369	472	59	846	2260

B.5 PCA (5 features) - Decision tree

PCA yields similar results to the other methods indicating that despite the fact that we include 44 features in our most detailed model they are actually only capturing a few different things.

Table 7: PCA - Decision Tree

Actual\Predicted	0	1	2	3	4	Total
0	20%	30%	20%	2%	27%	443
1	18%	31%	19%	3%	28%	449
2	21%	26%	19%	2%	31%	486
3	19%	31%	18%	2%	31%	402
4	18%	25%	18%	3%	35%	475
Total	437	646	426	57	689	2255

B.6 Binary - Decision Tree

Binary decision tree does a fairly good job predicting which businesses will close, but it also predicts too many businesses will close.

Table 8: Binary - Decision Tree

Actual\Predicted	0	1	Total
0	71%	29%	1780
1	53%	47%	1828
Total	2237	1371	3608

C Sentiment Analysis

Sentiment analysis is a simple supervised task in this case. We do not have any training data, so we are considering annotated restaurant reviews data taken from metashare (Restaurants_Train.xml). From this dataset we have 1315 reviews with sentiment score equal to -2, -1, 0, +1 or +2.

Then, we preprocessed the training data (above data) and test data (reviews and tips “text” fields) as per the text preprocessing functionality that we have in our code, which does the following tasks:

- Remove repetitions: cooool -> cool
- Spelling correction within 1 edit distance: majic -> magic
- Expansion: don’t -> do not, etc (check replacers.py for exhaustive list of expansion operations)

After getting preprocessed train and test data, we vectorize the whole corpus using HashingVectorizer, removing stop words and considering unigrams, bigrams and trigrams as features:

HashingVectorizer(tokenizer=word_tokenize, stop_words=’english’, ngram_range=(1, 3))

Then, we divided the vectorized corpus into train and test based on the above original train/test datasets.

For the train vectorized corpus we select K best features using F-measure and train a linear SVM classifier with optimal set of hyperparameters found using grid search with F-measure as scoring function and 5 fold cross validation:

Grid Search: params = { ’f_score__k’: [800, 1200, 1600, 2000, ’all’], ’svm__C’: [0.01, 0.5, 1, 10], ’svm__tol’: [1e-2, 1e-3, 1e-4], ’svm__dual’: [True, False] } gs = GridSearchCV(clf_SVM, params, cv=5, scoring=’f1’) gs.fit(X_tfidf, y) print gs.best_score_ print gs.best_estimator_.get_params()

Feature selection and classifier instantiation:

clf_SVM = Pipeline([(’f_score’, SelectKBest(f_classif, k=’10000’)), (’svm’, LinearSVC(C=0.5, tol=1e-3, dual=False))])

Finally, we applied this classifier on the vectorized test corpus to get the predicted scores.[11]

[11] We reviewed McAuley and Leskovec [2013], Huang et al. [2014], and Wang and Manning [2012] to understand how sentiment analysis works. Unfortunately we did not have time to incorporate many of their suggestions.