NEW YORK UNIVERSITY

Project Report- Intro to Data Science

Yelp Business Category Prediction

December 21, 2015

Classy-Flyers:

Yining Gao

Sean D'Rosario

Christina Bogdan

Raju Samantapudi

Instructor:

Prof. Brian D'Alessandro

Table of Contents

1 Introduction

Yelp.com publishes crowd-sourced reviews of local businesses. A business on Yelp can be classified into different categories and we aim to predict the category/categories that a business can belong to in a timely and accurate manner.

Each business on Yelp can be listed under one or more categories, depending on the type of service. For example, “Coffee & tea” is the category for Starbucks. A business can also belong to multiple categories (for example, The Cinnamon Snail, as shown in the image below). Businesses that are categorized have a wider reach and higher ranking in search results.

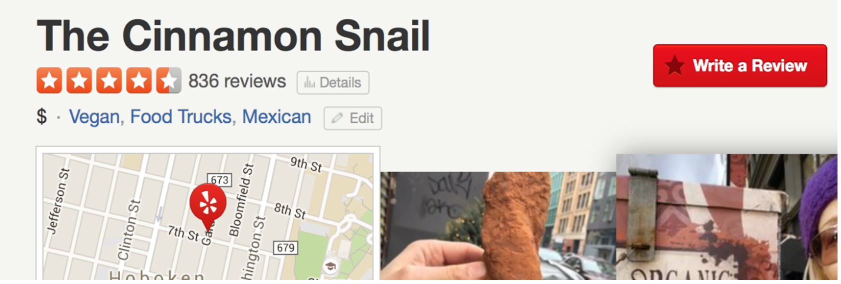


Figure 1.1 Snapshot of a business’s page on Yelp

Many small businesses on Yelp are still uncategorized, and are yet to reap the full benefits of being listed on Yelp. Manually categorizing them is a time-inefficient and labor-intensive task.

The aim of this project is to use available knowledge of the business to automatically assign it a category. As there are multiple categories that are associated with one business, this classification task is multi-label in nature. One of the main challenges with multi-label classification is the ambiguity and high degree of overlap between categories.

2 Data

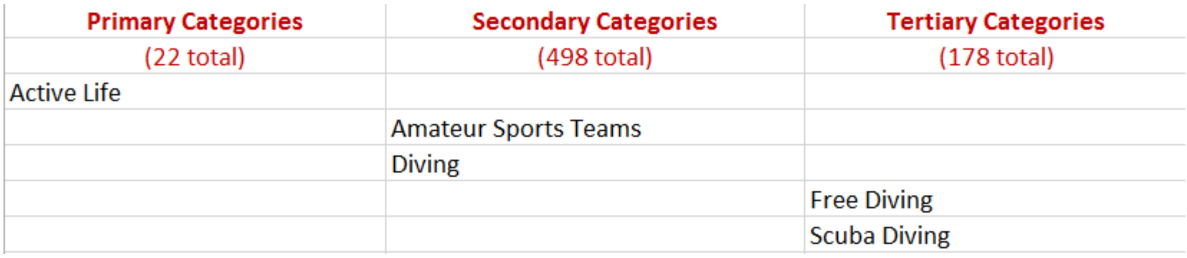
2.1 Data Understanding

We acquired the data set through Yelp’s Dataset Challenge in the form of five datasets in JSON format, each containing different dimensions of Yelp’s data (business, user, review, tip, and check-in). Of these five, we found that the data from the business and review documents were most relevant to our task.

In total, our data is comprised of 1.6M reviews by 366K users for 61K businesses. We also had the option to utilize Yelp’s API to access more data, but decided to limit ourselves to the smaller dataset because of computing constraints on our systems.

Going into our project we knew that the ‘review’ field (found in the review dataset) would be the basis for our models, and that we would be doing a lot of feature extraction from that one field. Reviews have many key words in them that are indicative of the business category they correspond to. This is confirmed by the word clouds that we created (seen in Appendix II). The business data file also contains other features that could be of interest to our model, including business hours of operation, business attributes, and business name. The full list of fields found in our dataset is listed in Appendix I.

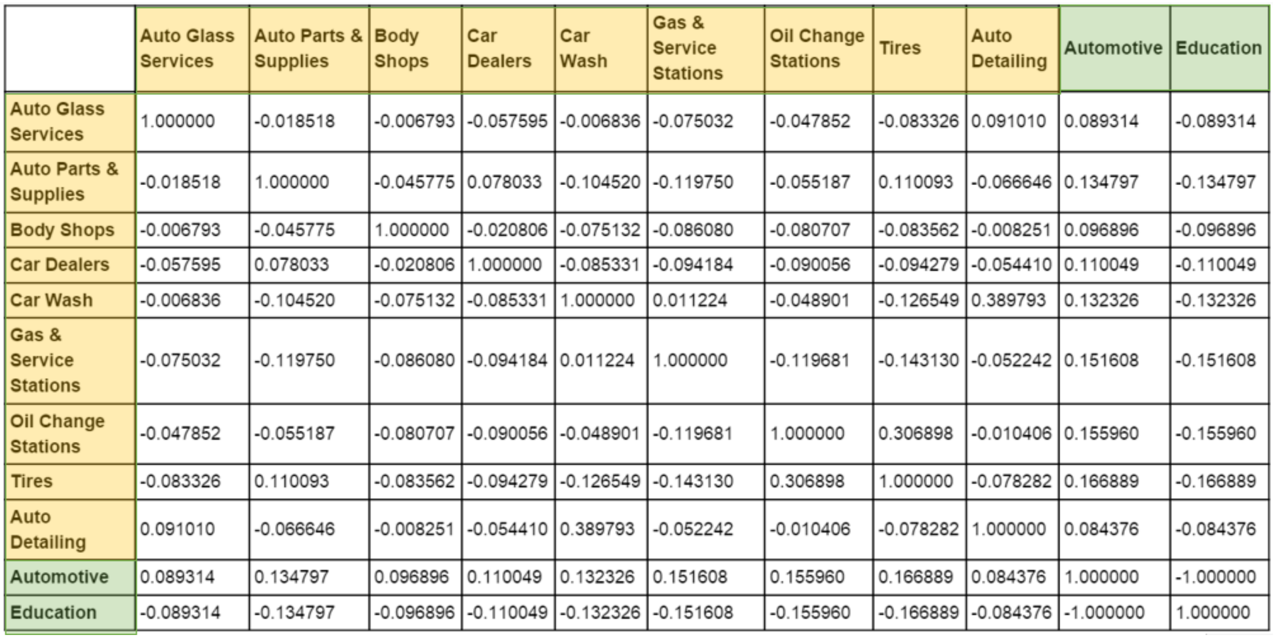
Exploring the distribution of businesses across different categories was an important step in understanding our data. Yelp allows businesses to be placed into more than one category, with some categories being general and others specific. The distribution of categories can be understood as a hierarchy of primary, secondary, and tertiary levels [3]. Categories in the primary division occur the most frequently, with the number of businesses per category decreasing as you move down the hierarchy. An example is shown in figure 2.1:



Primary Categories	Secondary Categories	Tertiary Categories
(22 total)	(498 total)	(178 total)
Active Life		
	Amateur Sports Teams	
	Diving	
		Free Diving
		Scuba Diving

Figure 2.1. Category hierarchy

The category hierarchy is distributed such that secondary/tertiary categories (sub-categories) are less likely to appear together in a business's categorization, and are obviously more likely to appear together with their corresponding primary category. An example of this can be seen in the correlation matrix below, which represents the co-occurrence of categories. Here, the two primary categories are 'Automotive' and 'Education', and the rest are sub-categories of the 'Automotive' primary category.



	Auto Glass Services	Auto Parts & Supplies	Body Shops	Car Dealers	Car Wash	Gas & Service Stations	Oil Change Stations	Tires	Auto Detailing	Automotive	Education
Auto Glass Services	1.000000	-0.018518	-0.006793	-0.057595	-0.006836	-0.075032	-0.047852	-0.083326	0.091010	0.089314	-0.089314
Auto Parts & Supplies	-0.018518	1.000000	-0.045775	0.078033	-0.104520	-0.119750	-0.055187	0.110093	-0.066646	0.134797	-0.134797
Body Shops	-0.006793	-0.045775	1.000000	-0.020806	-0.075132	-0.086080	-0.080707	-0.083562	-0.008251	0.096896	-0.096896
Car Dealers	-0.057595	0.078033	-0.020806	1.000000	-0.085331	-0.094184	-0.090056	-0.094279	-0.054410	0.110049	-0.110049
Car Wash	-0.006836	-0.104520	-0.075132	-0.085331	1.000000	0.011224	-0.048901	-0.126549	0.389793	0.132326	-0.132326
Gas & Service Stations	-0.075032	-0.119750	-0.086080	-0.094184	0.011224	1.000000	-0.119681	-0.143130	-0.052242	0.151608	-0.151608
Oil Change Stations	-0.047852	-0.055187	-0.080707	-0.090056	-0.048901	-0.119681	1.000000	0.306898	-0.010406	0.155960	-0.155960
Tires	-0.083326	0.110093	-0.083562	-0.094279	-0.126549	-0.143130	0.306898	1.000000	-0.078282	0.166889	-0.166889
Auto Detailing	0.091010	-0.066646	-0.008251	-0.054410	0.389793	-0.052242	-0.010406	-0.078282	1.000000	0.084376	-0.084376
Automotive	0.089314	0.134797	0.096896	0.110049	0.132326	0.151608	0.155960	0.166889	0.084376	1.000000	-1.000000
Education	-0.089314	-0.134797	-0.096896	-0.110049	-0.132326	-0.151608	-0.155960	-0.166889	-0.084376	-1.000000	1.000000

Figure 2.2. Correlation matrix of category co-occurrence

This unique structure of classes increased the complexity of our problem, and ultimately we chose to focus on effectively categorizing businesses based only on primary categories. Our decision was based on the following factors:

- Primary categories have significantly more data attached to them--for example, there are 21892 businesses in the primary category ‘Restaurants’, and only 51 businesses in the tertiary category ‘Cheese Shops.’ The limited data in secondary and tertiary categories makes it difficult to build a classifier.
- Primary categories are largely disjoint, while primary and secondary/tertiary categories occur together by nature, and have a high degree of overlap. These two cases should ideally be treated differently when applying a model.

2.2 Data Preparation

We used five fields from Yelp’s dataset in our final model: business ID, review text, business hours of operation, business name, and business category. We decided to simplify our problem by filtering the data on ten primary categories: Active Life, Home Services, Automotive, Pets, Education, Professional Services, Financial Services, Public Services & Government, Health & Medical, and Restaurants. We prepared our final data by filtering the business file on these ten categories, then joining it with our review data file.

3 Modeling

3.1 Baseline Model

Before diving into the textual review data, we built a baseline model based on business name and business hours of operation. These models are described below:

Business Name: A business on Yelp will always have its name recorded, even if it does not have any reviews. Furthermore, the name of a business can sometimes be telling of its category (e.g. ‘Tribeca Dental Design’). We apply text classification models on the business name to develop a baseline model. Some of the text classification features we used are k-NN, SVM, and Naive Bayes.

Business Operation Hours (BOH): We assume that the operating hours for each business will help us in determining the business category. An intuitive example is that a business classified as ‘Nightlife’ tends to operate later in the day in comparison to a business in the category ‘Education’. We wanted to use this information to predict the category information. The potential of using BOH as a feature is seen in the visualization below, which highlights the differences in distribution of BOH between two categories.

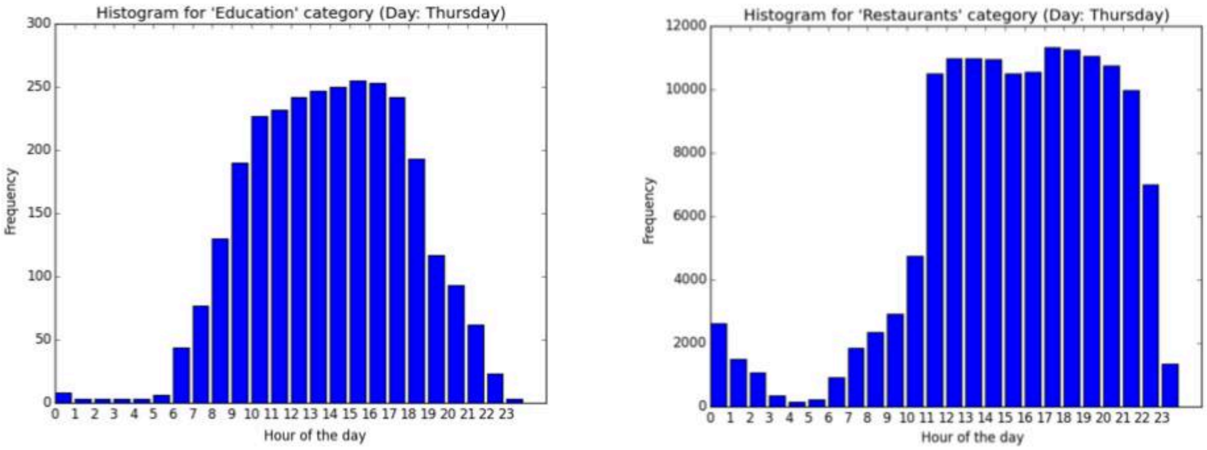


Figure 3.1. Distribution of frequency BOH for two categories

For the baseline, we used BOH to predict the category using the same three algorithms. However, using BOH as a feature is not ideal because there is a high degree of overlap in hours of operation between the categories.

3.2 Sampling Approach

For each primary category, we randomly sampled 10% of businesses from the original data set (which came to 3550 business units). The advantage of proportional sampling is that we maintain the same distribution of category in our sample, avoiding the pitfall of selection bias. A downside to using this method is that the Positive Response Rate (PRR) varies largely over all primary categories, with most of the categories having a PRR smaller than 5%.

One approach to deal with extreme small PRR is to use a down-sampling method, which is beyond the scope of our report. We will discuss the relationship of model performance with PRR later in the report.

3.3 Feature Engineering

We use both text and numerical features in our model:

Text features

Handling text features was a major part of our feature engineering process. We have business name and business reviews as text features. We processed all of our text features in the following steps:

(1) *Tokenization*: we transformed the documents into bags of words representation.

(2) *Stop words and punctuation filtering*: we removed words such as “a”, “the” and “and” that are common in all documents and thus less important in our modeling process.

(3) *TF-IDF transformation*: we used the term frequency - inverse document frequency (TF-IDF) method to transform the word count matrix. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the all documents.

(4) *Semantic modeling*: Since each review document can be expressed as a linear combination of topics, we thought these topics could help us better represent the underlying categories of the review documents. To enrich our feature set, we used LDA/NMF methods to extract topic coefficients and used them along with the previous TF-IDF features to train an updated model.

Numerical features

Business operating hours (BOH): Each business unit is made to a vector of 7(days)*24 (hour) length using 0 or 1 indicator representing whether it is operating at this time.

3.4 Classification Approach

We selected ten primary categories and then built ten classifiers for each of them in a one-vs.-all approach. Each classifier makes a simple binary prediction: whether a business belongs to this category or not. With a series of binary classifiers, we recovered the multi-label aspect of the problem by applying each binary classifier to a business. For example, if both the "Active Life" and "Pets" classifier predicted "yes", while all other classifiers predicted "no" for a business unit, our prediction would be that this unit belongs to category "Active Life" and "Pets", in the primary category level.

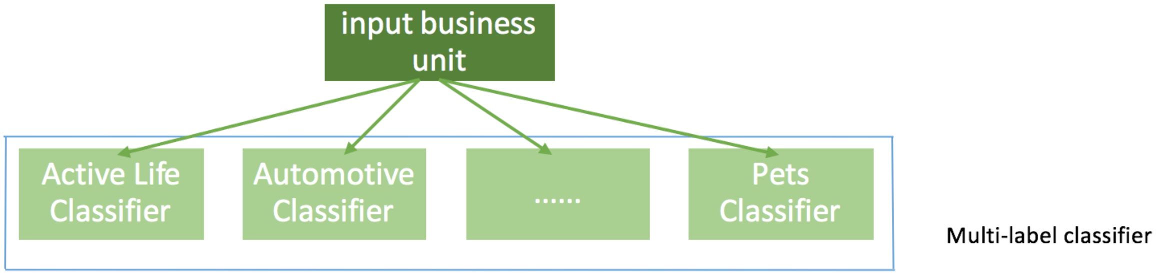


Figure 3.2 From Single Label Classifier to Multi-label Classifier

We applied Naive Bayes, k-NN, and linear SVM models to our data and then evaluated their performance.

4 Evaluation & insights

4.1 Binary Classifier Evaluation

We chose F1 score as our evaluation metric, which quantifies both precision and recall. We are interested in maximizing precision because we want to only recommend businesses to users that are correctly categorized. Recall is relevant because we want to categorize many businesses so that we can increase visibility of all businesses to users.

Below are the F1 scores for our three algorithms across different feature combinations:

	BOH-Baseline	*Business Name-Baseline*	*Review*	*Review+ BOH+ Name*	*Review+ LDA*	*Review+ NMF*
Naïve Bayes	0.22(0.24)	0.26(0.29)	0.09(0.29)	0.22(0.31)	0.10(0.30)	0.09(0.29)
k-NN	0.23(0.19)	0.59(0.26)	0.76(0.31)	0.53(0.28)	0.75(0.32)	0.75(0.30)
SVM	0.25(0.24)	0.67(0.26)	0.71(0.34)	0.73(0.31)	0.72(0.35)	0.72(0.31)

Table 4.1: Average F1 score (with Standard Deviation)

The above table presents the average F1 score of all 10 classifiers for three algorithms. For k-NN and SVM, review lift model performance greatly, while for Naive Bayes pure BOH or name does better prediction.

The success of the k-NN and SVM classifiers relative to the Naïve Bayes is mostly due to the fact that NB is a generative classifier, while k-NN and SVM are discriminative classifiers. Generative classifiers need large amounts of data to effectively estimate the probability distribution function of words given a document, while discriminative classifiers do not. Reviews are generally short so Naïve Bayes was not able to perform well.

The combination of various features (last three columns of table) doesn't always have better results. In our topic modeling approach, we initially thought that we would use ten topic coefficients because we were working with ten categories. Upon reflection, this was not ideal because we do not have control over the latent topic dimensions that were learned by the LDA and NMF models. A better approach would have been to explore building these models with different topic numbers and comparing the results, then choosing the number that yielded the best results.

To explore the performance of our models on secondary categories, we applied these same models, along with the F1 evaluation metric, to a set of ten sub-categories of the 'Automotive' primary category. Our classifier achieved similar results.

In summary, text features from reviews, along with a discriminative classifier are ideal for our category prediction problem.

4.2 Finding Optimal Hyper-Parameters

To find the best hyper-parameter (k) in our k-NN model, we performed a grid search for k in the range from 1 to 11 and tested each setting using cross-validation scheme. The optimal k is the majority vote of all 10 classifiers. Our validation set had the highest score when we set k=5.

4.3 Stepwise Feature Selection

In order to explore the optimal feature combination, we used a stepwise feature selection method.

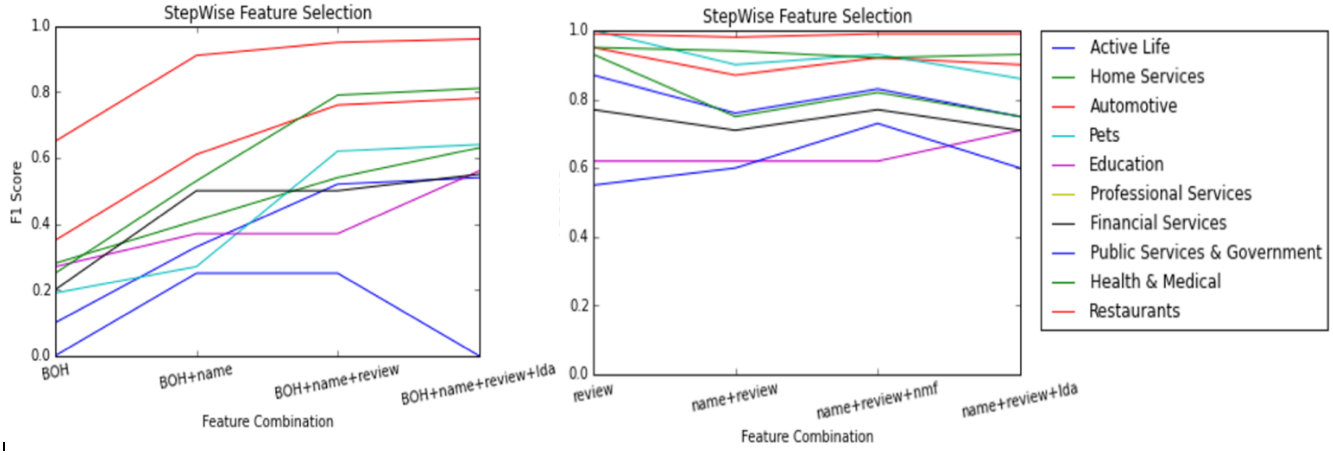


Figure 4.1 Stepwise Feature Selection

The figure on the left shows how much improvement adding text features (name and review) will bring to the result. Figure on the right shows that adding more text features (LDA, NMF) doesn't improve the result.

As a conclusion, model using review feature only gives the best results. We'll use this model to build the multi-label classifier.

4.4 Multi-Label Classifier Evaluation

We evaluate the multi-label classifier using the below formula.

$$Accuracy = \frac{\#Intersection(True, Predicted)}{\#Union(True, Predicted)}$$

For example, let's say that business' true categories are {A, B, C}, while our multi label classifier predicted {B, C, D, F}, the accuracy of this multi-label classifier is then the number of elements in the intersection of this two set: {B, C} divided by the number of elements in the union: {A, B, C, D, F}, i.e., 2/5 = 0.4.

For multi-label classifier evaluation, we created another test set (535 instances) from the 10% sample, with each instance belonging to more than one primary category. The accuracy result is shown below:

	k-NN	SVM
Accuracy	0.85(0.07)	0.85(0.06)

Table 4.2: Average Accuracy (With Standard Deviation)

These results suggest that our classifiers also performed well for multi-label categorization tasks.

4.5 More insights: Learning Curve

To better understand the difference between different binary classifiers and the effect of sample sizes, we plotted the Learning Curve (for SVM and k-NN) of business review for all ten primary categories.

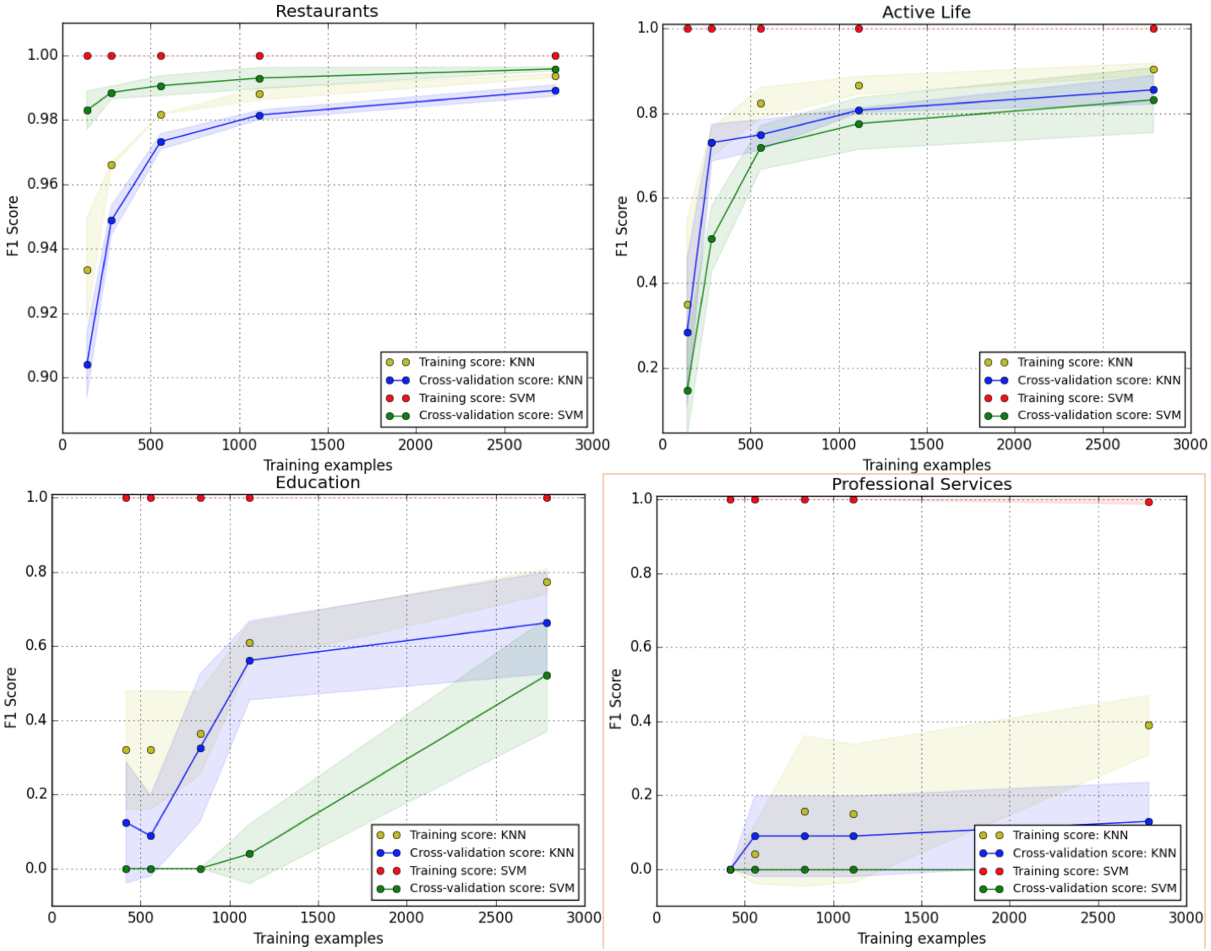


Figure 4.3: Learning Curve

Above, we present the learning curve for the four classifiers. The blue and green lines represent for the test scores for SVM and k-NN, respectively. When the sample size reaches about 2500, the gap between validation score and training score becomes stable.

Also notice the difference between various classifiers. Some classifiers have good curves that ultimately achieve small gaps between training and testing scores (e.g. Restaurant) while some remain a low score when the training size increases. The difference of Positive Response Rate partially explains this result. From “Restaurant”, “Active Life”, “Education” to “Professional Services”, the Positive Response Rate of the target value decrease dramatically: from 61% (“Restaurant”) to 6%, to 1.2% and 1.1%.

As a matter of fact, the following figure shows that when Positive Response Rate is less than 5%, F1 score drops almost perpendicularly.

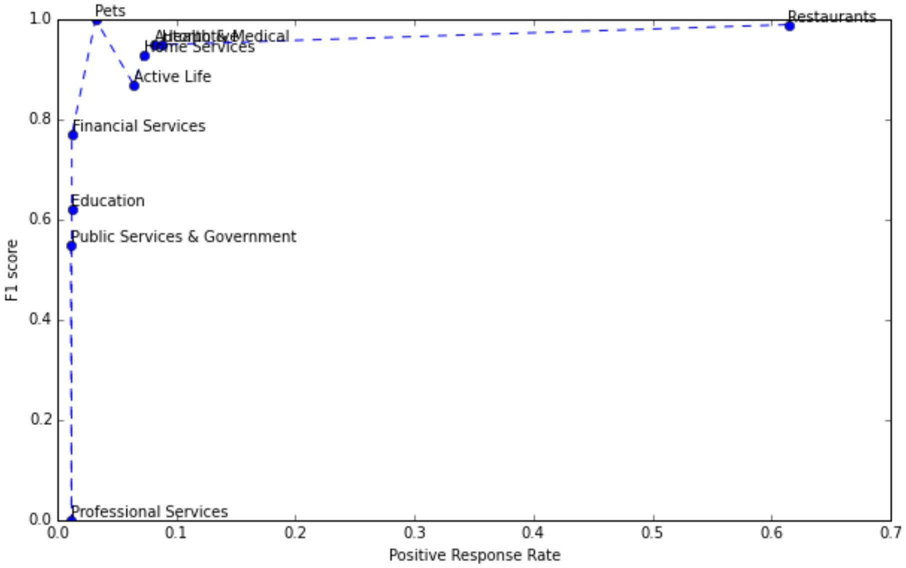


Figure 4.4: Relationship of PRR and F1-score

4.6 Word2vec

Finally, we wanted to understand the usefulness of word vectors [4] for the task of category prediction. We have built the word vectors from the entire training data available using the following configuration: vector dimension as 100, word2vec architecture as skip-gram model, context window size as 10. To evaluate the effectiveness of these word vectors we examined the Pearson correlation coefficient, which evaluates how close certain words are in the semantic space constructed by word2vec. The resulting correlation coefficient was 0.3, which was bad. The same Pearson correlation coefficient, when trained on a corpus of size of 1 billion words (not Yelp-specific) came to around 0.66. This gives an understanding of the inherent limitations of the Yelp reviews dataset, which suggests that the reviews do not follow proper grammatical structure, although they convey sentiment effectively.

Using the word vectors trained on Yelp reviews, we tried to classify each review to its closest category in the vector space. This was done by taking the average similarity of each word in the review across all the primary categories (after thresholding). Most of the reviews were classified in the “Restaurant” category, as this is the most predominant category present in our dataset. When we ignored the “Restaurant” category, the performance increased but did not match up to the accuracy of the previous experiments. This suggests that having a uniform distribution of data and well-organized data can help a lot when using word2vec as a feature for classification (which is not a real world situation). It is possible to explore improving word2vec for the skewed distribution of data, and we think it would help.

5 Deployment Plan

Below we discuss the different aspects of deployment and how we might implement them:

Categorization decision timeline:

- Businesses with no or very few existing categories: The biggest decision here is choosing whether to tag the categories online or offline. While online categorization is useful because it is faster and cost-effective, there is the risk of error. This would result in user dissatisfaction and a loss of credibility for Yelp. We recommend a two-tier categorization process, based on the prediction threshold. If our confidence of prediction is high, then we do the categorization online. Otherwise, we will do the categorization offline with a Yelp analytics team doing a secondary validation check.

- Businesses with existing list of categories: Since these already have a decent list of categories, our task is to identify new and emerging categories. Because these businesses have already been categorized in some way, the need to run a classifier should be of lower priority for Yelp. We suggest that Yelp run a model on these businesses less frequently to identify any new or emerging categories. We also recommend that Yelp use only new reviews to make the process more computationally effective. Even less frequently, Yelp should re-run their classifiers on the entire review datasets so that the business categories stay updated.

Decision Threshold: Coming up with a threshold in multi-class/multi-label classification is not straightforward and might need some insights from Yelp's business analytics team. We recommend, they make a clear observation of the issues that are being discovered internally or reported externally from customers. Depending on these useful insights the team can decide to update the threshold (to increase or decrease the threshold) and also determine the frequency at which they should be deciding.

Precision-Recall tradeoff: From the user's perspective, it would be in Yelp's best interest to build a model that maximizes precision over recall. A Yelp user would respond negatively if they were recommended an incorrectly categorized business, and this would hurt Yelp's credibility. On the other hand, a business would prefer a model that prioritizes recall, as it would increase their visibility to users. Yelp can determine what balance is optimal for their business by deploying several models and then performing cost-based analysis.

Updating production databases: The cost of updating production databases with these additions or changes in business category information is huge and for these to scale effectively with the overall business operations of Yelp, we recommend these changes to be done in bulk (during the period where the user activity is very minimal like 2.00 AM to 4.00 AM etc.) than individual updates.

References

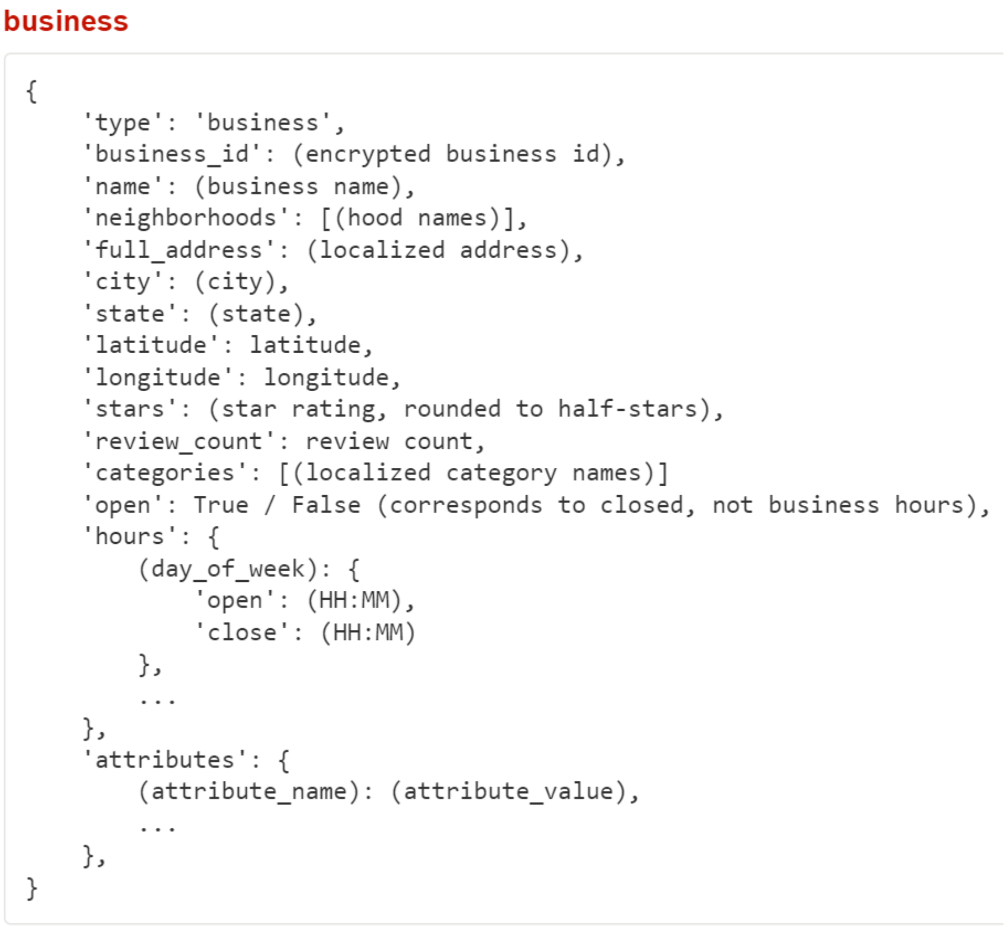
[1] “Data Science for Business” by Foster Provost and Tom Fawcett

[2] “Automatic categorizing Yelp Businesses”
http://engineeringblog.yelp.com/2015/09/automatically-categorizing-yelp-businesses.html

[3] Yelp business category list: (http://www.localvisibilitysystem.com/2013/07/19/yelp-business-categories-list/)

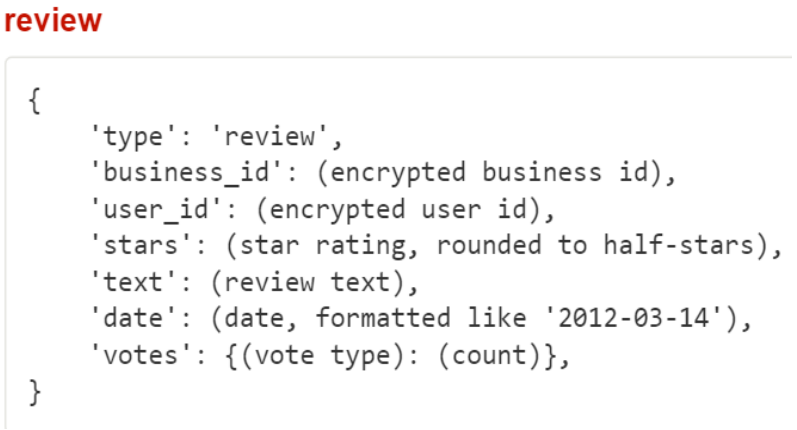
[4] word2vec (https://radimrehurek.com/gensim/models/word2vec.html)

Appendix I – Yelp Data Type Documentation



business

```
{
    'type': 'business',
    'business_id': (encrypted business id),
    'name': (business name),
    'neighborhoods': [(hood names)],
    'full_address': (localized address),
    'city': (city),
    'state': (state),
    'latitude': latitude,
    'longitude': longitude,
    'stars': (star rating, rounded to half-stars),
    'review_count': review count,
    'categories': [(localized category names)]
    'open': True / False (corresponds to closed, not business hours),
    'hours': {
        (day_of_week): {
            'open': (HH:MM),
            'close': (HH:MM)
        },
        ...
    },
    'attributes': {
        (attribute_name): (attribute_value),
        ...
    },
}
```

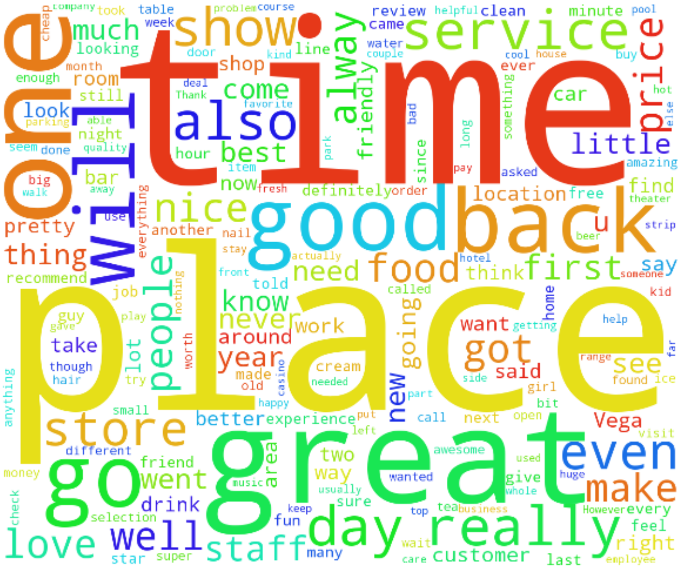


review

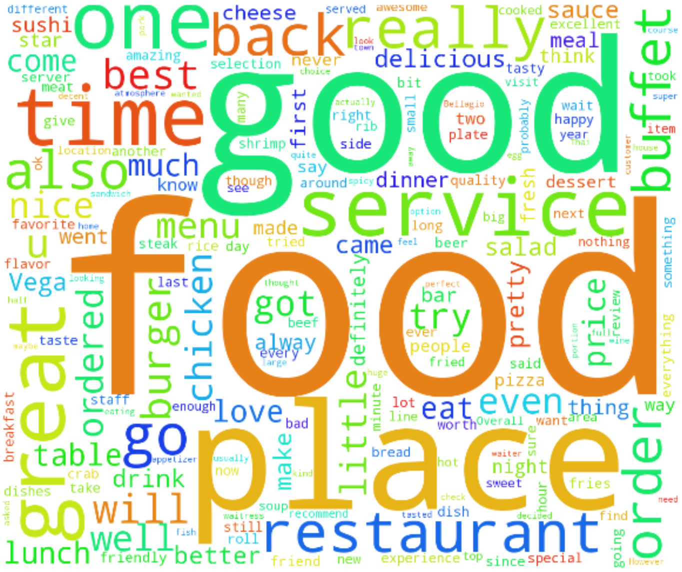
```
{
    'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)},
}
```

Appendix II – Word Clouds

Below are the word clouds for “Education” category reviews



Below are the word clouds for “Restaurant” category reviews


food
good
place
service
one
back
really
time
great
go
restaurant
buffet
order
also
best
much
nice
menu
chicken
burger
ordered
little
price
will
well
different
sushi
star
cheese
served
awesome
cooked
sauce
excellent
meal
course
come
amazing
selection
never
delicious
think
look
town
server
meat
choice
tasty
took
decent
atmosphere
wanted
bit
visit
super
give
many
actually
Bellagio
wait
right
two
happy
item
rib
plate
year
small
probably
Thai
location
another
shrimp
side
first
quite
say
egg
ok
know
though
see
around
spicy
dinner
quality
dessert
sandwich
customer
away
fresh
option
next
favorite
home
made
long
big
u
went
steak
rice
day
tried
came
feel
beer
salad
nothing
flavor
something
thought
last
perfect
Vega
looking
bar
got
try
half
definitely
review
beef
ever
pretty
full
portion
taste
wine
maybe
alway
every
people
large
fried
everything
said
huge
pizza
staff
enough
love
lot
even
thing
way
eating
minute
line
eat
area
breakfast
bad
worth
Overall
want
sure
appetizer
make
bread
waiter
hot
night
table
usually
kind
drink
crab
now
sweet
hour
fries
dishes
take
soup
check
waitress
recommend
tasted
dish
decided
find
need
asked
fish
still
roll
lunch
friendly
better
friend
new
experience
top
since
special
going
However
house