1. Abstract

Yelp is a local business directory service and review site with social networking features. It allows users to give ratings and review businesses. The review is usually short text consisting of few lines with about hundred words. Often, a review describes various dimensions about a business and the experience of user with respect to those dimensions. In this paper, we build a classifier that automatically classifies restaurant business reviews into those dimensions. We manually inspected a few hundred reviews for restaurant businesses and found 5 important dimensions and these include “Food”, “Service”, “Ambience”, “Deals/Discounts”, and “Worthiness”. We experimented with popular multi-label classification approaches using “unigrams”, “bigrams”, ”trigrams”, and “review ratings” as features and found that ensemble of classifiers produces the best results with a precision and recall of 0.72 and 0.71 respectively. The training and test dataset consisted of 8,000 and 2,000 data points respectively

2. Introduction

Yelp users give ratings and write reviews about businesses and services on Yelp. These reviews and rating help other yelp users to evaluate a business or a service and make a choice. The problem most users face nowadays is the lack of time; most people are unable to read the reviews and just rely on the business’ ratings. This can be misleading. While ratings are useful to convey the overall experience, they do not convey the context that led users to that experience. For example, in case of a restaurant, the food, the ambience, the service or even the discounts offered can often influence the user ratings. This information is not conceivable from rating alone, however, it is present in the reviews that users write. Our aim is to build a classifier that can classify the businesses into the five defined categories, based on the information present in the reviews. This information when presented to the user by classifying reviews into various relevant categories can prove to be very effective in making an informed decision. Moreover, such information can also be used to rank venues based on the categories.

Consider a Yelp review: **“They have not the best happy hours, but the food is good, and service is even better. When it is winter we become regulars”.** It is not difficult to identify that this review talks about only “food” and “service” in a positive manner, and “deals/discounts” (happy hours) are not that great. Extracting this information from this review and presenting it to the user, can help the user understand why the reviewer rated the restaurant “high” or “low” and make an informed decision, without reading the review. Although, the functionality described above is desirable and useful for any kind of business, we limit the scope of our classifier to only restaurants.

The classifier can be formulated as a learning problem, where the task is to build a learner. The learner can classify a given review into respective categories. However, since a review can be associated with multiple categories at the same time, it is not a binary classification or a multi-class classification. It is rather a multi-label classification problem.

A quick inspection of few hundred reviews helped us to decide important categories that are frequent in the reviews and worth extracting. We found 5 categories that include “Food”, “Service”, “Ambience”, “Deals/Discounts”, and “Worthiness”. The “Food” and “Service” categories refer to just that, food and service. “Ambience” relates to the décor, the look and feel of the place. “Deals and Discounts” correspond to offers during happy hours, or specials run by the restaurant. “Worthiness” can be summarized as value for money. Users often express the sentiment whether the overall experience was worth the money they paid. It is important to note that “worthiness” is different than the “Price” attribute already provided by Yelp. “Price” measures the overall expensiveness of a restaurant, whether it is “decently priced”, “expensive” or “very expensive”. It does not capture the sentiment or worth, which we are trying to attempt.

3. Problem Formulation

The classification of yelp restaurant reviews into one or more, “Food”, “Service”, “Ambience”, “Deals/Discounts”, and “Worthiness”, categories is the problem in consideration. Inputs are the Yelp restaurant reviews and review ratings. The multi-label classifier outputs the list of relevant categories that apply to the given Yelp review. Consider a Yelp review: **“They have not the best happy hours, but the food is good, and service is even better. When it is winter we become regulars”.** It is easily inferred that this review talks about “food” and “service” in a positive sentiment, and “deals/discounts” (happy hours) in a negative sentiment. Extracting classification information from the review and presenting it to the user, shall help the user understand why a reviewer rated the restaurant “high” or “low” and make a more informed decision, avoiding the time consuming process of reading the entire list of restaurant reviews. Although, the functionality described above is desirable and useful for any kind of business, we limit the scope of our classifier to only restaurants. The classifier can be formulated as a learning problem, where the task is to build a learner that can classify a given review into respective categories. However, since a review can be associated with multiple categories at the same time, it is not a binary classification or a multi-class classification. It is rather a multi-label classification problem.

Formal definition:

Let H be the hypothesis of multi-label classification and C are the set of categories, X is the review text and Y is output, then

$$C = \{\text{Food, Service, Ambience, Deals and Worthiness}\}$$

$$H: X \text{ -> } Y, \text{ where } Y \subseteq C$$

4. Background

Most text analysis research to date has been on well-formed text documents. Researchers try to discover and thread together topically related material in these documents. However short text mediums are becoming popular in today's societies and provide quite useful information about peoples current interests, habits, social behavior and tendencies ([4,5,8]). Large communities like Yelp are based on user generated content (UGC). Most of the time, the contributions from the users is in the form of short text contributions. Unlike traditional web documents, these text and web segments are usually noisier, less topic–focused, and much shorter. They vary from a dozen words to a few sentences. As a result, it is a challenge to achieve desired accuracy due to the data sparseness.

A number of classification research studies with various interests like detecting general emotion cues, empathy, verbal irony, certainty (or confidence) have been done [6]. These studies focus on binary classification. Other studies with interest to cluster items based on the short text are also done ([7.10]). Some research has also been done in the direction to improve the classification using personalized features, for example, ChatTrack [11], a text classification system, creates a concept based profile that represents a summary of the topics discussed in a chat room or by an individual participant. Use of domain specific features extracted from authors profile for short text classification in twitter is also explored to a certain extent [9]. Our work is different from these related work as we aim to classify a reveiw (short text) into multiple categories. So it is a multi-label classification problem. Nonetheless, we take inspiration from the above mentioned relevant work.

5. Data Set

The Yelp dataset released for the academic challenge contains information for 11,537 businesses. This dataset has 8,282 check-in sets, 43,873 users, 229,907 reviews for these businesses. For our study we have considered only the business that are categorized as food or restaurants. This reduced the number of business to around 5,000.

We selected all the reviews for these restaurants that had at least one useful vote. From this pool of useful reviews, we randomly chose 10,000 reviews. A labeling codebook describing what categories to include was developed through an initial open coding of a random sample of 400 reviews. The codebook was validated and refined based on a second random sample of 200 reviews. This exercise helped us to fix 5 categories that include food, ambience, service, deals, and worthiness. Once we identified these 5 categories, the 10,000 reviews were divided into 5 bins with repetition in each bin. 6 Graduate student researchers from our group then read and annotated each of these reviews in the identified categories. It took us approx. 225 man-hours to annotate all the reviews. We identified the conflicts in the annotation of reviews among different annotators. We removed all the reviews from the analysis where there were discrepancies among the annotators. This left us with 9019 reviews. We split these annotated reviews into 80% train and 20% test data.

6. Feature Extraction and Normalization

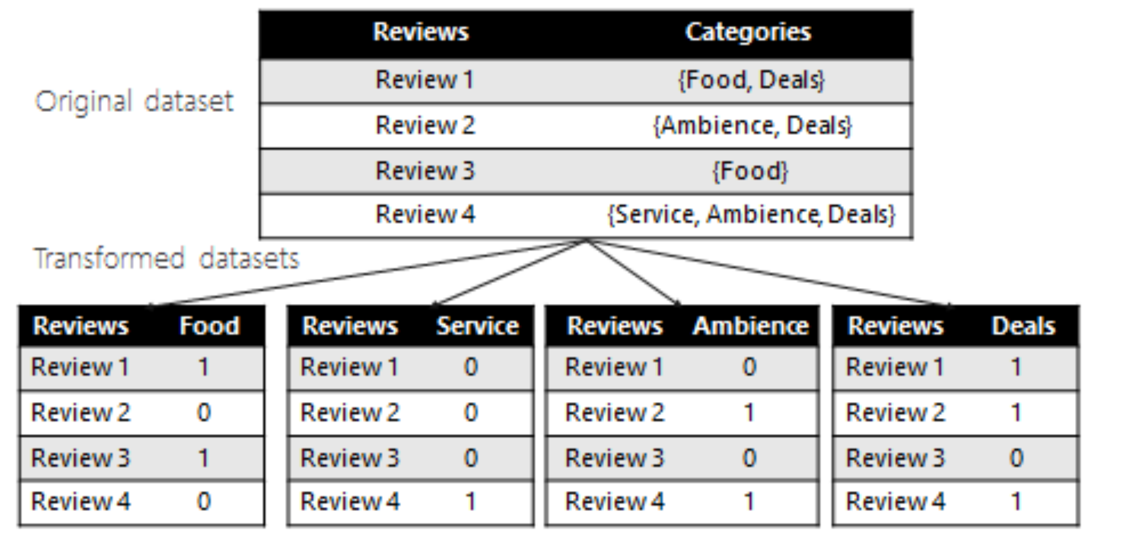
Two types of features are extracted from our corpus: (I) star ratings and (ii) textual features consisting of unigrams, bigrams and trigrams. For star ratings we created three binary features representing rating 1-2 stars, 3 stars, and 4-5 stars respectively. For extracting textual features, we first normalize the review text by converting it to lower–case, removing stop words and special characters. The cleaned text is then tokenized to collect unigrams (individual words) and calculate their frequencies across the entire corpus. This results in 54,121 unique unigrams. We condense this feature set by only considering unigrams with a frequency greater than 300, which results in 321 unigram features. This processing of textual features is similar to that of most prior work. We did similar processing to extract bigrams and trigrams.

7. Multi-label Classification

The problem of classifying a review into multiple categories is a multi-label classification problem, as a reviewer can like or dislike a place for many reasons, e.g., Food, ambience, service, etc. Formally, multi-label learning can be phrased as the problem of finding a model that maps inputs x to vectors y, rather than scalar outputs as in a ordinary classification problem. There are two main methods for tackling multi-label classification problems that are discussed in the literature: Problem Transformation methods and Algorithm Adaptation methods [1]. Problem transformation method transforms the multi-label problem into a set of Binary classification problems. The Algorithm adaptation method adapts the algorithms to directly perform multi-label classification.

Problem Transformation methods are further divided into two categories as follows:

- **Binary Relevance (BR)**
  This is the most widely used problem transformation method and it considers the prediction of each label as an independent binary classification task. It learns one binary classifier h: X --> {True, False} for each different label. It transforms the original dataset into |L| data sets such that each dataset contain all examples of the original dataset, labeled as “True”, if the labels of the original example contain the label and as “False” otherwise. The classification of a new data point is the union of the labels that are output by each of the |L| classifiers.



Original dataset

Reviews	Categories
Review 1	{Food, Deals}
Review 2	{Ambience, Deals}
Review 3	{Food}
Review 4	{Service, Ambience, Deals}

Transformed datasets

Reviews	Food
Review 1	1
Review 2	0
Review 3	1
Review 4	0

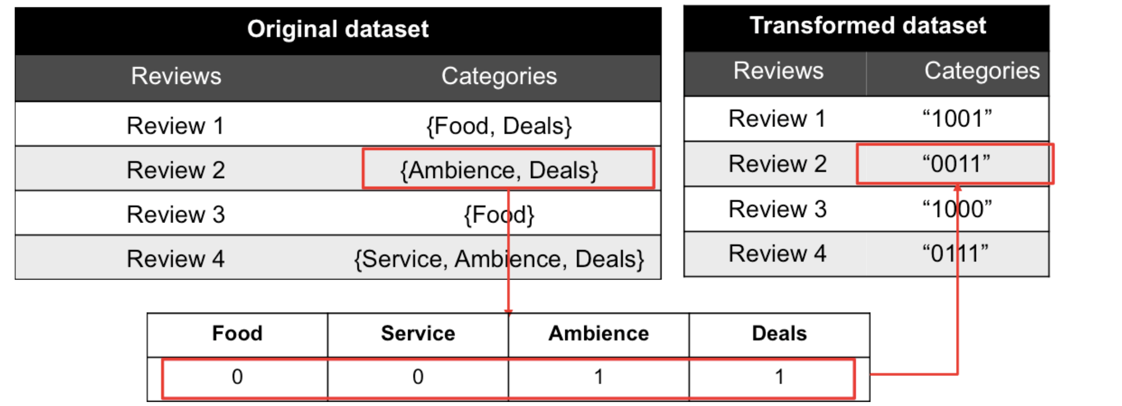
Reviews	Service
Review 1	0
Review 2	0
Review 3	0
Review 4	1

Reviews	Ambience
Review 1	0
Review 2	1
Review 3	0
Review 4	1

Reviews	Deals
Review 1	1
Review 2	1
Review 3	0
Review 4	1

Binary classifiers for each category

- **Label Power Set (LP)**
  BR ignores label correlation as it looks at each class independently. A less common problem transformation method considers each different subset of L as a single label. It so learns one single-label classifier h: X --> P(L), where P(L) is the power set of L, containing all possible label subsets. LP has the advantage of taking label correlations into account, but suffers from the large number of label subsets, the majority of which are associated with very few examples.



Original dataset	
Reviews	Categories
Review 1	{Food, Deals}
Review 2	{Ambience, Deals}
Review 3	{Food}
Review 4	{Service, Ambience, Deals}

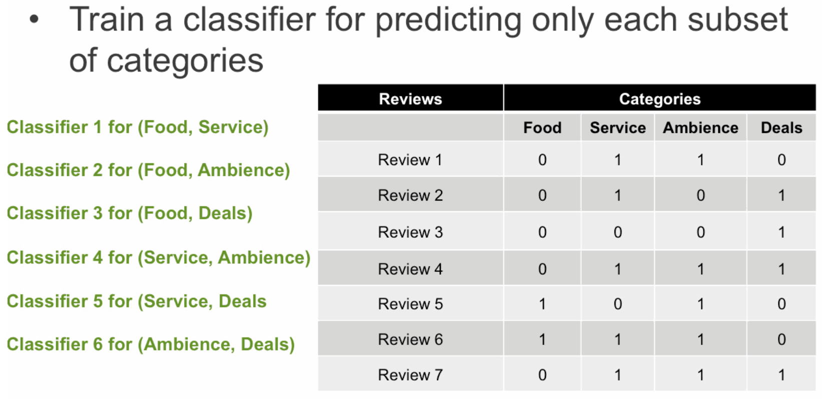
Transformed dataset	
Reviews	Categories
Review 1	"1001"
Review 2	"0011"
Review 3	"1000"
Review 4	"0111"

Food	Service	Ambience	Deals
0	0	1	1

Classifier for each subset of categories

- **Ensemble of Classifiers**
  Due to limited dataset for training, we had to rely on efficient training of our data. Hence, we use ensemble of LP classifiers where each classifier is trained using a different small random subset, say *k* of the set of labels. This approach aims at taking into account label correlations and at the same time avoiding the aforementioned problems of LP. Ensemble combination is accomplished by setting a threshold on the average zero-one decisions of each model per considered label. For example, when the total number of labels is 4 and k

= subset size = 2. Hence, we generate a total of six classifiers learning to predict this combination of labels {(L1, L2), (L1, L3), (L1, L4), (L2, L3), (L2, L4), (L3, L4)}. During prediction, you take prediction of all the six classifier and then take a majority vote.



- Train a classifier for predicting only each subset of categories

Classifier 1 for (Food, Service)

Classifier 2 for (Food, Ambience)

Classifier 3 for (Food, Deals)

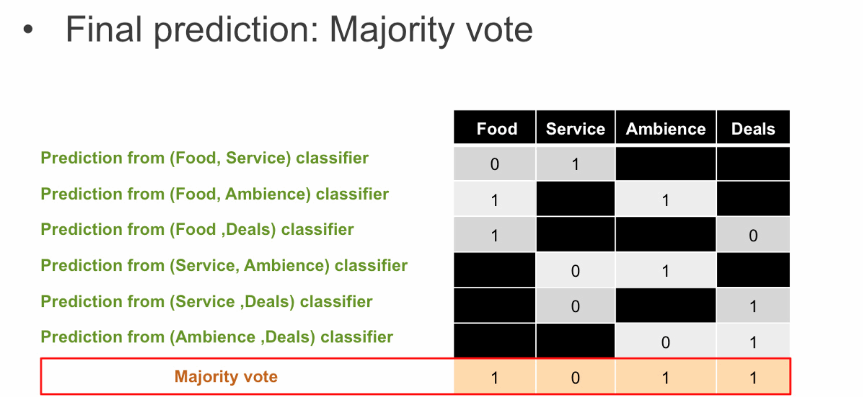
Classifier 4 for (Service, Ambience)

Classifier 5 for (Service, Deals

Classifier 6 for (Ambience, Deals)

Reviews	Categories			
	Food	Service	Ambience	Deals
Review 1	0	1	1	0
Review 2	0	1	0	1
Review 3	0	0	0	1
Review 4	0	1	1	1
Review 5	1	0	1	0
Review 6	1	1	1	0
Review 7	0	1	1	1

Ensemble of subset classifiers
Total 6 classifier for subset of size of 2 categories



- Final prediction: Majority vote

	Food	Service	Ambience	Deals
Prediction from (Food, Service) classifier	0	1	■	■
Prediction from (Food, Ambience) classifier	1	■	1	■
Prediction from (Food ,Deals) classifier	1	■	■	0
Prediction from (Service, Ambience) classifier	■	0	1	■
Prediction from (Service ,Deals) classifier	■	0	■	1
Prediction from (Ambience ,Deals) classifier	■	■	0	1
Majority vote	1	0	1	1

Ensemble of classifiers: Prediction

Due to limited data-set owing to manual annotations, the data observed is highly skewed and as expected Yelp reviews are more biased towards Food, Service and Ambience and thereby to model the classifiers better we duplicated the data belonging to Deals/Discount and Price for our model to predict those categories as well. Binary relevance algorithm used for multi-label classification here was a set of |L| Naïve Bayes classifier and for LP we used Decision trees and for Algorithm Adaption based algorithms we used Multi-Label KNN algorithm. Other multi-label classifiers Random K-label set (Rakel) and SVM were also modeled to train the data.

8. Performance Measures

Recall and Precision are the primary evaluation criteria used. Recall and Precision (R & P) have been evaluated for train and test data set over each of the above-mentioned machine learning classifiers. R & P index was also calculated simultaneously over each of the 5 categories as well and it was observed that all the classifiers were doing very well on predicting Food and Service categories and not very effective in predicting other three categories. Our problem is to classify a yelp review into 5 categories and thereby we built a classifier and hence, there is strict hypothesis. Our data set has manually annotated un-conflicted 9019 reviews and we split these annotated reviews into 80% train and 20% test data. Feature vectors (section 2.3) are generated for train and test data. Train data is used for modeling and test feature vectors are evaluated on the trained model and the R & P results show the effectiveness of the technique. To understand, what precision and recall means in our context, consider (X, Y) to be a data point where X is the review text and Y is the set of true categories, $Y \subseteq L$, where L = {Food, Service, Ambience, Deals, Worthiness}.

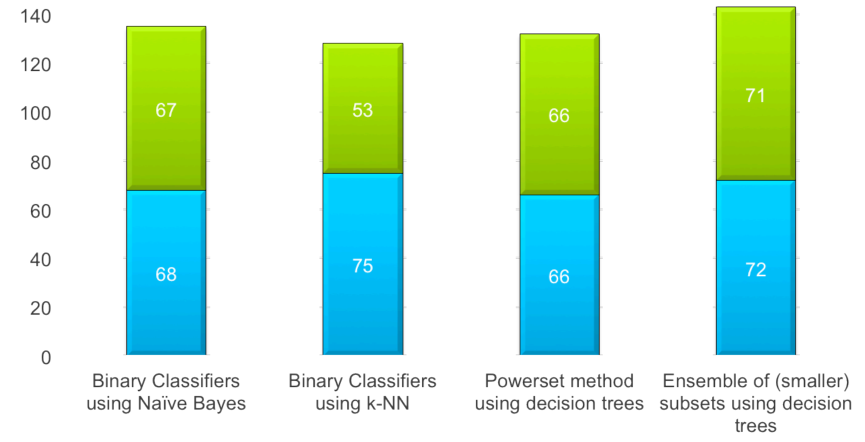
Let h be a classifier. Let Z = h (X) be the set of categories predicted by h for the data point(X, Y). Then,

$$\text{Precision} = |Y \cap Z|/|Z|$$
$$\text{Recall} = |Y \cap Z|/|Y|$$

9. Experimental Results

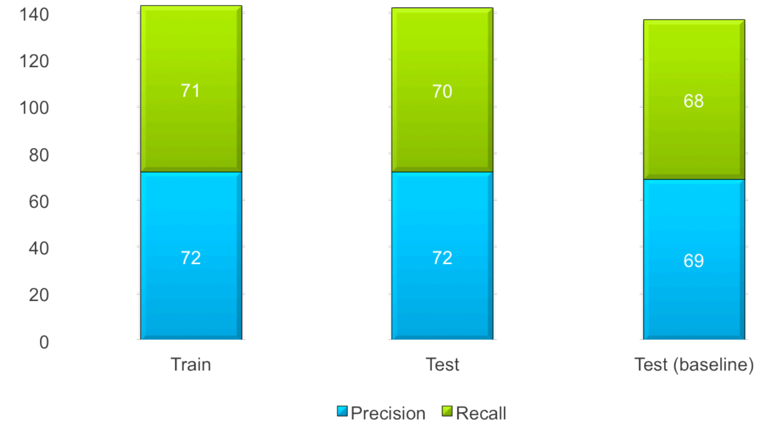
We experimented with all the three approaches discussed above in Section 7 - Multi Label Classification. We used Precision and Recall as our evaluation metrics. For each review, the comprehensive set of experiment configurations (different approaches, different classifiers, different feature sets, parameter settings) can be found later in this section. In the first approach of using L binary classifiers, where L is the total number of categories, we used Naive Bayes, k-Nearest Neighbor, Support Vector Machines (SMO implementation), decision trees, and Neural Networks. In the figure below we report results for Naive Bayes and K-NN for this approach, as only they were competitive. In the second approach, we consider label correlations and predict the power set of labels. Decision trees performed the best in this category. We also experimented with ensemble of classifiers approach using decision trees that gave us the best results overall.

It is observed that the Precision and Recall improvement after including bigrams and trigrams as features is marginally significant than when only unigrams with ratings were considered as features. This implies that the classifiers were able to learn enough from the dataset just by using unigrams.

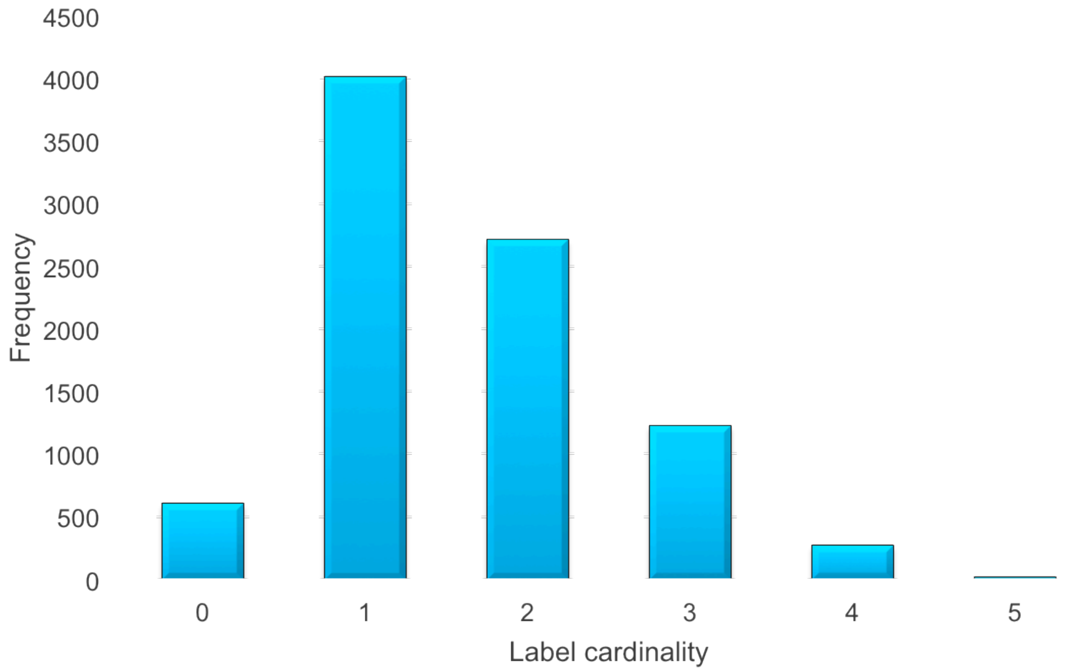


Training performance on 3-fold cross validation

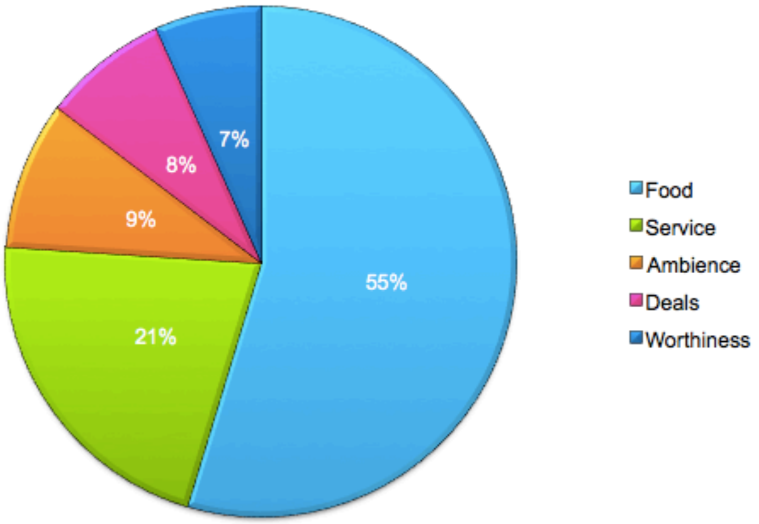
As ensemble of classifiers has the best results among all the techniques evaluated, the test data was run on this technique.



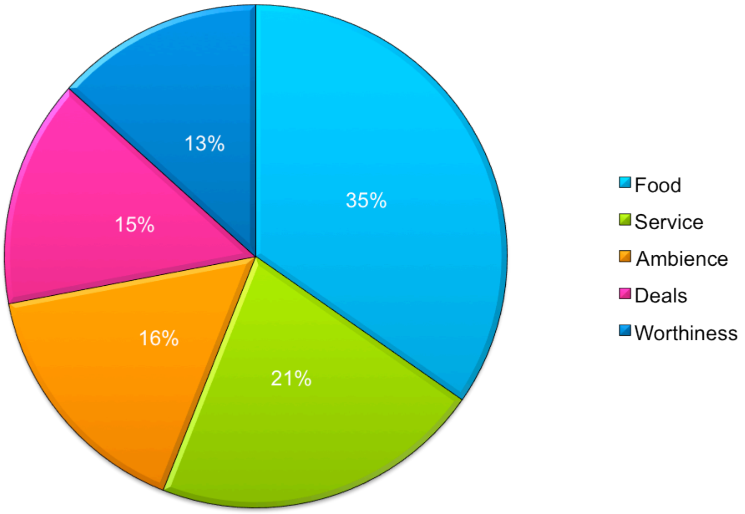
Train vs. Test performance of Ensemble of (smaller) subsets using decision trees



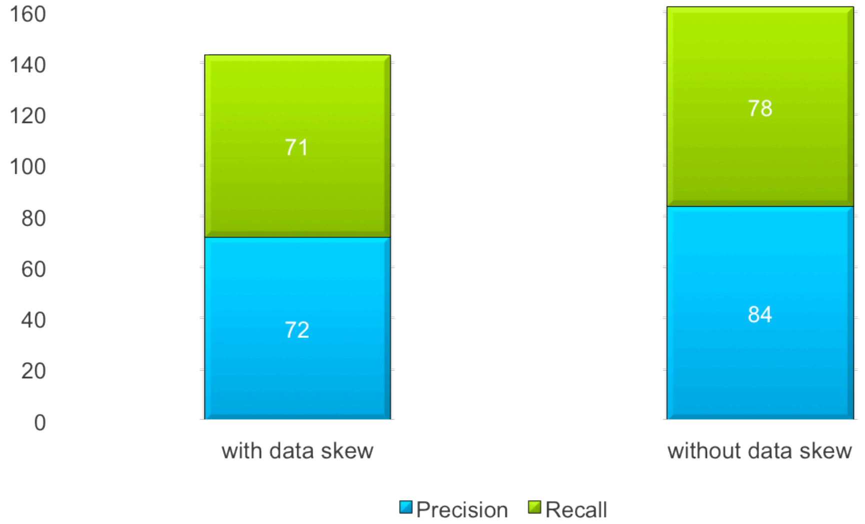
Average number of Labels/Categories per review is 1.74



As can be seen from above that distribution of data set is more biased towards a few categories



To adjust the skew nature of data distribution, replication of reviews for lower frequency categories led to such a normalized distribution and ensemble of classifiers was evaluated on this normalized data set and results are better than earlier (shown below).

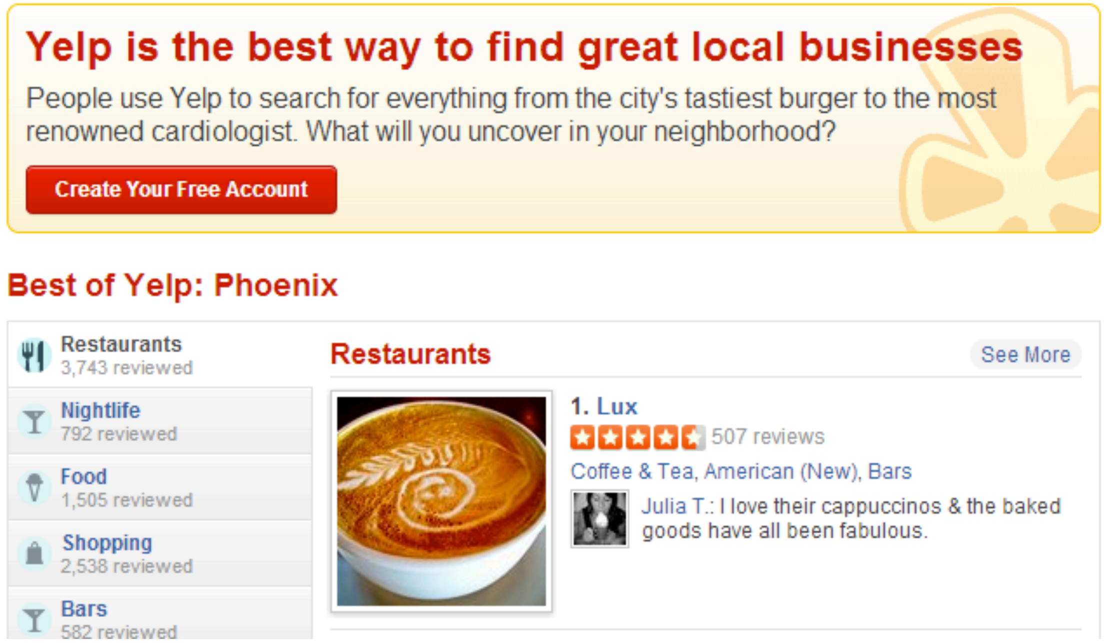


Cross-validation performance of Ensemble of (smaller) subset using decision trees. Significant improvement in classification when evaluated against the normalized data.

10. Implications of using this feature on yelp website

Yelp website offers various features to their users but lack the fundamental aspect of providing the reviews distribution based on categories like the problem discussed in this paper. Yelp can introduce this feature and provide the user with features like sort by Ambience, Service, Food, Deals and Worthiness and also show distribution of reviews of a restaurant. The following images detail them.

Case 1: Facilitate sort by categories



Yelp is the best way to find great local businesses

People use Yelp to search for everything from the city's tastiest burger to the most renowned cardiologist. What will you uncover in your neighborhood?

Best of Yelp: Phoenix

Restaurants	3,743 reviewed
Nightlife	792 reviewed
Food	1,505 reviewed
Shopping	2,538 reviewed
Bars	582 reviewed

Restaurants

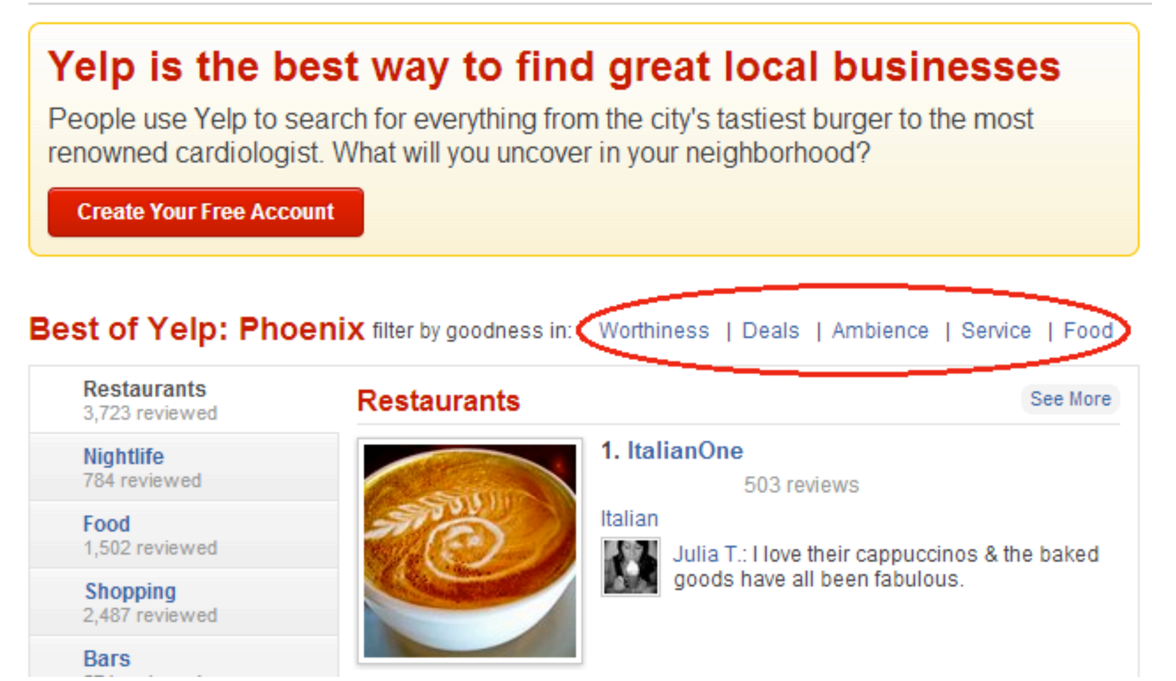
1. Lux

507 reviews

Coffee & Tea, American (New), Bars

Julia T.: I love their cappuccinos & the baked goods have all been fabulous.

Yelp website



Yelp is the best way to find great local businesses

People use Yelp to search for everything from the city's tastiest burger to the most renowned cardiologist. What will you uncover in your neighborhood?

Best of Yelp: Phoenix filter by goodness in: Worthiness | Deals | Ambience | Service | Food

Restaurants
3,723 reviewed

Nightlife
784 reviewed

Food
1,502 reviewed

Shopping
2,487 reviewed

Bars

Restaurants

1. ItalianOne

503 reviews

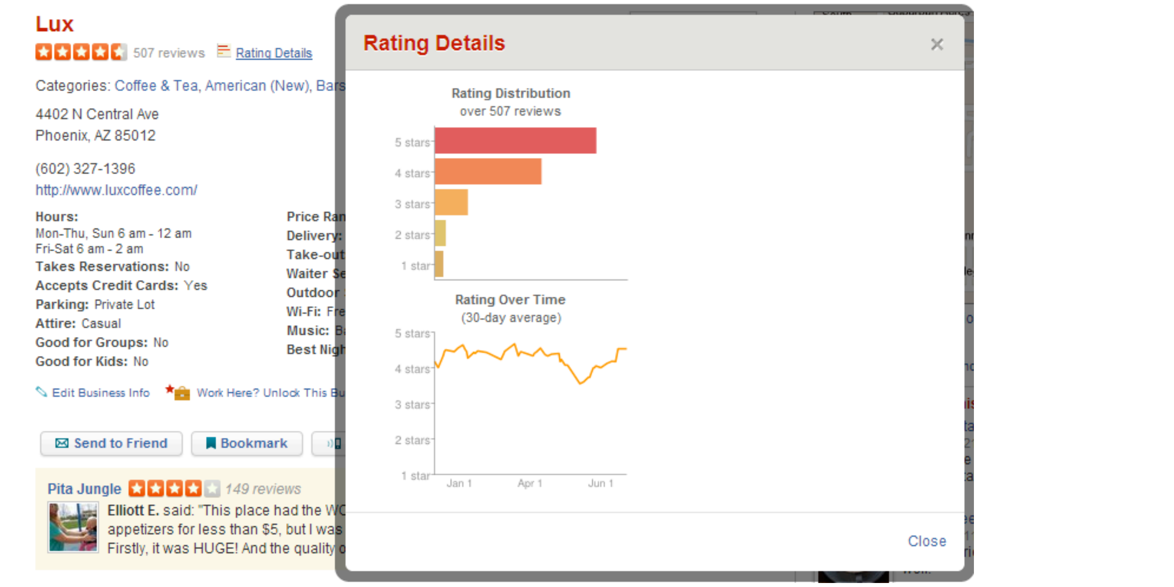
Italian

Julia T.: I love their cappuccinos & the baked goods have all been fabulous.

Mocked Yelp website

Case 2: Show Review Distribution

Yelp website shows the review distribution based on rating of users on the business. We propose to introduce review distribution based on the categories so that the Yelp user is aware of reason of popularity of restaurant based on the distribution of reviews on categories as shown below.



Lux

507 reviews Rating Details

Categories: Coffee & Tea, American (New), Bars

4402 N Central Ave
Phoenix, AZ 85012

(602) 327-1396
http://www.luxcoffee.com/

Hours:
Mon-Thu, Sun 6 am - 12 am
Fri-Sat 6 am - 2 am
Takes Reservations: No
Accepts Credit Cards: Yes
Parking: Private Lot
Attire: Casual
Good for Groups: No
Good for Kids: No

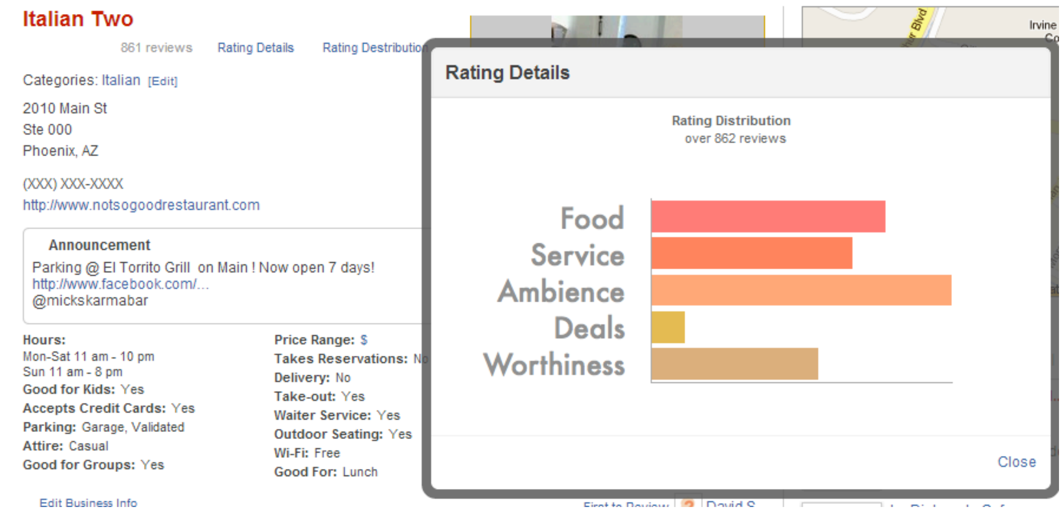
Price Ran
Delivery:
Take-out
Waiter Se
Outdoor
Wi-Fi: Fre
Music: B
Best Nigh

Pita Jungle 149 reviews

Elliott E. said: "This place had the W
appetizers for less than $5, but I was
Firstly, it was HUGE! And the quality o

Rating Details

Yelp website showing restaurant review rating distribution



Italian Two

861 reviews Rating Details Rating Destributio

Categories: Italian [Edit]

2010 Main St
Ste 000
Phoenix, AZ

(XXX) XXX-XXXX
http://www.notsogoodrestaurant.com

Announcement
Parking @ El Torrito Grill on Main ! Now open 7 days!
http://www.facebook.com/...
@mickskarmabar

Hours:
Mon-Sat 11 am - 10 pm
Sun 11 am - 8 pm
Good for Kids: Yes
Accepts Credit Cards: Yes
Parking: Garage, Validated
Attire: Casual
Good for Groups: Yes

Price Range: $
Takes Reservations: No
Delivery: No
Take-out: Yes
Waiter Service: Yes
Outdoor Seating: Yes
Wi-Fi: Free
Good For: Lunch

Edit Business Info

Rating Details

Rating Distribution
over 862 reviews

Close

Mocked Yelp website showing reviews distribution over 5 categories

11. Scalability

We did not face any issue with the runtime as well as memory requirements for running the experiments. We performed all the experiments on a laptop with 4 GB ram, 1.7 GHz i7 processor. One of the reasons is limited size of the dataset as our dataset consisted of approximately 10,000 reviews. However, we argue that the approach is scalable even in the case of larger datasets. The base classifier in our algorithm is a weak classifier – decision trees. It is known to classify datasets with millions of examples and hundreds of attributes with reasonable speed. Moreover, the algorithm for ensemble of classifier is an embarrassingly parallel problem because a classifier is trained for each subset of category independently. Hence, we can parallelize the task of learning k-subset classifier on different machines and combine their results during prediction. This ensures the scalability of the approach to ultra-large scale datasets as well. Moreover, we can abstract parallel processing using MapReduce, where, each mapper, is responsible, for a subset of categories.

12. Threats to Validity

The following are the threats to validity of our findings:

a) **Number of data points:**
Yelp data set has about 200,000 reviews over all businesses. Restaurants and food services account for 114,000 reviews. We have chosen 10,024 reviews that is a subset of 114,000 reviews of restaurants and food services. The results might be biased for the 10k reviews under consideration.

b) **Number of categories**
Our manual inference of 400 reviews to identify a list of probable discrete categories and further 200 reviews to shortlist to 5 categories Food, Service, Ambience, Deals and Worthiness. It is possible that many other prominent categories might exist if all the 114k reviews are explored.

c) **Manual annotation and verification**
6 researchers in the group manually annotated the 10,024 reviews and special focus and efforts were taken to avoid discrepancy in annotation by frequently taking a vote on categories for ambiguous reviews. Manual annotation has always sparked a debate and can pose threat to validity of our findings.

13. Conclusion

Yelp reviews and ratings are important source of information to make informed decisions about a venue. We conjecture that further classification of yelp reviews into relevant categories can help users to make an informed decision based on their personal preferences for categories. Moreover, this aspect is especially useful when users do not have time to read many reviews to infer the popularity of venues across these categories.

In this paper, we demonstrated how reviews for restaurants can be automatically classified into five relevant categories with precision and recall of 0.72 and 0.71 respectively. We found that an ensemble of two multi-label classification technique (Binary Relevance and Label Powerset) performed better than the techniques individually. Moreover, there is no significant difference in performance when using a combination of bigrams, unigrams and trigrams instead of only unigrams. We also showed how the results of this study can be incorporated into Yelp's existing website.

14. Acknowledgment

We would like acknowledge the efforts of our friends who helped us to annotate the data which made this study feasible. We also acknowledge Prof. Dave Newmann and Alexander Ihler for their support and useful discussions.

15. References

[1] Zhang, M.-L., and Zhou, Z.-H. 2007. Ml-knn: A lazy learning approach to multi-label learning, pattern recognition. Pattern Recognition 40(7):2038–2048.
[2] Tsoumakas, G.; Vilcek, J.; Spyromitros, E.; and Vlahavas, I. 2010. Mulan: A java library for multi-label learning. Journal of Machine Learning Research.

[3]Tsoumakas, G., Katakis, I., Vlahavas, I. (2010) “Mining Multi-label Data”, Data Mining and Knowledge Discovery Handbook, O. Maimon, L. Rokach (Ed.), Springer, 2nd edition, 2010.

[4] Khan, F.; Fisher, T.; Shuler, L.; Tianhao, W.; and Pottenger, W. 2002. “Mining chat-room conversations for social and semantic interactions”. In LU-CSE-02-011.

[5] Kose, C.; Ozyurt, O.; and Amanmyradov, G. 2007. “Mining chat conversations for sex identification”. In Proceedings of the 2007 international conference on Emerging technologies
in knowledge discovery and data mining, PAKDD’07, 45–55. Berlin, Heidelberg: Springer-Verlag

[6] Glazer, and Courtney. 2002. Playing nice with others: The communication of emotion in an online classroom. In 9th Annual Distance Education Conference

[7] Banerjee, S.; Ramanathan, K.; and Gupta, A. 2007. “Clustering short texts using Wikipedia”. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR’07, 787–788. ACM.

[8] Rosa, K. D., and Ellen, J. 2009. “Text classification methodologies applied to micro-text in military chat”. In International Conference on Machine Learning and Applications, ICMLA ’09, 710–714.

[9] Sriram, B.; Fuhry, D.; Demir, E.; Ferhatosmanoglu, H.; and Demirbas, M. 2010. “Short text classification in twitter to improve information filtering”. In Proceedings of the 33rd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR’ 10.

[10] O’Connor, B.; Krieger, M.; and Ahn, D. 2010. “Tweetmotif: Exploratory search and topic summarization for twitter”. In Proceeding of the International AAAI Conference on Weblogs and Social Media, ICWSM ’10.

[11] Bengel, J.; Gauch, S.; Mittur, E.; and Vijayaraghavan, R. 2004. “Chat room topic detection using classification”. In proceeding of 2nd Symposium on Intelligence and Security Informatics, 266–277.