Bringing you the Reviews that Matter

1. Introduction and Significance

Making a decision is hard and the wealth of information that people now have at their fingertips does not make things easier. Whether in search of a new place to eat or just to go shopping for the evening, people always wonder where to go. A resource that many turn to is Yelp, which has gone from users rating and reviewing restaurants to giving opinions about all sorts of businesses. However, given the large number of reviews written each day, it is often difficult to filter the relevant reviews.

Yelp offers a feedback mechanism on reviews which allows users to leave votes on an individual review. With the votes on a review as a measure of social endorsement and hence relevance of a review, we aim to answer the question - which reviews are likely to receive social endorsement and what aspects of a review promotes endorsement?

While social endorsement may not be a perfect measure of whether a review was relevant to a reader, previous studies have found links between social endorsement and actual influence on users and business outcomes (Leslie et. al., 2017), albeit in different domains and circumstances. We do not claim that such links hold true in our domain, but it seems plausible. Therefore, we think that predicting social endorsement is an appropriate starting point for gaining insight into what users are influenced by while reading reviews.

In a world filled with decision-making, people come to rely on tools that may help them make a decision. Through studies like these, review platforms can tailor themselves to highlight the important reviews. Ranking reviews by the votes they have received suffers from a cold-start problem - new reviews which have not received any votes cannot be ranked. Our work leverages the fact that an assessment of their relevance can be made even in the absence of this information, by looking at the text and user history. Thus the visibility of highly relevant reviews can be improved. Businesses can also gain potential insights from this research about what aspects are important to users.

2. Related Work

To encourage both the readers and the content creators in engaging with the platform, many online communities offer a feedback system. Facebook likes, Twitter's likes or retweets, Amazon's 'helpful' votes on reviews and Yelp's 'useful', 'cool' and 'funny' votes are all examples of such systems. Prior research by Archak et al. (2007) and Lu et al. (2009) suggests that attributes of the review text, the reviewer, and social context may contribute towards shaping the user's response to a review.

Bakhshi et al. (2015) attempted to understand what constitutes a high quality review, as perceived by readers. Some of the business, review, and reviewer features are used to study the feedback system. Using regression analysis, they found that active and regular members are the highest contributors to good quality reviews and longer reviews have higher chances of being popular in the community. The paper completely disregarded features about the exact content of the review and the network of the reviewer. We think that measures of activity of the social network of a user, and the themes and topics expressed in a review could be useful predictors of the votes received, hence we advance this work by adding these features. At the same time, we also validate the results of review and reviewer features by using both regression and classification models.

Ghenai et al. (2014) explored the most relevant features that makes a review 'useful', 'funny', and 'cool' and with the main objective to classify reviews between the three signals. Business, review, reviewer, and basic reviewer network features are used. The supervised learning techniques used in this study return a confidence score for every classification instance which gives an idea of how useful, cool, or funny a review is. Given the promising results, the paper proposed using the review classifier to order reviews based on their confidence score and recommend the ones with the highest score values. The paper runs into the problem of rare class classification since the proportion of ‘useful’ reviews is very high (99%), as the authors acknowledge. They also run into some issues downstream - when ranking reviews shown to users, a single score is required. We mitigate these two problems by using total votes received by a review as the target variable (for regression), and a binary value indicating whether the review received any votes (for classification). This results in a single score that can be used for ranking, and also results in a balanced class distribution. We also improve upon this work by adding features about a reviewer’s network and content-related features.

On Amazon platform, Mudambi et al. (2010) worked on customer reviews to understand what makes a helpful online review. An analysis of 1,587 reviews from Amazon.com across six products indicated that review extremity, review depth, and product type affect the perceived helpfulness of the review. The dependent variable is a measure of helpfulness calculated using the proportion of ‘yes’ votes to the total votes cast on helpfulness. They used a regression model to analyze the model. Our work is inspired from this work and adapts it on another platform Yelp.

Further, extensive work has been done towards predicting the number of retweets on Twitter. Petrovic et al. (2011) worked on predicting if a tweet will be retweeted to understand message propagation within large user communities. They found that the most important features for prediction are the identity of the source of the tweet and retweeter. Zaman et al. (2010) trained a probabilistic collaborative filter model to predict future retweets using data of who and what was retweeted. Kupavski et al. (2012) forecasted the number of retweets a given tweet will gain during a fixed time period using additional important features like the flow of the cascade and PageRank on the retweet graph. Yu et al. (2015) further argue that retweet prediction should be a regression analysis problem, not just a classification problem. Experimental results showed that the regression model has a better prediction accuracy in dealing with retweet prediction as compared to a classification model. Our work incorporates this argument with respect to the Yelp platform by incorporating both classification and regression models.

Chen et al. (2017) have considered post content and user engagement to predict Facebook likes, while Fang et al. (2016) predicted community endorsements in online discussions on Reddit Platform. Although extensive work has been conducted on endorsement on most popular social networking platforms, the Yelp social network and how social endorsement works on it has not been exhaustively explored. We, therefore, propose to study this in more detail, particularly by taking into account a reviewer’s network features and the content of the review to predict social endorsement of a given review, in addition to other commonly used features. Lastly, prior work on Yelp data does not take into account the category of the business that a review is written for (e.g. restaurants, shopping etc.). We think that this may influence endorsement received by reviews, hence, we have also attempted to study social endorsement patterns across different business types.

3. Objectives, Goals and Outcomes

The problem that we aimed to tackle through the project was predicting the social endorsement received by a review. We had four research questions that we wished to answer:

RQ1: Is it possible to build a model to accurately predict social endorsement of reviews?

RQ2: Does the content, style, and tone of a review affect endorsement?

RQ3: Does the “who” behind the review influence social endorsement?

RQ4: Do the same features matter for predicting endorsement across business types?

We examine the answers to each of these questions in Section 6: Results.

In addition to commonly used features like past reviewing activity of the user and high-level characteristics of the review, we wanted to examine features based on the content of the review as well as the social network of the reviewer. Broadly, our aim was to see whether the sentiment, content, readability, length of the review, the past activity of the reviewer, and the activity of the social network of the reviewer were relevant in predicting the social endorsement received by their reviews.

We have successfully achieved these objectives. Our final results include regression models for three different business types that predict the number of votes received by a review, by using all the aforementioned features. We have also quantified the importance of each feature, and compared the roles of features across different business types.

Since our expected outcome of predicting the number of votes received by a review did moderately well, we also attempted to classify whether a review will receive any social endorsement or not. A major point of difference between our expected and actual outcome was the fact that the textual content of the review did not heavily contribute towards the performance of the models, while the features related to the reviewer did.

4. Description of Work Accomplished: Data

Yelp is a local search service that offers crowd-sourced reviews and ratings of local businesses, allowing users to view, upvote, and write reviews about the experiences they have had. Spread across different rounds, the Yelp Dataset Challenge gives academics an opportunity to explore the Yelp data corpus and derive interesting insights from it. For the purpose of our project, we have used the 12th Round of Yelp Data. The dataset is fairly extensive, comprising of 5,996,996 reviews of 188,593 businesses, written by 1,518,169 users who are spread across 10 metropolitan areas.

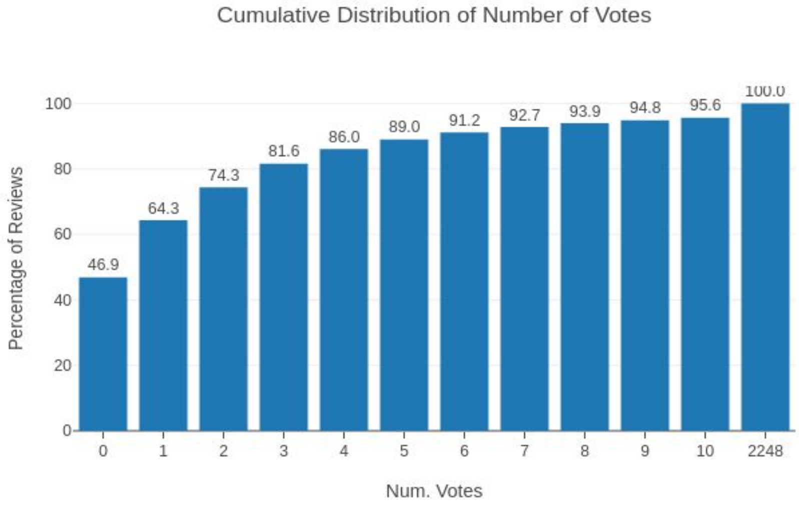


Figure 1: Cumulative distribution of the number of votes received by ~6M reviews

For the purpose of predicting social endorsement that a review is likely to receive, we consider the total votes received by a review as our target variable, measured by the sum of cool, useful and funny votes received by it. As shown in Figure 1, 90% of reviews have votes in the range of 0-5. For classifying whether a review will receive ‘any’ social endorsement or not, this data is fairly balanced given that 46.9% of the reviews have received no votes at all while 53.1% of the reviews have gained one or more votes.

Given the size of our dataset, our first task was to sample the data to a reasonable size while still ensuring that we have an even distribution. Since we aimed at comparing the performance of our classifier across different business types, we chose ‘Restaurants’, ‘Home Services’ and ‘Beauty & Spas’ as the categories we planned to study. The three categories were chosen because of the mutually exclusive nature of the business types, and since they were among the most popular business types on Yelp. The count of businesses belonging to the 'Restaurants', 'Home Services' and 'Beauty & Spas' category were 57,191, 18,634, and 18,967 respectively. Given that the size of the data was too big to handle locally, we decided to sample the data based on the business and then extracted the reviews and the reviewers corresponding to the sampled businesses. All businesses having greater than 100 reviews were sampled, resulting in a dataset of 180 businesses under 'Restaurants', 248 businesses under 'Home Services', and 255 businesses under 'Beauty & Spas'. The corresponding review counts were 50,218, 47,941, and 50,340 respectively. We did not filter or sample based on geographical location, since we wanted our results to be generalizable and not restricted to a specific subset. We used a 25% held-out test set for each of the three business types.

5. Description of Work Accomplished: Approach

5.1 Feature Engineering

The features used to train the models can be classified into four broad categories, namely: *Review Content, Other Review Characteristics, Reviewer History, and Reviewer Network*.

The *'Review Content'* category captures the text content of the review, and thus, any semantic information present. For the purpose of capturing this information, we applied bag of words and LDA topics. Prior to applying these transformations, we filtered out words belonging a stopword list to eliminate common meaningless words, and also words present in less than 50 reviews, to eliminate the long tail of words which are present in too few reviews to be useful. For bag-of-words, we used the 2,000 most frequent words after this filtering to keep the dimensionality reasonable, and 100 topics for LDA. The intuition behind using these features was that there could be specific words or themes that review readers care about which leads them to leave a vote on the review. The *'Other Review Characteristics'* category included the following features: *Review Stars*, *Sentiment*, *Readability Score,* and *Length*. *Review Stars* and

Sentiment are a potential measure of the polarity of the review and can be used to gauge if readers prefer extreme reviews or balanced reviews. *Readability Score* was measured using the text_standard function from the textstat library - we wanted to see whether easy-to-read reviews with simpler language were more endorsed by readers. Lastly, *Length* was calculated using character count.

The *'Reviewer History'* category takes into account the user's past reviewing activity. It consists of: *review_count*: Number of reviews written in the past, *user_score:* Average votes received on past reviews, *elite_count*: Number of years where user held the Elite status, *days_on_yelp*: how many days the user was on Yelp before posting the review, *avg_stars_given*: average star rating of the reviews given by the user.

Lastly, the *'Reviewer Network'* category measures how active the user is on Yelp socially, and also how active the user's friend network is. It includes: *fans*: number of fans that follow the user, *friend_count:* how many friends the user has, *max_network_score* and *avg_network_score.* The *user_score* of any user is the average number of votes received on past reviews. We calculate this score for all friends of the user, and take the maximum and average of this score respectively to obtain *max_network_score* and *avg_network_score.* The goal of these two features is to capture whether or not the user's friends are prolific writers on Yelp and check if their Yelp activity affects the activity of the original user.

5.2 Regression

We first trained regression models on the review data to predict the total number of votes received by each review. Our reasoning behind this was that any such vote is an indicator of endorsement.

We ran different regression models on the data - Ordinary Least Squares (OLS), Weighted Least Squares, Negative Binomial, Poisson, and Tobit. As shown in Figure 1 above, the target variable has a highly non-uniform distribution, and hence we thought simple linear regression might not be a great fit. We intended OLS to be a simple baseline, and Weighted Least Squares (with the log of the vote count as the weight) to improve upon OLS by assigning higher weights to reviews with more votes, which are fewer in number. We used Poisson and Negative Binomial regression models because they are commonly used to model count variables, and Negative Binomial also takes into account over-dispersed count variables.

As for Tobit regression, it has the additional property of correcting for selection bias. Yelp does not indicate the number of persons who read the review, only the number of individual votes on a review. Since it is unlikely that all readers of the review voted on these votes, there is a potential selection problem. According to Kennedy (1994), if the probability of being included in the sample is correlated with an explanatory variable (e.g. length or extremity of review), the OLS and Weighted Least Squares estimates can be biased. We used R-squared as our evaluation metric for all regression models.

5.3 Classification

We also trained classification models to predict whether or not a review receives any votes at all. We wanted to see the best performance we could get, but also be able to interpret the models to understand the contribution of individual features. Tree-based models (Random Forest and XGBoost) were more likely to do better, but simple linear models (Linear SVM and Logistic Regression) are easier to interpret, so we used all four models. Accuracy was used as the evaluation metric, since we had a fairly balanced class split - with 46.9% of the reviews having no votes and 53.1% of reviews having one or more 1 votes.

5.4 Feature Importance and Business Types

We quantify the importance of each feature by looking at the value of its coefficient in the best-performing regression and linear classifier models. We only consider features with a p-value < 0.001. In order to ensure that the difference in ranges of the features does not affect the coefficient values, we perform normalization of the data before performing either classification or regression. To analyze feature importance for different business types, we trained individual classification and regression models for each business type and compared the most important features for each business type. To train these individual models, we fit a separate bag-of-words model vectorizer and topic model for each business type, so the resulting topics and vocabulary are different for each model.

6. Description of Work Accomplished: Results

Following are the results generated from both our classification and regression models on restaurants using a combination of various feature sets. We then go on to discuss how these results answer each of our research questions.

Features	**Train Accuracy**	**Val. Accuracy**	**Test Accuracy**
Content (Bag of Words)	0.652	0.631	0.631
Content (Topics)	0.663	0.641	0.641
Review (Sentiment, Readability, Stars, Length)	0.641	0.639	0.637
Reviewer (Fans, Friends, Elite, Network Score)	0.68	0.676	0.671
Topics + Review	0.666	0.642	0.647
Topics + Reviewer	0.705	**0.689**	**0.688**
Review + Reviewer	0.705	**0.699**	**0.691**
Topics + Review + Reviewer	0.712	**0.693**	**0.693**

Table 1: Classification accuracies of best-performing models for restaurants

Business	**Model**	**All Features**		**Topics + Review Features**		**Reviewer Features (including network)**	
		R^2 Train	**R^2 Test**	**R^2 Train**	**R^2 Test**	**R^2 Train**	**R^2 Test**
Restaurants	Tobit	0.365	**0.337**	0.117	0.116	0.342	**0.312**
	Poisson	0.380	0.233	0.124	0.098	0.269	0.232

Table 2: Regression R-squared values of the two best-performing models for restaurants

RQ1: Is it possible to build a model to accurately predict social endorsement of reviews?

The motivation behind the question was to get an idea of the limits of prediction in this problem. Given that endorsement depends on a lot of factors, simply predicting it from data on the review and reviewer may not always be possible.

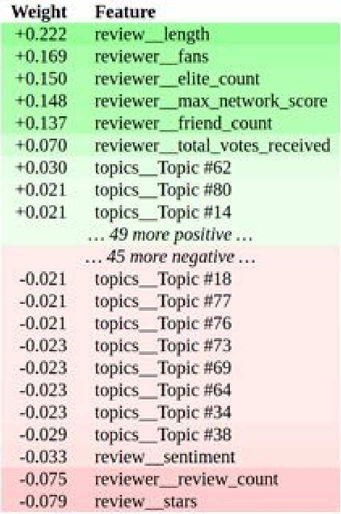
The best accuracy we achieved on the held-out test set was 69.3% for restaurants, obtained by an XGBoost model trained on the complete feature set. The best R-squared value was 0.337. This level of performance is reasonably good, but we think there is scope for improvement. So, we consider why our machine learning models may be making errors.

One of the potential explanations for this is that voting on Yelp can be fairly arbitrary and noisy - users might not always think much before voting on a review. We have two points that support this explanation. Firstly, looking at the distribution of vote counts on reviews shown in Figure 1, we notice that a large number of reviews have very few votes to no votes. By manually looking at the reviews that receive few or no votes, we see that many of them contain content that could potentially be useful to a prospective customer. This indicates that the voting data may be quite noisy. Secondly, the train accuracies are fairly close to the validation and test accuracies. This indicates that it is not that our models are overfitting to the data, but that it is a genuinely hard machine learning problem to solve. Again, a potential explanation for this is that the input data is too noisy. Of course, there may be other reasons too - lack of expressivity of our models or unsuitable feature representations. For example, topics and bag-of-words only capture semantics of the review at a very high level. To gain a more in-depth understanding of why our models are not doing as well as we expected, we move on to the next questions that look into the impact of different kinds of features.

RQ2: Does the content, style, and tone of a review affect endorsement?

In any case, when it comes to predicting the social endorsement of business reviews, it only stands to reason that we would start by looking at the reviews themselves to try and extract features that would be useful to our models. However, to our surprise, we found that such features alone were not particularly good indicators. Not only did the models that used review features have lower accuracies and R^2 values (compared to those that utilized reviewer and network features), but only two review features ranked high when we analyzed feature importance. A discussion of these features and the possible reasons we have for their relative unimportance can be found below.

The review-based features that we extracted from the reviews were: bag of words, review length, number of stars a review received, readability, sentiment, and topics. We started simple (bag of words and review length) and began moving to more complex features such as sentiment and topics. All these features were run through the aforementioned regression and classification models and the results can be seen in Tables 1 and 2. In addition, the heatmap below indicates the features with the highest coefficients for our linear classifier.



Weight	Feature
+0.222	review__length
+0.169	reviewer__fans
+0.150	reviewer__elite_count
+0.148	reviewer__max_network_score
+0.137	reviewer__friend_count
+0.070	reviewer__total_votes_received
+0.030	topics__Topic #62
+0.021	topics__Topic #80
+0.021	topics__Topic #14
	… 49 more positive …
	… 45 more negative …
-0.021	topics__Topic #18
-0.021	topics__Topic #77
-0.021	topics__Topic #76
-0.023	topics__Topic #73
-0.023	topics__Topic #69
-0.023	topics__Topic #64
-0.023	topics__Topic #34
-0.029	topics__Topic #38
-0.033	review__sentiment
-0.075	reviewer__review_count
-0.079	review__stars

Figure 2: Heatmap and ranking of feature importance in the classification model

The heatmap of the regression model can be found in the RQ4 section of this report (see Table 5), but suffice it to say that the relative feature importance heatmap was very similar to Figure 2. We see in Tables 1 and 2 that adding review-based features (including topics) to purely reviewer-based features does not help our cause very much. We also notice in both tables that the models that used just review-based features had the worst performance. Lastly, from Figure 2, we notice that the features that tended to be the most important were primarily not review-based features. This suggests that there is not much correlation between the content, tone, and style of a review and the amount of social endorsement the review will get.

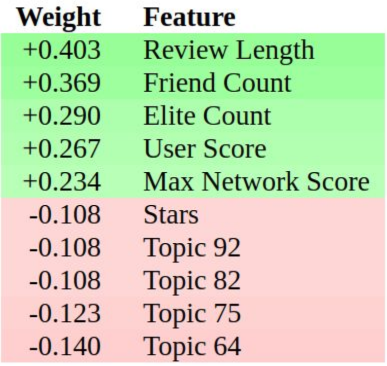
So now for the question that we spent a fair bit of time answering: why in the world would the content of a review not matter when it comes to predicting how much social endorsement a review gets? The team believes this may have to do with the nature of Yelp itself and how the use of Yelp's sort option may introduce this disconnect. Intuitively, voting behaviour depends on how technological features of the platform make reviews visible, hence voting could depend on factors unrelated to the review itself. A more detailed discussion regarding this topic can be found in the 'Discussion of Outcomes, Implications, and Conclusions' section.

However, we notice that user features do tend to help models perform better and encompass many of the important features, so we next go into some more detail about those.

RQ3: Does the "who" behind the review influence social endorsement?

Given that the content and readability of Yelp reviews are not sufficient to produce highly accurate predictions of social endorsement, we considered taking information about the reviewer into account for our two problems. The intuition behind the approach was that the visibility of a review strongly depends on the popularity of the reviewer and the network he is a part of, given the social nature of the Yelp platform.

We included features like votes per review received by the reviewer in the past, his/her past ‘elite’ status on Yelp and details of the friend network he/she had. This led to a considerable increase in the accuracy of our classifier - around 5% absolute improvement.



Weight	Feature
+0.403	Review Length
+0.369	Friend Count
+0.290	Elite Count
+0.267	User Score
+0.234	Max Network Score
-0.108	Stars
-0.108	Topic 92
-0.108	Topic 82
-0.123	Topic 75
-0.140	Topic 64

Figure 3: Feature importance for regression model trained on restaurant reviews

Further, an analysis of the feature importance of our model (from Figure 2 and 3) indicates that reviewer features do have strong positive correlations with the results. Amongst the top features that we observed, reviewer related attributes score high, like the number of friends the reviewer has on Yelp, the number of years he/she has been awarded Elite status, the votes per review he/she has received and the most popular friend in his/her network in terms of votes per review received. Given that the user score serves as an important feature, it is indicative of the fact that reputation on Yelp is built over time with each review one writes and every vote one gets. Another interesting observation was that out of the two network features that we used, namely *average network score* and *maximum network score*, the *maximum network score* was found to be more significant. This means that a user who has a single prolific friend, rather than a larger number of less prolific friends is more likely to gain endorsement. Of course, this is probably not a causal relation - this feature is likely linked to other characteristics of the user's network. We hypothesize that most users have a large number of friends who are not very active writers on Yelp, and hence the average network score is fairly low for all users. However, the fact that a user is connected to a highly prolific user probably indicates that the user has an active network in general, and thus gains more social endorsement on their reviews.

While the reviewer features did significantly boost the classifier, its overall performance yet has space for improvement, at ~70% accuracy. However, given the nature of the Yelp platform and the manner in which it curates a user's feed and a business' review feed, it seems inherently difficult to predict the voting preferences and behaviour of Yelp users. Most of the reviews that users view is by selecting a specific business, which are in turn sorted based on Yelp's custom sorter or user-defined preferences. Reviewer features have a strong correlation with social endorsement since the default Yelp feed shows reviews of people in the user's network. Hence, the problem statement that our classifier soughts to answers is fraught with challenges ranging from how users interact with the Yelp platform to how Yelp delivers content to them.

RQ4: Do the same features matter for predicting endorsement across business types?

We follow the same process for two other business types - 'Home Services', and 'Beauty & Spas' - as we did on restaurants, and obtain the following results.

Business	**All Features**	**Topics + Review Features**	**Reviewer Features (including network)**
Restaurants	**0.693**	0.647	0.671
Home Services	**0.686**	0.643	0.648
Beauty & Spas	**0.685**	0.651	0.63

Table 3: Best classification accuracies by feature set for 3 business types

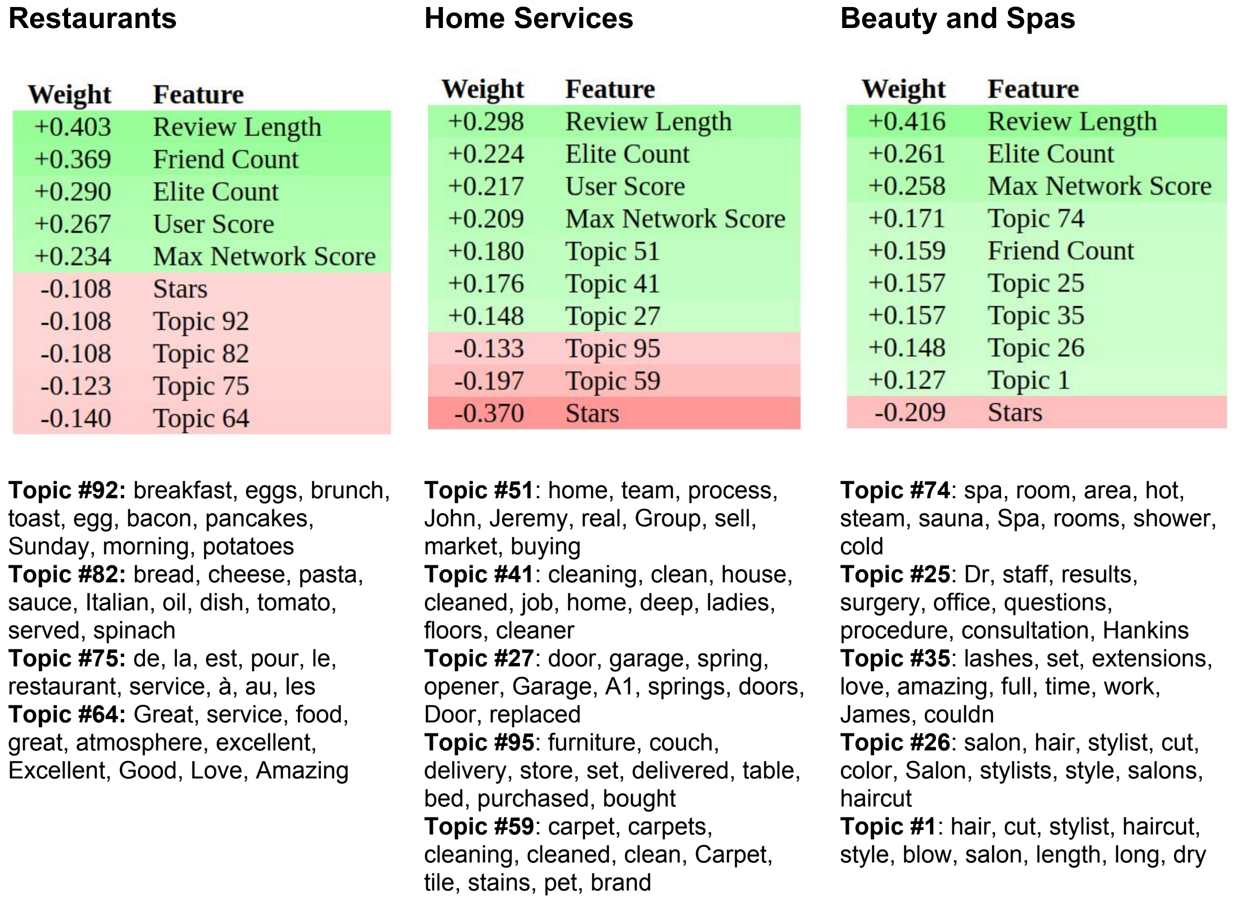
Business	**Model**	**All Features**		**Topics + Review Features**		**Reviewer Features (including network)**	
		R^2 Train	**R^2 Test**	**R^2 Train**	**R^2 Test**	**R^2 Train**	**R^2 Test**
Restaurants	Tobit	0.365	**0.337**	0.117	0.116	0.342	0.312
Beauty & Spas	Tobit	0.236	**0.254**	0.112	0.119	0.162	0.186

Home Services	Tobit	0.207	**0.148**	0.086	0.083	0.141	0.076

Table 4: Regression R-squared values of best-performing model with all feature types

In case of 'Beauty & Spas' and 'Home Services', we notice that using either reviewer features or review features in isolation did not work very well. Combining these two sets of features significantly improved performance. However, for the case of ‘Restaurants’, using only reviewer features worked very well. Adding content and review-based features improved performance to a lesser extent, when compared to the other two business types. These results are consistent for both regression and classification models. This can be explained intuitively based on the type of business. Intuitively, we think that this could be because in the case of restaurants, the reviewer and the network influence the users to read the review and vote for them. For other business types such as ‘Home Services’, a user may not be as interested in reading the reviews by friends. Our hypothesis is that the social aspect of voting is much more significant in the case of restaurants - a user posting a review about where they went out for dinner or drinks is potentially a more social activity than a user posting about a carpet cleaning service, for example.

In order for an online review community to be effective to both users and businesses, it is important to understand what constitutes a popular review as perceived by people. After understanding that a combination of review and reviewer features are good predictors of social endorsement, we want to understand if the same features are important across different business types. After analyzing the top 10 statistically significant features from Tobit Regression results (Figure 4) of Restaurants, Home Services, and Beauty and Spas, it was interesting to find that the same features (apart from content features) played an important role in gaining social endorsement.



Restaurants

Weight	Feature
+0.403	Review Length
+0.369	Friend Count
+0.290	Elite Count
+0.267	User Score
+0.234	Max Network Score
-0.108	Stars
-0.108	Topic 92
-0.108	Topic 82
-0.123	Topic 75
-0.140	Topic 64

Topic #92: breakfast, eggs, brunch, toast, egg, bacon, pancakes, Sunday, morning, potatoes
Topic #82: bread, cheese, pasta, sauce, Italian, oil, dish, tomato, served, spinach
Topic #75: de, la, est, pour, le, restaurant, service, à, au, les
Topic #64: Great, service, food, great, atmosphere, excellent, Excellent, Good, Love, Amazing

Home Services

Weight	Feature
+0.298	Review Length
+0.224	Elite Count
+0.217	User Score
+0.209	Max Network Score
+0.180	Topic 51
+0.176	Topic 41
+0.148	Topic 27
-0.133	Topic 95
-0.197	Topic 59
-0.370	Stars

Topic #51: home, team, process, John, Jeremy, real, Group, sell, market, buying
Topic #41: cleaning, clean, house, cleaned, job, home, deep, ladies, floors, cleaner
Topic #27: door, garage, spring, opener, Garage, A1, springs, doors, Door, replaced
Topic #95: furniture, couch, delivery, store, set, delivered, table, bed, purchased, bought
Topic #59: carpet, carpets, cleaning, cleaned, clean, Carpet, tile, stains, pet, brand

Beauty and Spas

Weight	Feature
+0.416	Review Length
+0.261	Elite Count
+0.258	Max Network Score
+0.171	Topic 74
+0.159	Friend Count
+0.157	Topic 25
+0.157	Topic 35
+0.148	Topic 26
+0.127	Topic 1
-0.209	Stars

Topic #74: spa, room, area, hot, steam, sauna, Spa, rooms, shower, cold
Topic #25: Dr, staff, results, surgery, office, questions, procedure, consultation, Hankins
Topic #35: lashes, set, extensions, love, amazing, full, time, work, James, couldn
Topic #26: salon, hair, stylist, cut, color, Salon, stylists, style, salons, haircut
Topic #1: hair, cut, stylist, haircut, style, blow, salon, length, long, dry

Figure 4: Feature importance comparison across Business types

Firstly, in terms of review features, review length plays a critical role in all three business types. Our own intuition was that people might prefer shorter and concise reviews, but the high positive coefficient for review length indicates that longer reviews have a higher chance of being popular in the community, which is also in line with the results of Bakhshi et al. (2015). This may be explained by the hypothesis that longer reviews are likely to contain more information about the business, and such reviews are likely written by more dedicated reviewers. Second, stars play an important role in determining the review's level of endorsement, and have a negative coefficient. It is kind of intuitive that users may vote more often on critical reviews, which could be useful in making a decision, though this is opposite to the result obtained by Bakhshi et al. (2015). Another interesting point to note about the 'Stars' feature is that it is more negative for

‘Beauty & Spas’ and especially ‘Home Services’, as compared to restaurants, indicating that people vote on critical reviews more often for these two business types.

Since we had different topics in the three types of businesses, we can’t directly compare them. However, we do see that for ‘Home Services’, the topics quite clearly align with specific kinds of home services (such as ‘real estate’, ‘furniture’, ‘carpet cleaning’ etc.). The coefficients for these topics indicate that reviews for certain kinds of home service gain more votes than for other kinds. We don’t see a clear pattern like this for the other two business types.

In terms of reviewer features, user score and the number of times a user has been an elite member have a positive significant relationship in gaining endorsement. The first probably reflects the fact that users who have received votes in the past are likely to continue doing so. The significance of ‘Elite’ could result from multiple reasons - Yelp has a feature to sort reviews by Elite, and also displays the Elite status of a user next to their reviews. We can explain the positive relationship between past history and votes received through the knowledge and experience an active user gains from being active in the Yelp community: active users have the advantage of gaining knowledge about the community and its interests. Such knowledge may reflect in the quality of reviews experienced members write. In terms of reviewer network features, reviews of users who have many friends, and also friends with a high user score (Maximum Network Score) stand a higher chance of gaining more votes in the community.

The reviewer and reviewer’s network features indicate that popular, active and regular members and users having a popular network are the highest contributors to popular reviews. It is also interesting to see a similar pattern among three different types of businesses. This can be attributed to Yelp’s system in making reviews of popular users and friends more visible to a user, as mentioned before.

7. Discussion of Outcomes, Implications, and Conclusion

While we have done a great deal of analysis on our results in the previous section, there are clearly some points that merit more discussion. In addition, it is worth outlining what future work remains to be done, what limitations we have with our study of the problem, and what content from class discussions relates to the problem at hand.

The first research question has been analyzed quite thoroughly. As we saw, the final accuracy and R^2 scores of models were not as high as the team was hoping they would be, but a surprising number of insights came from looking at feature importance. With this in mind, we look at each set of features by going through the remaining research questions.

The first set of features that we analyze are review-based features, which were not as useful as the team was hoping, but certainly taught us something about how the design of the platform plays a major role in shaping user engagement. As discussed during the class lectures, the design of a social platform determines, to a large extent, who joins the platform, who stays and how engaged they are. From Whyte's study on how open spaces influence social interactions to the research conducted by Bakshy et al. (2011) on how information diffuses through Twitter via retweets, we have often pondered over how the design features of a platform is closely knit with the type of interactions it aims to foster.

Through our study, we observed that models with only review-based features did not perform as well as those with reviewer and review features, and that these features were few when it came to the list of most important features. As mentioned before, the team has the idea that this may be due to the design of Yelp. We know Yelp is flooded with far more reviews than an average user can manage and no user would scroll through the hundreds of reviews that pertain to any one business. Rather, many users would use the sort option that Yelp provides, which sorts reviews by five things: 'Newest First', 'Oldest First', 'Highest Rated', 'Lowest Rated', and ‘Elite’. A user may be more likely to sort reviews by their date or look for reviews from those ‘elite’ users, as many would consider those users trustworthy. Taking this factor into account, our results look very reasonable. But considering Yelp’s sort filters, there is the possibility that reviewer features work well.

Reviewer features were the most important features when it came to both classification and regression and was the set of features that did well individually. Thus, having other people vote on your review is also a function of who you know and your reputation. This has important implications in that it reveals how Yelp can be viewed as a social network where not only do you share reviews with others, but there is a sort of social hierarchy and online reputation associated with the platform.

Finally, we learned that not only are certain features more important than others, but it is the case that the same features are generally the most important ones across types of businesses. The results agree with intuitions like people tend to vote for critical reviews, review length matters, different topics matter for different businesses, and popular reviewers generally get more votes. Again, we can draw more comparisons between Yelp and other social networks in this way, even if such comparisons may not be intuitive at first glance.

This problem is hard to solve, and the team’s approach yielded promising, but not fantastic, results. As this is understandably due to our approach, perhaps there is a feature or model that we did not think of and we limited ourselves to just text-based features, so more things could be incorporated, like photos or the nuances of the platform (for example, how behaviour varies across Yelp on mobile vs web). The results we garnered give us hope that the problem may not be entirely unsolvable, and we hope that one day, that is certainly the case.

8. List of Teammate Names (0.5 pg)

This project was done by Mukundan Kuthalam, Jayant Jain, Nikhita Karnati, and Asra Yousuf.

References

1. Archak, N., Ghose, A., and Ipeirotis, P. G. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (2007), 56–65.

2. Bakhshi, S., Kanuparthy, P., and Gilbert, E. Demographics, weather and online reviews: a study of restaurant recommendations. In Proceedings of the 23rd international conference on World wide web (2014), 443–454.

3. Bakhshi, S., Kanuparthy, P., and Shamma. D., 2015. Understanding Online Reviews: Funny, Cool or Useful?. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15). ACM, New York, NY, USA, 1270-1276. DOI: https://doi.org/10.1145/2675133.2675275

4. Lu, Y., Zhai, C., and Sundaresan, N. Rated aspect summarization of short comments. In Proceedings of the 18th international conference on World wide web (2009), 131–140.

5. Ghenai, A. (2014). What makes a review useful , funny or cool on Yelp.

6. Mudambi, S.M., & Schuff, D. (2010). What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. MIS Quarterly, 34, 185-200.

7. Fang, H., Cheng, H., & Ostendorf, M. (2016). Learning Latent Local Conversation Modes for Predicting Community Endorsement in Online Discussions. arXiv e-prints , arXiv:1608.04808.

8. Petrovic, S., Osborne, M., & Lavrenko, V. (2011). RT to Win! Predicting Message Propagation in Twitter. ICWSM.

9. Zaman, Tauhid R. and Herbrich, Ralf and Van Gael, Jurgen and Stern, David. (2010). Predicting Information Spreading in Twitter. Computational Social Science and the Wisdom of Crowds Workshop (colocated with NIPS 2010)

10. Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., and Kustarev, A. 2012. Prediction of retweet cascade size over time. In Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12). ACM, New York, NY, USA, 2335-2338. DOI: https://doi.org/10.1145/2396761.2398634

11. Yu, Haihao & Feng Bai, Xu & Huang, ChengZhe & Qi, Haoliang. (2015). Prediction algorithm for users Retweet Times. 9-13. 10.14257/astl.2015.83.03.

12. Chen, Wei-Fan & Chen, Yi-Pei & Ku, Lun-Wei. (2017). How to Get Endorsements? Predicting Facebook Likes Using Post Content and User Engagement. 190-202. 10.1007/978-3-319-58484-3_15.

13. Kennedy, P. 1994. A Guide to Econometrics (3rd ed.), Oxford, England: Blackwell Publishers.

14. Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone’s an influencer. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining - WSDM 11. doi:10.1145/1935826.1935845