Analysis on Yelp Friend Network

Chuang Ke and Minghao Zhu

September 29, 2023

Analysis on Yelp Friend Network

Chuang Ke
Electrical and Computer Engineering
Carnegie Mellon University Silicon Valley
Moffet Field, California
chuangk@andrew.cmu.edu

Minghao Zhu
Electrical and Computer Engineering
Carnegie Mellon University Silicon Valley
Moffet Field, California
minghaoz@andrew.cmu.edu

***Abstract*—Social networking functions like word of mouth when it comes to reviews and ratings. Given the data of users and reviews from the Yelp dataset, we first build a network of friends to evaluate the influence of social networking on user reviews and stars, and then another network of reviews on businesses to try to identify characteristics of successful businesses in their corresponding network of user reviews. For milestone 2, we percept the influence of friend relations among users given the context. Meanwhile, we rebuild the second network mentioned above by taking temporal information into account. We mined the new network and extracted feature embeddings for nodes before combining them with common network metrics and building the final feature vectors for the new networks of businesses. Finally, we trained several classifiers for the task of determining whether a business is successful or not and made evaluations of the results.**

Keywords—Yelp data analysis, social networking, online user behavior, natural language processing, network feature mining.

I. Introduction

The services-based industry evolves rapidly with the help of modern technology. Local online review apps like Yelp, TripAdvisor, and Facebook business reviews, in particular, have played an important role in manipulating user behavior. Yelp, our main consideration in this project, is reported to have 45% of all customers checking business reviews on the app before actually visiting it while 35% of searches on Yelp lead to a visit within 24 hours[1]. In this project, we focused on the friend network of Yelp that is driven by home feed updates and direct review sharing. By constructing and analyzing this friend network, we hope to find positive feedback forwarding in successful businesses and extend our findings to predict potential ones.

II. Prior Work

Yelp provides a rich dataset[2] that drives a lot of research on user behavior, social network analysis, and predictions. However, only a few shade lights on friend influence, and their results turned out to be rather discouraging.

Our fundamental assumption is that friends on Yelp do have an impact on user behavior. Statistical analysis was performed on user ratings and showed that the user's rating is unaffected by that of his/her friends by correlation and regression techniques[3]. While replicating the same result on this correlation, we think reviews are the more important factor in user behavior due to their text nature and contain more information and dimensions compared to just a rating between 1 to 5.

In terms of how to define a business as successful, Feng, Kitade, and Ritter suggested that a business with more than 37 reviews and average ratings over 3.5 can be classified as successful[6]. As we investigate the distribution of ratings for businesses with more than 200 reviews, we find 4.0 a better dividing point and this observation aligns with our common mindset towards a Yelp rating.

Finally, the prediction of successful businesses in terms of ratings has been done through data mining and machine learning algorithms. The combination of text (review) features and non-text features contributes to the successful prediction of stars, but including the stars of reviews as a feature undermines the results due to its natural connection and large information ratio[4]. Using simple text features extraction methods such as unigram and bigram may result in rather poor performance on the validation set[5]. Here we propose to first identify the feedback forwarding network among friends within a certain business before making any predictions instead of predicting using all reviews.

To conclude our literature reviews, there exist few research papers on building a friendship-review network for Yelp dataset analysis and basic n-gram models might not be enough to represent the textual aspect of the review dataset. Thus the combination of network metrics and textual embeddings could be a better feature extraction of a business before trying to predict its success.

III. Approach

We used the Yelp public dataset[2] which contains a large collection of businesses, users, and reviews in separate JSON files. The *yelp_academic_dataset_business.json* and *yelp_academic_dataset_user.json* consist of details of businesses and users indexed by *business_id* and *user_id*, while *yelp_academic_dataset_review.json* is our main concern here that has both a *business_id* and a *user_id* and features a review text and a rating. Each user data point has a list of *user_id*s of his/her friends, enabling us to investigate the influence of these connections.

The approaches we take are different in terms of network building and analysis in the two tasks we proposed while the similarity computation shared similar methods.

A. Task 1: Evaluate the influence of friend relationships on online user behavior

In task 1, we first build the network with users as nodes and whether friends with each other as edges/connections. The resulting network is hard to analyze directly due to the number of nodes (~2M). So we first perform random downsampling among users from around 2 million to 20 thousand while keeping the remaining edges and filtering out users with less than 5 pieces of reviews.

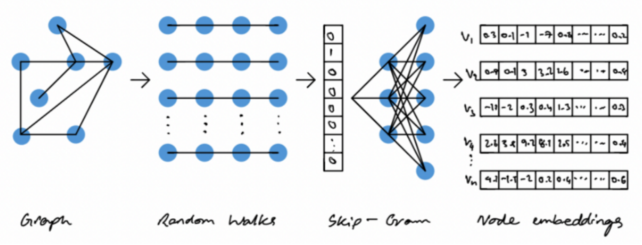
For the similarity of user reviews (text similarity for now), we first concatenate all reviews of a user and treat the result as a document. We adopt the Doc2Vec or Paragraph Vector model introduced by Le, Q(2014) for computing fixed-length vector representations before calculating the cosine distance between these vectors. With this calculation, we can then gather the similarity between a user and his/her friend or a user and other non-friend users. We intuitively evaluated the pre-trained Doc2Vec model on the "text8" dataset[9] as well as another model trained on our own corpus.

We evaluate these two models by comparing the average similarity between a user and his/her friends and the average similarity between the same user and sampled non-neighbors in the network (strangers) to form a distribution of the differences. Both results are shown in the result section.

B. *Task 2: Identify characteristics of a successful business by network analysis and predictions*

In task 2, we are building networks of Yelp reviews of businesses. For each network, we use all the reviews of the corresponding business as the nodes and we connected the two nodes with an edge if the reviewers of the two reviews are friends with each other. The network is directed and weighted, and our approach is using the similarity between the two reviews as the weight while all edges are pointing from the older reviews to the newer ones. A time gap factor is also introduced considering that the influence of a review is undermined by its "age": the longer the time gap between two reviews we have, the smaller weight should be set to that edge, the detailed settings can be found in the following section.

After constructing the networks, we assume that there are patterns that existed in Yelp reviews (and how it affects their friends) for successful businesses. Thus, we would like to identify the key factors of the business's success through network feature mining and network analysis. To be more precise, we hope that we can build a relationship between the fact that certain businesses have higher stars and are successful, the features extracted by network mining models, and key attributes of their network such as the degree distribution, the number of connected components, and the number of edges that have higher weights.



For network mining and representations, we would utilize the Node2Vec[12] model that provides node embeddings for networks with consideration of edge directions and edge weights by performing random walks. The fact that these representations from Node2Vec are scalable enables us to build fixed-length feature vectors as shown in the process above.

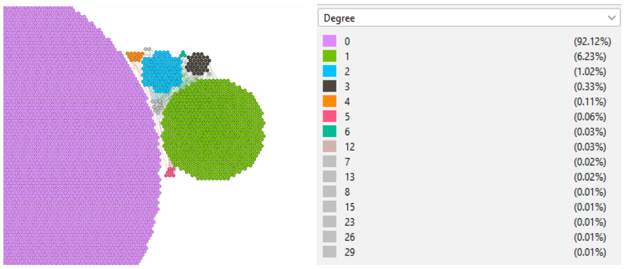
Finally, we will feed our constructed feature-label combination to popular machine-learning models for the classification task to see if these features can represent the characteristics of the defined successful businesses.

IV. Experimental Setup and Results

A. *Task 1: Evaluate the influence of friend relationships on online user behavior*

1) User network and downsampling

The current downsampling is performed randomly among users, resulting in a sparse network of users as shown below.

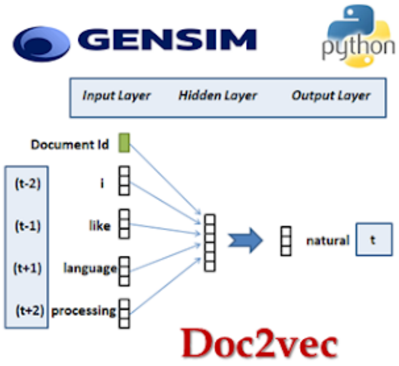


Degree

🟪	0	(92.12%)
🟩	1	(6.23%)
🟦	2	(1.02%)
⬛	3	(0.33%)
🟧	4	(0.11%)
🟥	5	(0.06%)
🟩	6	(0.03%)
▨	12	(0.03%)
▨	7	(0.02%)
▨	13	(0.02%)
▨	8	(0.01%)
▨	15	(0.01%)
▨	23	(0.01%)
▨	26	(0.01%)
▨	29	(0.01%)

2) Context corpus building and Doc2Vec training

Doc2Vec[7] is a modified version of Word2Vec[9] that is more robust and specializes in identifying similar paragraphs which better fits our needs for vectorizing and computing review similarities.



We treat each piece of review as a single document and first preprocess the text by eliminating punctuation, changing it to lowercase, and removing stopwords. For Milestone 2, we failed to train the model with the whole corpus (~7 million) due to the limitation of memory and used 50% of the overall corpus for the training of our final Doc2Vec model.

3) Similarity computation (text)

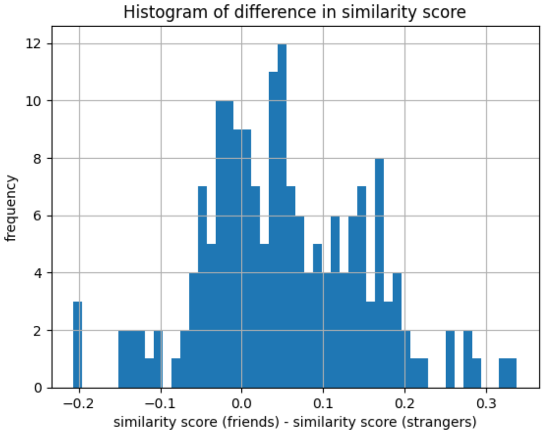
We denote the Doc2Vec model as M, the concatenated reviews of two users S_0 and S_1, and the similarity score is computed as:

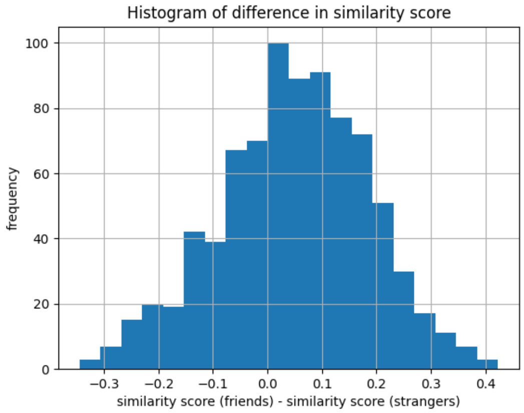
$$Score = 1 - CosineDistance(M(S_0), M(S_1))$$

Therefore, we have the difference of similarity scores:

$$Diff = avg(Score_{Friends}) - avg(Score_{Strangers})$$

The resulting histogram of difference is shown below, the first histogram is the results of the model trained on text8 dataset, the second one is on the corpus we build.

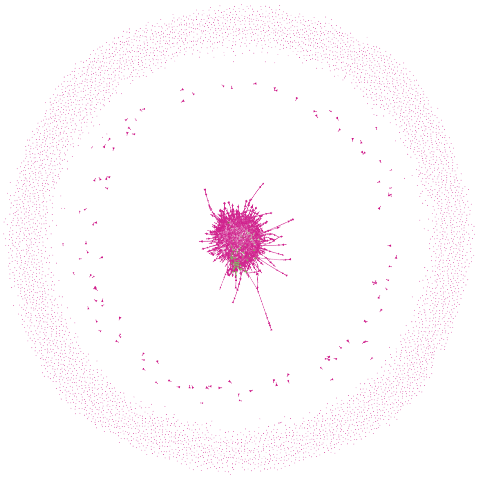




Notice that a positive difference means that a user and his/her friends tend to have similar reviews. In contrast, a negative value indicates the opposite. From the above graph, we can conclude that friends do have similar reviews in terms of text meaning. Moreover, given the context of all Yelp reviews in the dataset, the Doc2Vec model performs better than the pre-trained one and thus we decide to proceed with the later one for our later task.

B. *Task 2: Identify characteristics of a successful business by network analysis and predictions*

As an example, we first pick Acme Oyster House as an example of a successful business. It has an overall rating of 4.0 and 7674 related reviews. We build the network based on the reviews and connect the edges where the reviewers are friends. For each edge, we calculate the weight based on the similarity between the two reviews using the NLP method similar to task 1. Among 6001 edges, 5898 edges weight 0.5, and 1927 edges weight 0.8. Besides, for the 7674 nodes that we have for this network, over one-third of the nodes are connected to at least one other node, while 136 is the largest amount of nodes that one single node is directly connected with. Below shows the visualization of the network in Gephi[10] with a ForceAtlas 2 layout colored by the ranking of degrees.

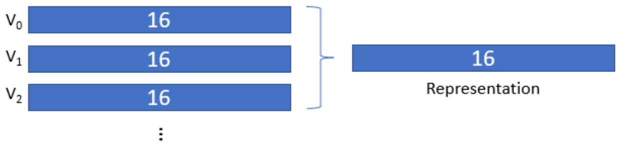


Furthermore, we constructed over 1300 networks like the one above based on all businesses that have more than 500 reviews. Similar to the network we built in milestone 1, we use reviews for each business as the node, and we connect two nodes with an edge if the reviewers are friends. While the weights are the similarity between the two reviews, we introduced the time difference between the two reviews as a factor for the weight. To be more precise, if the time gap between the two reviews is over one year, we would multiply the weight by 0.5 as it is less likely that the older review has an impact on the newer one. However, the factor would vary from 0.5 to 1 based on the time gap between the two reviews, where a smaller time gap would result in a larger factor as it is more likely to have an impact.

$$factor = \begin{cases} 1.0 - 0.5 \times \dfrac{gap}{12} & if\ time\ gap < 12\ months \\ 0.5 & otherwise \end{cases}$$

$$w_{u,v} = similarity \times factor$$

With the networks we build, we now train Node2Vec on each of them before fetching the node vectors from it. Given each node has a unique feature vector of the same length but the number of nodes inside each network varies, we decide to take the average of these node vectors for the feature representation of a network. Notice that the choice of vector size here is based on the observed sizes of the networks and online materials, we might not further discuss the impact of different vector sizes given the scope of this project.

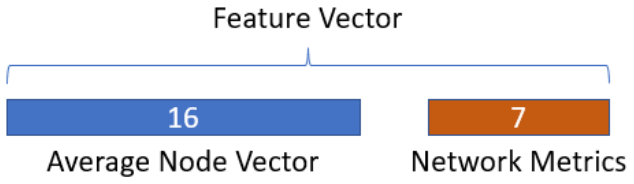


Besides, we collected seven important attributes/metrics of each network, the choice of metric is shown in the chart below. Notice that the choices here are not fully discussed and await further analysis.

TABLE I. Network Analysis Metrics

Metric	Dimension
Number of Nodes	1
Number of Edges	1
Average Degree	1
Average Edge Weight	1
Average Clustering Coefficient	1
Number of Weakly-connected Components	1
Size of the Largest Weakly-connected Component	1

Together with the vectors we collected using the Node2Vec method, we end up with the final feature vector (as shown below) of length twenty-three for the training of our classification task.



To identify the successful ones out of all businesses, we picked the rating of the business as the indicator. For each business, we give it a label of one if its rating is larger than or equal to four out of a five-star scale. Otherwise, we give the business a label of zero if its rating is smaller than four.

$$label = \begin{cases} 1 & if\ stars \geq 4.0 \\ 0 & \quad otherwise \end{cases}$$

We utilized seven machine learning models from the Scikit-Learn libraries, which are Decision Tree, Random Forest, and three SVC models with Sigmoid, RBF, Poly kernel, KNN, and RNN. For each model, we randomly selected fifteen percent of the data as the testing dataset, and the rest are used for the training set. To evaluate our models, we choose accuracy and f1-score to take both precision and recall rate into consideration. The training result for each model is shown in the table below.

TABLE II. Training results

Acc Rank	Model Performance		
	Scikit-Learn model name	***Accuracy***	***F1-Score***
1	SVC with RBF Kernel	0.82	0.90
2	SVC with Poly Kernel	0.82	0.90
3	RNN	0.82	0.90
4	Random Forest	0.81	0.89
5	SVC with Sigmoid Kernel	0.73	0.85
6	KNN	0.72	0.86
7	Decision Tree	0.67	0.79

V. Conclusion and Short-Term Plans

To conclude, we developed two tasks based on the Yelp Dataset to identify the characteristic of a successful business. In milestone 1, we are more focused on task 1, where we built a network of users, and verified that friends do have similar behavior in rating and text comments. For task 2, we started by building the network with one chosen business, and we performed an initial analysis. For milestone 2, we build more than 1300 networks for all businesses that have more than five hundred reviews and build the new network by introducing a new way to calculate the similarities as the weight and the time difference between two reviews as a new factor applied to all the weight.

We achieved an accuracy from 67 percent to 82 percent among the seven machine learning models, and we hope we can still improve them in milestone 3 by performing the following improvement. First, we are going to build more networks to provide more data for training, thus we would build networks for all businesses that have more than two hundred reviews. By doing that, we would have more than 6000 data points, which is about five times more than what we have for milestone 2. Plus, we would improve the Node2Vec vectors as well as the network analysis matrix as we identify more attributes that may have an impact on the result by, for example, investigating gain ratios for each feature dimension. In addition, we are having an unbalanced dataset since there are about two times more data that we defined as a successful business than the one we defined as not that successful. Thus, we would be exploring ways to rebalance our dataset. Furthermore, we will work on hyper-tuning the models and may introduce a new network if we found useful.

For division of labor, Chuang focuses mainly on Doc2vec, Node2vec models, and data preparation for the node vectors, while Minghao focuses on network analysis, introducing time gaps to the new network, and machine learning model training and evaluation.

References

[1] Marinova, I. (2022). *25+ Groundbreaking Yelp Statistics to Make 2022 Count*. Review42. https://review42.com/resources/yelp-statistics/

[2] Yelp Dataset. (2021). https://www.yelp.com/dataset/

[3] Gupta, S., Desai, V. and Thakkar, H. (2017). *Social Data Analysis: A Study on Friend Rating Influence.* Lighting Talk at WWC Connect India 2017. https://arxiv.org/abs/1702.07651

[4] Y, Chen., F, Xia. (2020). *Restaurants' Rating Prediction Using Yelp Dataset*. 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), 2020, pp. 113-117, doi: 10.1109/AEECA49918.2020.9213704.

[5] Asghar, N. (2016). *Yelp Dataset Challenge: Review Rating Prediction*. https://arxiv.org/abs/1605.05362

[6] Feng, J., Kitade, N., & Ritter, M. (2015). *Determining Restaurant Success or Failure.* https://www.cs.dartmouth.edu/~lorenzo/teaching/cs174/Archive/Winter2015/Projects/finals/fkr.pdf.

[7] Le, Q. Mikolov, T. (2014). *Distributed Representations of Sentences and Documents*. https://arxiv.org/abs/1405.4053

[8] Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations.

[9] Mahony, M. (2011). *Large Text Compression Benchmark*. http://mattmahoney.net/dc/textdata.html

[10] Bastian M., Heymann S., Jacomy M. (2009). *Gephi: an open source software for exploring and manipulating networks.* International AAAI Conference on Weblogs and Social Media.

[11] Grover, A., & Leskovec, J. (2016). *node2vec: Scalable Feature Learning for Networks.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.