Yelp Mining Proposal

Tyler Lugger Amogh Jahagirdar Elias Bezanis

1. PROBLEM STATEMENT

In this project, our groups motivation is to use the data collected by the review service Yelp to help restaurants that are being reviewed. Yelp has been around for a long time and has an enormous number of reviews for restaurants all across the globe. These can be created by anyone with access to the website and can help or hurt a restaurant. It has been thought that Yelp reviewers can sometimes hold restaurants hostage by holding the power to influence how others perceive a restaurant. Good reviews can help the popularity of restaurants by bringing in more people that want to try it for themselves.

We want to use Yelp data to give more power back to the restaurants being reviewed. By mining data about the users that frequently use the review service, we hope to create a predictive model of user behavior. This would allow us to find out several aspects about Yelp users and pass that information on to the restaurants themselves.

2. LITERATURE SURVEY

Since the Yelp Data set challenge, is a public challenge all across the world, there are many existing works. Furthermore, the competition is on its 9th round, so there is a plethora of existing work from the previous rounds. An analysis, focusing primarily on the winners of the previous rounds, their questions, as well as their approaches would certainly help guide the direction of this research.

One group of winners in the first round of the competition utilized the Latent Dirichlet Allocation algorithm in order to extract latent subtopics from Yelp restaurant reviews[1]. The LDA model is one which allows a set of observations to be described by "hidden" groups, where these groups explain why there is some similarity in the data. In this group's case, it would allows them to specifically isolate which features customers care about when they are about to rate a restaurant. In terms of pre-processing, this group isolated businesses which were only restaurants, and reviews only targeted at restaurants. Since LDA, relies on the concepts of topics, the text of a review had to be assigned into different subtopics. Essentially, this splitting of text and assigning into various categories is a form of Natural Language Processing. This group also used tools such as Python and other standard visualization/numerical analysis libraries. After, their analysis, they discovered that the keyword "service" in the review [1], was essentially the main factor for a given rating.

A group of winners in the fifth round describe a multi-instance classification technique that goes beyond classifying at the group level[2]. In other words, the primary focus for this group, is not necessarily the mining results, but the generation of a new technique that does more than simply take a given object and classify it into a group (for example positive or negative); it can use these group level classifiers to train what they call "instance-level classifiers"[2]. These instance level classifiers can extract specifics from a given object and then perfect the model. For example, a positive review can still have negative comments. Instance level classifiers would use the group classifier to first put the positive review in the right category, but then extract the negative comments, and further perfect the existing model(See [2] page 1). This can be particularly useful, to perfect a business. Even though many may find that a given business has high quality, analyzing keywords that point out small errors, can help this business elevate their customers' experience. They describe their model, by using cost functions and other relatively complex mathematical expressions. For pre-processing, they had a combination of the Yelp data set, as well as data from Amazon reviews. Then they performed natural language processing, in combination with a group classifier, to extract words into a basket of categories (including "Very Positive", "Positive", "Neutral", "Negative", and "Very Negative") (See [2] page 6). After rigorous analysis, they concluded that their technique was not only accurate on the real world data sets, but also were relatively scalable.

A group of winners in the seventh round also won the grand price for developing a new technique, which they called "Semantic Scan"[3]. Essentially, the goal of semantic scan is to provide quick detection and characterization of "emerging topics in text streams" (see [3] page 1). They argue that existing similar techniques have shortcomings for rapid detection of these emerging topics. They specifically mention that LDA, Latent Dirichlet Allocation is too slow for analyzing ”web scale text streams” (see [3] page 1). Primarily, semantic scan focuses on utilizing a ”contrastive LDA model with spatial scan statistics” (see [3] page 3), where topic modelling and assignment is fundamentally changed from different models, with a focus on rapid analysis as well as analysis with noisy text. To pre-process, they setup the Yelp data in such a way that the data simulates an on going business event, as their technique is focused on analysis of a text stream. For example, they simulate a surge of business for a given restaurant and set up a text stream of the reviews, and apply Semantic Scan on it, to see how rapidly and accurately it can assess the noisy text data(See [3] page 8). At the end, they concluded that their model was also effective, and provided statistical tests to further illustrate this.

It’s important to note that there were many more winners, and general participants in this competition [4]. Analysis of even more participants, would further aid in not only perfecting the research/evaluation process, but in gaining an understanding of why certain models work in specific cases.

3. PROPOSED WORK

Our team has developed a variety of strategies to process the Yelp data set in an innovative and efficient way. With the help of MongoDB we can begin processing the JSON formatted data without setting up that much of our own infrastructure. In our pre-processing phase we plan to use many different methods for a variety of effects. We are first going to reduce our data. The Yelp dataset isn’t massive, but for our purposes certain subsets such as the picture subset aren’t necessary for predicting user behavior. Then we will clean our data for restaurants only. Based on the restaurants we find, we are discussing integration of a weather dataset to cross-reference reviews and types of weather, because we believe there may be a correlation here. Finally, we need to transform the data such as text reviews or check-ins into a more manipulatable and standard type of result. To transform raw text into manipulatable data we have a program called ”Heavy Moose”[5] developed by a group member that can be reconfigured to parse text strings and count important values.

In terms of an end result, our goal is different than those of prior competitors. From the data we process we not only want to predict user behavior, but actively incentivize consumers and producers based on our predictions. This can be done in a variety of ways such as couponing before a customer goes to a restaurant, or letting a restaurant know that they may have a spike in customers today so they should get more staff. The whole purpose of Yelp is to make the service industry more effective. By bringing the consumer and producer closer together through predictive modelling, a type of communication that will be beneficial for both parties can take place. To evaluate the effectiveness of our result, we would need a few months to look at the ”newer” version of the Yelp dataset and verify our predictions.

4. DATA SET

Our data set comes from the Yelp data challenge. Yelp has selected a portion of their review data that anyone can use for academic purposes. This data set which is nearly five gigabytes in size, contains millions of tuples spread across five JSON files. These files contain all of the information we will need about Yelp users, reviews on businesses, and information on the businesses themselves.

Since this data set is part of a challenge, our team will have the opportunity to submit our data mining results to the challenge to win one of the many prizes that yelp gives to teams that work on their data set. If we are able to find anything useful from this data set, we have the potential to create real change in Yelp and how businesses use Yelp to maximize customer satisfaction.

5. EVALUATION METHODS

To evaluate the results of our data mining on the Yelp data set, our team plans to begin our mining on smaller portions of the data set. Since we are creating a predictive model, we can compare our predicted user behavior with actual user behavior that occurs later in our data set. Using this technique, our group will be preforming cross-validation on our predictions to evaluate how accurate our predictive model can get.

Our group can also perform validation on our test data set as it grows in size. We could see if our user behavior predictions grow in accuracy as the size of our data set being mined increases. This would tell us that our predictive model is accurate and would become more accurate with an increased amount of user data.

6. TOOLS

For our data mining project, our group plans to utilize several tools. We will use data mining software to help us with our initial data analysis. Using these, we hope to begin finding trends within the Yelp data and get an idea of what else we could look into that we may have not considered. We will use a NoSQL database to store our data such as MongoDB to store our data for easy access. While this is desired, it is not quite necessary and can be pushed back if needed. Python will be our main programming language for in depth data mining. This is an easy to use programming language that is great for data computation and analysis. Python also has plenty of useful packages that can be used to simplify analysis. Numpy, ScyPi, and Pandas provide functions to help with numerical computation. scikit-learn provides easy to use tools for data mining and analysis.

7. MILESTONES

There are 5 major milestones for our project, each with a few sub-goals:

1.Integration of datasets and tools - Once the Yelp dataset and weather dataset have been transferred into MongoDB they will be ready for manipulation. Also, the reconfiguring of the Heavy-Moose program will make text manipulatable

2.Pre-Processing of data- Here we apply a number of preprocessing techniques on the data with the aim of reducing it to only that which predicts consumer and producer behavior. The methods for this are as prescribed in the Proposed Work section.

3.Analysis of user behavior- After the data has been processed we can now perform a number of functions on it to determine the user behaviors that we hypothesized to see their frequencies and patterns.

4.Compilation of data- After we have identified frequencies and patterns they must be reduced even more into generalities that are presentable to Yelp, along with strategies we develop for using this compiled data.

5. Presentation of data- our work must be condensed into presentation form for class, and into a form acceptable for Yelp. In our presentation we will show to what extent and how we can predict user behavior, and will describe the techniques we used to get those final answers.

8. SUMMARY OF PEER REVIEW SESSION

During our peer review session, our group received some great feedback on our project. We learned that there would be another group working with this Yelp data as well but they would be mining different information from it. It would still be good to communicate with this group throughout the project as they will still be working with the exact same data. If they are using any analysis or mining methods that are different than ours, it would be a good idea to compare and see if we could get assistance from them. And we would of course be glad to share any information that we find with their team as well.

Our group received a comment about our plan to incorporate weather data into our mining to see if reviews are affected by sources that restaurants can't control. It seems like this is something that could produce new and useful results. Our team plans to look into this more and secure a source for our weather data.

9. REFERENCES

[1]Huang, James, Stephanie Rogers, and Eunkwang Joo. ”Improving Restaurants by Extracting Subtopics from Yelp Reviews.” (2013): 1-5. Web.

[2]Kotzias, Dimitrios, Misha Denil, Nando De Freitas, and Padhraic Smyth. From Group to Individual Labels Using Deep Features. N.p., n.d. Web. <http://mdenil.com/media/papers/2015-deep-multi-instance-learning.pdf>.

[3]Maurya, Abhinav, Kenton Murray, Yandong Liu, Chris Dyer, William W. Cohen, and Daniel B. Neill. Semantic Scan: Detecting Subtle, Spatially Localized Events in Text Streams. N.p., n.d. Web. <https://arxiv.org/pdf/1602.04393.pdf>.

[4]https://www.yelp.com/dataset_challenge

[5]Elias Bezanis, Oliver Hanna, Yang Wang
https://github.com/omnific-h/HeavyMoose