Yelp Review Summarization of Health Care Services

Amish Shah
Graduate Student
School of informatics and computing
Indiana University, Bloomington
shah7@indiana.edu

Dakshi Kumar
Graduate Student
School of informatics and computing
Indiana University, Bloomington
dakumar@indiana.edu

ABSTRACT
Yelp users provide reviews for local businesses and star ratings. However, it fails to explain why the businesses are good or bad. Through this work, we try to extract phrases from reviews that display valuable comments for user or business owners. Our work revolves around giving constructive opinion of public on health care services in U.S. cities to health monitoring and planning committees. But the fundamentals of the model, particularly, feature extraction is not limited to our problem and it can be tweaked to incline with a wide variety of problems that can be solved. This model can be a great add-on to the Yelp review system.

1. INTRODUCTION
When one wants to seek information for some local business or hospitals or health care services in a city on search engines, it comes up with numerous links. Yelp provides reviews confined to businesses and services. On Yelp, user can search for the hospitals and health care services in a city. It gives a page full of hospitals and care centres which can be sorted with the maximum star rating of the list. After visiting the yelp profile of the business user can go through its reviews provided by the people who have already visited that place.

To know about a place through the reviews of the users, one must go through all the reviews and know what is good and bad in that business to make any judgements to visit that business.

For instance, a patient just want to know if a dermatologist in the hospital or care centre is good enough. In that case, user has to go through all the reviews and make a decision. It is not only time consuming but also needs concentrated efforts.

Another example, if someone is seeking to have a new business in the city or how to know about what is already lacking the business? If one conducts an analysis of the Yelp reviews, it does not solve a new problem but create confusing options.

The constitutional problem for all the cases above is that the user cannot efficiently find attributional parts of the reviews that express more and state less. For our work, we are viewing Yelp from the standpoint of health planning and monitoring committees. The review data on Yelp provides both the star rating and the text for each review and are particularly targeted towards products and businesses, through our project we try to know if it would be possible to shed light on the following using sentiment analysis and machine learning models on business (health and medical) attributes and their corresponding reviews:

- The attributes that makes health care centre famous
- The reviewers' most likeable quality about the health care services
- The comfort-ness and quality in regards to equipment and supplies offered by health care services
- Departments of a city's healthcare service centres that need improvements

One potential way to try to derive to these questions are extracting phrases that reveal good and bad aspects of hospitals across cities that would help communities to build various models and plans.

2. RELATED WORK
Gaining knowledge from reviews using sentiment analysis has been an active field of research. In 2012, Chahuneau et al. [3] worked on predicting polarity of each review by training a logistic regression model. But it fails to determine why a review given was positive or negative. It lacked qualitative feedback. In [1] Peter D. Turney, did excellent sentiment analysis where he came out with the formula to take out the average of the ratings for the business type that provided satiable accuracy. In [4], Farhan describes how well a restaurant performs based on restaurant attributes by using a linear regression model. They describe what features a good restaurant has. But the attributes are limited to Yelp filters, we try to go beyond that and take into consideration reviewer's perspective that can have many dimensions more than Yelp filters. In [5], the authors try to summarize the text to aspect extraction and sentiment detection. With much work done on reviews of restaurants and local services, we plan to prepare an accurate model by doing sentiment analysis and classification of reviews that will help healthcare and urban planning communities. Freedman and Jurafsky[2] how price of food express reviews of a customer. Through this project, we will explore how various health center features affect patients or customers. Hu and Liu [6] in their research conduct opinion mining on cohesive set of products, we try to span a variably bonded heterogeneous factions of health care which can help reinforce public health and vitality.

3. METHOD

3.1 Filtering Data
The dataset for the project was collected from Yelp Dataset Challenge [8]. This JSON data is provided by Yelp to reduce the efforts for data collection and providing a great platform for innovative research. It is a plethora of data comprising of 2.7 million reviews by 687,000 users for a total of 86000 local businesses. These businesses are located in many cities, however for this work we consider cities in the United States - Charlotte, Las Vegas, Madison, Phoenix, Pittsburgh, Urbana.

For this project, we would like to collect expressive phrases from reviews that fall into business category - “Health & Medical” For convenient analysis, we break down this category into two:

1. Hospital
2. Care Centres

Following, is the format for our features of interest from the dataset[8]:

1. business:

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

2. review:

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

Initially, we collect the business data, particularly, “business_id” for the business type “Health & Medical”. This collected data is sub-divided into two categories as mentioned before: (1) Hospital, that includes business type: “Hospital” (2) Care Centres, which is a collection of business types: "Assisted Living Facilities", "Counseling & Mental Health", "Habilitative Services","Home Health Care", "Medical Centers", "Rehabilitation Center","Weight Loss Centers", "Child Care & Day Care", "Urgent Care" and "Skin Care".

Once a list of “business_id” is compiled in the above step, we use them to collect reviews pertaining to those businesses using the review object. After filtering, we are left with 1260 reviews for Hospitals and 4516 reviews for Care Centres across 6 U.S cities.

3.2 Phrase Extraction

The goal of this project is to extract the subjective and qualitative evaluation of the hospitals and care centres by the reviewers who have visited these centres. There has been several studies that have tried to extract phrases from reviews to get some feedback on products [9] and [1] But these extract phrases provided in the combination of Parts-Of-Speech tags:

- JJ|NN
- JJ|NNS
- RB/RBR/RBS|JJ|
- RB/RBR/RBS|JJ||NN
- RB/RBR/RBS|JJ/RB/RBR/RBS|NN/NNS
- RB/RBR/RBS|VBN/VBD
- RB/RBR/RBS|RB/RBR/RBS|JJ
- VBN/VBD|NN/NNS
- VBN/VBD|RB/RBR/RBS

In our study, we want to avoid this large number of combination because eventually we will end up evaluating the whole sentence rather than expressive phrases/bigrams. Hence, we use “NN|JJ” “NNS|JJ”, “JJ|RB” and “JJ|RBR” patterns and it was observed that we still achieve good accuracy. At the same time, we observed that there is a trade-off between the quantity of phrases and the quality of phrases developed. An ideal model should provide less phrases (quantity) that highly express views (quality) of the customers.

To retrieve the phrases based on the sequence of Parts-Of-Speech (POS) tags, there is a process called Chunking[10] We followed the process of chunking with regular expressions. So, the phrases to be extracted are represented in the form of tag patterns. ”JJ| NN” is an example of a tag pattern which states that if the review contains a phrase with an adjective followed by an noun. This facility is provided by Chunker package in NLTK where the tag patterns are regular expressions.

The NLTK Chunker is very useful to fetch phrases from each of the reviews. We had to go through the reviews and experiment with the chunk POS tag format. For this work, we have kept the tag patterns similar for extraction of positive and negative sentiments from a review.

3.3 Prior Polarity Formula

In sentiment analysis, it is a prevalent trend to make use of polarity values of words from out-of-context that is called as the prior polarity value[11]. There exists packages like SentiWordNet and Pattern that gives us polarity values of each word. In our work, we have used the Pattern package to calculate the polarity values of a phrase. Pattern package includes a function which returns the polarity value between +1 and -1 for a given word. Unlike Pattern, SentiWord does not give a single polarity value. It provides a positivity and a negativity score for each Parts-Of-Speech of a word. This would make deriving polarities for phrases more complicated, hence we opted Pattern package which gave us good results. For a phrase, Pattern [12] calculates an average score of all the words. As done by Peter in [1], for all the phrases in each review, we calculate the polarity values, sum them and divide by the total number of positive and negative phrases in each review [1]. A phrase is tagged as positive if polarity score is greater than or equal to 0.2 and negative if polarity score is less than or equal to -0.2. If the subjectivity value for each score is greater than 0.5. Subjectivity is kept high because reviews are subjective hence we want to consider phrases that are highly subjective. We ran the model for several trials and found that the model achieves good accuracy given that we have set strict values. Table 1. shows the selected phrases and their respective polarity values for review - *"Today is the first time I've gone here and even though I was feeling really bad it was a pleasant experience.*

Everyone there was genuinely nice and helpful. I can't say enough about the staff. The doctor was very knowledgeable and just as nice. He had an X-ray taken and had it read and called me within 15 minutes after I left to tell me the result.

Great place to go!"

Table 1. Phrases collected from a review with their polarity scores

Phrases	Polarity Score
Really bad	-0.699
Great place	0.8
pleasant experience	0.733
genuinely nice	0.6
Avg. score	0.358

3.4 Classify

The scores calculated for each review as mentioned in Section 3.3, we take the score for each review and check it against the corresponding rating. It is necessary to draw a relation between the review polarity score and star rating to affirm that phrases retrieved weigh to determine the ratings, and hence the stance of the user towards the local hospital or care centre. We achieved positive correlation with no exceptions and hence with the confidence achieved in the model we proceed to calculate the accuracy of the phrases retrieved.

While rating a business on Yelp, following are the guidelines for rating[13]:

- 1 Star : “Eeks! Methinks not”
- 2 Stars: Meh. I’ve experienced better.
- 3 Stars: A-OK
- 4 Stars: Yay! I’m a fan.
- 5 Stars: Woohoo! As good as it gets!

With these guidelines and the work in [1] on Epinion reviewing site, we classified the star ratings into two groups: “Thumbs Up/Positive/Good” - (3 Stars, 4 Stars, 5 Stars) and “Thumbs Down/Negative/Bad” - (1 Star, 2 Stars). Conclusively, we check the accuracy of non-negative and positive polarity scores against “Thumbs Up/Positive/Good” and “Thumbs Down/Negative/Bad” for each review respectively.

After achieving a considera ble accuracy over baseline, we proceed to the categorization step. Once we have a collection of extracted phrase it depends on the question in consideration to form categories. For this work, we tried to evaluate a few metrics for the hospitals and classified the phases into various categories. After grouping, the data was represented and visualized to spotlight on the issues raised in the Introduction section.

4. ANALYSIS OF RESULTS

In this section, we drew a correlation between star ratings and polarity values. Following, we checked the accuracy of the model for both the categories in various cities. After convincing results, we proceeded to visualize these phrases/tagged phrases in various plots.

4.1 Correlation between polarity scores and star rating

We checked the correlation between polarity scores and star ratings because it conveyed that the prior polarity formula used to calculate the polarity score is a good measure to pick the phrases and classify the review.

Table 2. Correlation values of hospitals and care centres reviews

City	Hospitals reviews	Care Centres reviews
Charlotte	0.495	0.527
Las Vegas	0.535	0.572
Madison	0.387	0.668
Phoenix	0.553	0.571
Pittsburgh	0.597	0.401
Urbana	0.693	0.709

In table 2., the correlation values between star ratings and polarity scores is shown. In “Hospital” category, the correlation ranged between 0.495 and 0.709. Charlotte gave a low value but it was still satisfactory. In “Care Centres” category, Pittsburgh gave a low correlation value: 0.401 while Madison gave the highest correlation. We are not considering Urbana because the review set was too small to determine a reliable correlation value.

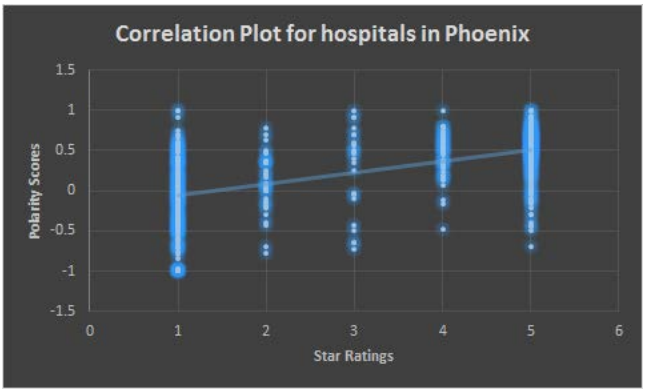

Figure 1: Correlation between star ratings (1-5) and polarity scores (-1 to +1) for all the reviews about hospitals in Phoenix

To give a granular perspective, we plotted a correlation in Figure 1. that states if the polarity score is negative then they mainly fall in the rating (1,2) and positive scores fall in the range (3,4,5)

This was observed for all the other cities as well. Hence, this adds to the justification of classifying labels as “Positive/Thumbs up” and “Negative/Thumbs down” reviews instead of a 5-point star rating.

4.2 Model Evaluation

Table 3 : Accuracy of the model on reviews of hospitals across cities

City	Baseline	Accuracy
Charlotte	75%	82.50%
Las Vegas	53.76%	72.48%
Madison	85%	80%
Phoenix	54.69%	72.05%
Pittsburgh	68.96%	79.31%
Urbana	75%	87.50%

As seen in Section 3.4 and Section 4.1 with sufficient justification, we classified our models against two labels. The baseline has been calculated by considering if we randomly guess all the reviews to be positive (majority class) Thus, we tried to adjust the polarity formula, subjectivity and positivity threshold and POS tag for phrases so that our model gives a good accuracy without overfitting the reviews. In table 3, we expressed the accuracy of the model and it can be observed that the accuracy is sufficiently greater than baseline. These results bolster the claim that the phrases extracted are expressive and add weight to the sentiments of the reviews. However, we were not able to achieve good accuracy for Madison and the potential reason for this is the lack of reviews.

Table 4 : Accuracy of the model on reviews of care centres across cities

City	Baseline	Accuracy
Charlotte	80.96%	87.88%
Las Vegas	65.21%	78.96%
Madison	70%	85%
Phoenix	67.99%	79.41%
Pittsburgh	75.71%	71.42%
Urbana	62.5%	62.5%

In Table 4, we have expressed accuracy for care center reviews, this has also achieved considerable results. Since the data for Urbana was comparatively less, we virtuously consider the accuracy inadequate to draw any inference. Also, it was observed that Pittsburgh is not able to cross the baseline by a few points.

Category A	Category B	Category C	Category D	Category E
excellent experience, excellent experiences, good experience, good experiences, great experience . . .	amazing environment, comfortable gym, comfortable seating, good atmosphere, great food, great office, healthy meal	amazing communication, amazing counselor, amazing doctor, amazing Gastroenterologist, courteous staff, extremely friendly, experienced physician, great humility .	amazing job, fast service, fantastic results, fantastic medication, good massage, good meds, Good service, good trainers, great massage . .	Happy hour, great price, great incentives, great deal, great deals, Good prices, good deal . . .

Figure 2: Example of phrases extracted as per category

4.3 Phrase Evaluation

In this phase, we categorized the extracted phrases and tried to present some useful analysis. We first figured out what all categories we should form to pigeonhole the phrases. Following are the categories and the type of phrases they include:

1. Category A contain phrases that describe experience
2. Category B contain phrases that express criticism for environment, facilities, atmosphere, equipment, waiting rooms, bedsides, etc.
3. Category C include phases that comment on the doctors, nurses, receptionist, staff, etc.
4. Category D contain all the phrases that talk about care, treatment and service

5. Category E consists of phrases that speak about the economic aspects of a visit

We then tagged the phrases manually into these categories and if they don't fall into any of the above categories we consider those phrases as futile. Table 5 gives the unique phrase count for Thumbs Up and Thumbs Down phrases from all the phrases. Since we had to manually tag the phrases we were not able to group all phrases for all the cities.

Table 5: Unique thumbs-up and thumbs-down phrases for all the hospitals and care centres in Phoenix, Pittsburgh and Madison

	Unique-Useful Thumbs Down Phrases	Unique-Useful Thumbs Up Phrases	Total Unique Phrases Extracted
Phoenix Hospitals	205	90	721
Phoenix Care Centres	528	101	1565
Pittsburgh Care Centres	64	10	128
Pittsburgh Hospitals	48	16	130
Madison Hospitals & Care Centres	31	5	80

We are considering unique number of phrases since we want to know feature-oriented feedback. Table 5 provides the count only for the 5 categories, if the categories to group the phrases change then the count will change accordingly. We have combined the counts for Madison Hospitals and care centres because the number of phrases emitted by the model were less to consider them in isolation.

4.4 Phrase Visualization

This phase of the work depends totally on the categories. Hence, depending on the answer one wants to seek, we can use the phrases as fundamentals for a given problem. Focusing on our problem of care centres and hospitals, we start by modeling a comparison of Thumbs Up and Thumbs Down remarks for all care centres in Phoenix in Figure 3

Such ratios will have help determine how the performance of care centres in specific categories.

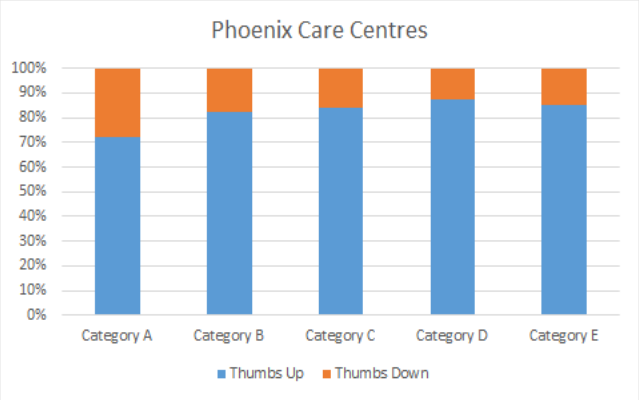

Figure 3: Ratio of Good and Bad remarks for care centres in Phoenix

We can also determine if we want to compare 'Thumbs Down' or 'Thumbs Up' remarks for different cities either for some or all the categories. Figure 4: Shows the ratio of number of remarks received for each category for two cities.

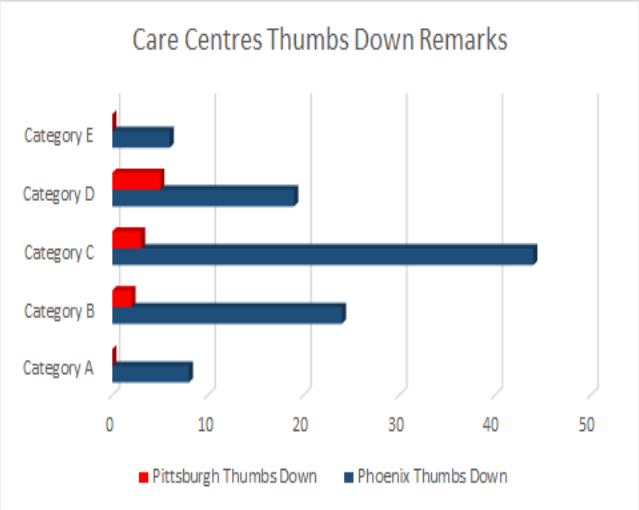

Figure 4: Good and Bad remarks for care centres in Phoenix

Figure 2 is a sample of coarse-grained extraction example of a few "Thumbs Up" phrases that are generated at the time of classifying the reviews for care centres in Phoenix.

5. FUTURE WORK

A commonplace issue observed in various sentiment analysis model is to predict sarcasm. Our model is vulnerable for phrases that involve sarcasm like "fantastic wait". Also, some positive reviews contain words that are relevant to illness which makes the polarity score negative, hence affecting the accuracy, specially in the case of Madison hospitals and Pittsburgh care centres.

Currently, we are manually tagging the phrases which would be time consuming for cities like Phoenix and Las Vegas. One can make use of synsets from SentiWord or Pattern package that could be helpful for deriving similarities between the categories and phrases without human intervention. But the model will become computationally expensive.

There has been a lot of research done to derive the semantic orientation for a sentence by prior polarity formula. A sentiment classification research done by Yan Dang et. al [14] use a polarity

formula that can be replaced with the current model and the models can be compared.

6. CONCLUSIONS

The model generates valuable phrases that can be categorized into various categories depending on the problem in consideration. For our problem, we picked only noun phrases (JJ|NN, JJ|NNS) and adverb phrases (RB|JJ, RBR|JJ) for each review, still the accuracy obtained across various cities in the categories of “Hospital” and “Care Centre” is favorable. Providing such qualitative feedback and flexibility of grouping the phrases not only can be applied to “Health & Medical” but it can be applied to various other businesses a broad range of problem space. The application of this model is not limited to Yelp but various business review or ecommerce websites can be considered for analyzing reviews. Such models can be used for knowing the summarized reviews of one’s own or competitor’s product/business.

7. ACKNOWLEDGMENTS

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed. Vincent Malic, for splendidly teaching us sentiment analysis and how to think critically about it.

8. REFERENCES

[1] Peter D. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”, National Research Council of Canada, Ontario, Canada.

[2] J. Freedman and D. Jurafsky. 2011. Authenticity in America: Class distinctions in potato chip advertising. Gastronomica, 11(4):46–54.

[3] V. Chahuneau, K. Gimpel, B. R. Routledge, L. Scherlis, and N. A. Smith. Word salad: Relating food prices and descriptions. In EMNLP-CoNLL, pages 1357–1367, 2012.

[4] W. Farhan, “Predicting Yelp Restaurant Reviews,” UC San Diego, La Jolla, 2014.

[5] Blair-Goldensohn, Sasha, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, and Jeff Reynar.

[6] Building a sentiment summarizer for local service reviews. in Proceedings of WWW-2008 workshop on NLP in the Information Explosion Era. 2008.

[7] M. Hu and B. Liu, “Mining opinion features in customer reviews,” in Proceedings of AAAI, pp. 755–760, 2004.

[8] Yelp Dataset Challenge details: https://www.yelp.com/dataset_challenge

[9] S. S. Htay and K. T. Lynn , “Extracting Product Features and Opinion Words Using Pattern Knowledge in Customer Reviews,” *Scientific World Journal*, Dec. 2016.

[10] https://www.eecis.udel.edu/~trnka/CISC889-11S/lectures/dongqing-chunking.pdf

[11] Guerini M, Gatti L, Turchi M. Sentiment analysis: how to derive prior polarities from SentiWordNet. In: Proceedings of empirical methods in natural language processing (EMNLP); 2013. p. 1259–69.

[12] http://www.clips.ua.ac.be/pages/pattern-en

[13] https://www.yelp.com

[14] Dang, Yang, Yulei Zhang, and Hsinchun Chen. "A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews." N.p., n.d. Web.