Literature review

Abhay Pawar(asp2107) Conrad De Peuter(cld2167)
Vinayak Bakshi(vb2424) Varun Aggarwal(va2344)

12 November 2016

1 Introduction

With the increasing web connectivity and development of websites that encourage participation of users, huge chunks of text data voicing public opinion is being generated daily. Most importantly users provide reviews about their experience of using a product, or watching a movie, or dining at a restaurant. These reviews are important in context to a new user to help him/her make an informed decision. However, many reviews are lengthy with only few sentences expressing the author's opinions. Therefore, it is not easy for people to collect relevant information. Moreover, for every information unit to be reviewed, such as a restaurant, there are numerous reviews and reading them all is unfeasible. Again, reading few reviews could give a biased opinion. Thus, automatic review mining and summarization has become a popular research topic which we would like to explore.

On most of the review websites, each review has an overall rating. However, the overall rating fails to capture the finer details of the review. For example, two reviews can have the same 2-star rating, but due to different reasons. A user more particular about the ambiance of a restaurant may give it a poor rating but for another user, food quality could be the driving criteria. From the perspective of a user reading the reviews to get information about a restaurant, the evaluations of the specific aspects are equally important as the overall rating of the product

2 Goal

- We will predict ratings for various aspects present in reviews. We will first identify sentences which talk about an aspect and use these sentences to create features. Example of aspects could be food, service, value for money for restaurants. TripAdvisor has aspect-level ratings and we will use this data to train our model.
- To validate the identification of sentences talking about an aspect we will use an annotated dataset, of 652 restaurant reviews from CitySearch.com, introduced by G. Ganu et al. 2009[5]. The dataset contains annotation of the aspects that each sentence is talking about in a review.
- The second major task will be predicting overall rating of a review. We will use aspect level ratings predicted from previous task as well as features created for the review as a whole.

3 Research Hypothesis

- We aim to formulate a novel methodology, building on previous research, for mining aspect level rating from product reviews.

- We will validate the effectiveness of bootstrapping in segmenting aspects from overall reviews. We will test if LDA and its variations can identify various keywords for broad aspect categories.
- We will then test the importance of variables in predicting the rating class of aspects.
- We will also test if the aspect level rating predictions can predict the overall rating of a review. We also use other features used for aspect rating prediction.

4 Previous Work

Human performance in identifying sentiment behind reviews was studied by Pang and Lee (2005)[14]. They performed 2 subject studies on a movie review dataset. Both the individuals were able to discern perfectly well when asked to compare which amongst 2 reviews from same author was more/less/equal positive when the difference in rating of review was at least 3 notches. This research validates the need for aspect level ratings because humans are able to differentiate between positive and negative reviews very well. We shall now summarize the previous work on three major sub-problems of our project and describe how it can be useful to us.

Aspect Mining and Aspect Rating Prediction:

Past research has focused on both identifying the broad aspects present in reviews and identifying the sentences talking about a specific aspect. LDA and its variations have been widely used to identify aspects present in reviews.

One of the earliest research on aspect mining was conducted by M. Hu and B. Liu (2004)[9]. Their aim was to mine product features that have been commented on by customers and then to collect positive/negative sentences about each such feature. A POS tagger[1] is used to parse reviews. They hypothesise that when customers discuss a common feature, their words will converge. Thus, the authors use the Classification Based on Association miner [11] to identify frequent nouns and noun-phrases and categorise these as the product features. The feature set is pruned for redundancy and compactness. After compiling a list of the common features, they identify opinion words as those adjectives that modify the feature word/phrases and then look at the orientation of these words to summarise the positive/negative sentences about the feature.

Opinion mining has also been applied in other domains such as Movie reviews. Li Zhuang et. al. (2006)[10] extracted feature-opinion pairs from movie reviews and identified the polarity of the opinion. A class-level keyword list is created from the labelled reviews for features and opinions based on frequency. Next a dependency grammar graph is generated to identify relationship between opinion and feature (Ex: NN - amod - JJ, VB - advmod - RB). Four dependency relation templates with highest frequency are taken. The keyword list and dependencies are used to mine explicit feature-opinion pairs from the unlabelled data. This is an improvisation over Hu Liu et. al (2004)[9] in extracting valid opinion-feature pairs by using dependency graphs over choosing nearest opinion word. We intend to use similar dependencies as one of the feature in our feature set. A challenge in mining aspects in movie reviews was that reviewers also comment on the movie-related people viz. Director, Actor, etc. apart from movie related features. Therefore the feature set needs to be accurately defined.

In the absence of well defined aspects, techniques like LDA, PLSA [2] can be used to identify aspects before predicting their ratings. Although in many cases aspects identified are not ratable. In order to account for this shortcoming, a Multi-grain Latent Dirichlet Allocation was carried out to extract local and global features of product reviews by Titov McDonald (2008)[16]. They built a model which takes a review and returns the various aspects discussed in a review, ratings for these aspects, as well as textual evidence supporting the rating for these aspects. To label textual evidence supporting their aspect ratings they use this MG-LDA model, which models a distribution of topics for overlapping windows of phrases in the document, and chooses the topics which have maximum probability, allowing a phrase to be related to multiple aspects. We plan on using this method to extract phrases which relate to our given topics.

Hongning Wang et al.(2010)[7] designed a bootstrapping algorithm for aspect segmentation. They manually selected a set of keywords for each aspect. The review set D, aspect keywords T, vocabulary V, threshold p and iteration limit I are given as input to this algorithm. The algorithm works by first splitting the reviews into sentences. It then matches the aspect keywords with words in each sentence and records their frequency. Based on aspect frequency, the sentence is labelled with an aspect. The keyword list is updated by ranking the words in V by their chi-squared value within each aspect and selecting the top p words as the new keyword list for the next iteration. The above procedure is repeated until top p keywords remain unchanged. We will use this technique for aspect segmentation for our corpus with an additional improvisation on selecting seed words. We can obtain seed words based on tf-idf which will give us more accurate results. Once, we segment out sentences talking about an aspect, we will use these sentences to create features to predict the rating of that aspect.

Hongning Wang el al.(2011)[8] in their following work on Latent Aspect Rating Analysis designed a unified generative model for aspect segmentation and aspect rating prediction by incorporating topic modelling techniques into LRR. We won't be doing the task of topic modeling to identify the list of aspects present in reviews because the aspects for various domain are generally well-defined(Eg. Aspects of Food, Service, Ambience,Value for Money, etc. are well defined for restaurants and it is unlikely that we will find any new aspects).

Feature Selection

A simple bag of words representation is not always effective. For example: 'Good service in a very bad restaurant' and 'Very Bad service in a good restaurant' convey different things but have the same BoW representation. Thus, there is a need to incorporate more complex features. Gupta, Fabbrizio, Haffner (2010)[12] in their work on predicting ratings for service and product reviews, experimented with different features, viz. unigrams and bigrams occurring more than 3 times in the corpus; word chunks obtained by processing Nouns, Verbs and Adjectives; POS of the word chunks; and overall rating. Average rank loss was used as their performance metric to identify prominent features. Using only unigram features gave best model performance which was attributed to sparseness in data. They compare the performance of 3 models viz. numerical regression(neural network), ordinal regression(PRank algorithm) and classification(MaxEnt classifier). Minimum rank loss was obtained for MaxEnt classifier. They also enhanced their predictions by modeling interdependence among aspect ratings. For this, they trained MaxEnt classifiers for each possible pairs of ratings. This classifier assigned a probability value to getting the observed difference between ratings of any two reviews. Finally, to predict aspect ratings for a given review text, they used both the original rating predictor and the difference predictor.

The research by Baccianella, Esuli, Sebastiani (2009)[15] deals with feature vector representation of text for multi-aspect review rating using ordinal regression (ϵ-support vector regression). They first extract PoS patterns (E.g, ADJ NN = 'horrible hotel'). The extracted features are made more robust by reducing them to canonical forms and replacing with lexical equivalents. The most important features are selected through techniques devised specifically for ordinal regression based on the variance of the features over the ratings. The try to minimise the mean absolute error. They found that their sophisticated and robust features performed better than simple BoW features.

Inspired by the papers mentioned above, we plan to use a number of different feature sets for our models. We aim to find out which features prove to be the best predictors. We will also perform ablation studies to determine the contribution of each group of features.

Rating prediction and Classifiers

Pang and Lee 2005[14] compare an SVM one-versus-all classifier and a linear, ϵ-insensitive SVM regression on movie review data to get 3-class and 4-class ratings. Their main aims is to improve these classifiers by exploiting the fact that similar items should get similar labels. Their model increasingly penalises the classifier if it assigns divergent labels to similar items by minimising the following cost function

$$\sum_{x \in test} [-\pi(x, l_x) + \alpha \sum_{y \in nn_k(x)} d(l_x, l_y) sim(x, y)] \tag{1}$$

Here, $\pi(x, l_x)$ is the initial label preference obtained by any of the two methods described above. $d(l_x, l_y)$ is the distance between labels. $nn_k(x)$ are the nearest neighbors of x based on similarity metric sim(x,y). The similarity of items has traditionally been described as the cosine between term-frequency-based document vectors. However, they found that items with high similarity measured by this similarity metric does not correlate with the similarity of the labels. Instead, the authors define their own similarity metric called the positive-sentence percentage (PSP) under the hypothesis that items are similar if they get similar labels, and would thus contain a similar ratio of positive to subjective terms. Good results were obtained with 3 way classification. But with 4-way classification, regression + similarity metric gave comparable results to regression alone. Their research showed encouraging results and so, we plan to exploit similar inter-dependencies among labels.

Goldberg and Zhu(2005)[6] introduce a graph based approach for the scenario where there is a limited amount of labeled data and a lot of unlabeled data, similar to the Wang dataset for aspect-level ratings. They built a graph where each node is a document/review, and have a similarity measure for each of the reviews. They draw an edge between a node and its k closest neighbors, and this edge has a weight which is calculated by a similarity measure between the documents. In addition they include ratings predictions from a separate learning algorithm for the unlabeled documents. Every labeled document is connected to a node which has its ground-truth rating score whose only neighbor is the labeled document, the same can be done for the unlabeled documents and their scores from the separate learner which we can get from any unsupervised approach. After defining the loss over the graph as the sum of the difference between the predicted scores and the actual scores plus the sum of the difference between each node and its nearest neighbors multiplied by their edge weights, they find the predicted weights which minimize this loss. We will attempt to build a similar graph for each aspect, using the aspect-specific rating as our labeled data and reviews without aspects rated as unlabeled documents. For the separate learner, we will use one of the other approaches listed above.

Yue Lu et. al (2009)[18] rated aspects using a simplistic approach by Local and Global prediction. In Local prediction, they assumed each phrase mentioned in the comment shares the same rating as the overall rating of the comment. In Global prediction, an aspect level rating classifier is learnt using the global information of the overall ratings. The phrases extracted are in the form of modifier-aspect pair where the modifier is an adjective/adverb expressing an opinion towards the aspect. First, we obtain an empirical distribution for modifier given aspect and rating. Next, each phrase is classified by choosing the rating class that has highest probability of generating the modifier in the phrase (Naïve Bayes) It is inferred that the ratings predicted by global prediction algorithm are more accurate in ranking the aspects. For our corpus, we can use the local prediction as one the baseline models to compare against our classifier.

Wang and Ester (2014)[17] introduce a sentiment aligned topic model(SATM), which models aspect ratings as a multinomial distribution. They first create an aspect-word distribution. Then for each sentiment they create an aspect-sentiment label-word distribution. Using a review's overall rating as a Bayesian prior, and a labeled sentiment lexicon, they use Gibbs sampling to predict aspect-level ratings. Using the TripAdvisor aspect-level ratings as a gold standard they show their model can predict aspect-level ratings with good success.

Fang and Zhan (2015)[4] introduce SVM, Naive Bayes, and Random forest models to predict sentiment polarity. Building on Hu and Lui's (2004)[9] work, which created a list of positive and negative words used in reviews, they attempt to predict sentiment polarity using a BOW approach which compares the number of positive and negative words in a sentence. This approach was performed at the document level, but after splitting a review into its various aspects we can apply it to the aspect level.

Overall Rating prediction:

The second part of the goal is predicting the overall rating of a review. We will use features similar to those used for aspect rating prediction. The whole review will be used for creating these features.

Hen se and Rajat (2015)[13] used vector space transformation of bag of words and used these as features for Neural network model. They compared their results with Naive Bayes model using Bag of words. The neural network model performed better than the Naive Bayes model.

Duyu Tang, et. al. (2015)[3] have also taken into consideration the user-specific meaning of a word. Some users use certain words with much higher frequency and do not imply as much emotion as someone who uses it rarely. They trained a UWCVM (user-word composition vector model) to infer these user-specific meaning. The final word vectors which are used to train the model come from this UWCVM.

We intend to use the word vector representations to use as features for our models. Instead of using all the words we also intend to use only those words which have high polarity. To find these words, we will find the ratio of frequency of each word in positive reviews to negative reviews. Only words with a ratio greater than 2 will be chosen. Similarly, only those words with ratio of frequency in negative to positive reviews greater than 2 will be chosen.

Features: We will use all the features that we discussed for aspect rating prediction for overall rating prediction as well. We would use the aspect rating predictions as well for this task as one of the features and test their efficacy. For a lot of reviews, nothing would be explicitly mentioned about an aspect. In this case, we will test two hypotheses: use these overall features for that specific review or exclude that review to train the aspect model. The assumption for first hypothesis is that the reviewer has same sentiment towards the aspect. Previous research hasn't discussed this issue and we will try to solve this by testing these hypotheses.

5 Summary of what we intend to do

Aspect Mining

We will test the bootstrapping algorithm from Hongning Wang et al.(2010)[7] to find sentences relevant to an aspect. We will also use LDA [2] to see if we find any new aspects and if aspects can be clubbed into a broad aspect category. These aspects can also be used as seeds for the bootstrapping algorithm or keywords to search for relevant sentences. We will use data from G. Ganu et al. 2009[5] to validate our results.

Features

We found that large number of papers focused on unigrams and POS features. Unigrams typically gave good results. We intend to do a holistic analysis of various features. There is no research which does a comparative study of all features for this task. Broadly, we will look at following features:

- Bag of Word features - Unigram, Bigram
- Word phrases as feature-modifier pair
- POS/dependency trees to find useful phrases
- Word2vec features

We also need to tackle the issue of reviews not containing any sentences about an aspect. We will test if review level features work well in such cases. If not, we will exclude such reviews from the training data.

Classifiers

Largely the papers have used classifiers like SVM, Naive Bayes, etc., linear/ordinal regression techniques. We will build one vs all models to classify the review into one of the many ratings. We will use the typical classifiers like SVM, Naive Bayes, Logistic regression. If time permits we will try out Neural Networks with word2vec features.

Overall Prediction of rating

We will test if the aspect level predictions are a good predictor of overall rating. We will also use features similar to that used for aspect rating prediction. The features will be created at review level. We will use the same classifiers that were used for aspect rating prediction.

6 Team Member contribution

Vinayak Bakshi(vb2424): Compiled introduction to project statement. Read most papers and summarized research on aspect segmentation [7,8,10] and rating prediction[18]. Developed ideas from the above research applicable to our project. Edited final sections on Aspect Mining and Aspect Prediction, Feature selection.

Conrad De Peuter(cld2167): Reviewed and summarized various papers in the rating and various classifiers section, as well as the aspect mining and prediction section to determine their applicability for our project. Performed initial exploratory data analysis of the Wang TripAdvisor data set.

Varun Aggarwal(va2344): Reviewed and summarized multiple papers concerning feature selection and sentiment classification and determined their applicability to our project. Broke down our problem statement into major sub-problems that needed to be solved. All further relevant papers were searched and narrowed down to solve these broad problems.

Abhay Pawar(asp2197): Did literature review and summarization of all papers for the overall rating prediction section. Read papers from other sections to get an overview of the work. Narrowed down on ideas that should implemented and are applicable for the project. Compilation of all ideas and document editing to create the report.

References

[1] *NLProcessor – Text Analysis Toolkit.* 2000.

[2] A.Y. Ng D. M. Blei and M.I. Jordan. *Latent Dirichlet allocation.* Journal of Machine Learning Research, 3(5):993–1022,, 2003.

[3] Ting Liu Yuekui Yang Duyu Tang, Bing Qin. *User Modeling with Neural Network for Review Rating Prediction.* Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), 2015.

[4] Xing Fang and Justin Zhan. *Sentiment analysis using product review data.* Journal of Journal of Big Data, Springer,, 2015.

[5] N. Elhadad G. Ganu and A. Marian. *Beyond the stars: Improving rating predictions using review text content.* 2009.

[6] A. Goldberg and X. Zhu. *Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization.* In HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing,, 2006.

[7] Y. Lu H. Wang and C. Zhai. *Latent aspect rating analysis on review text data: a rating regression approach.* In Proceedings of the 16th KDD, pages 783–792. ACM,, 2010.

[8] Y. Lu H. Wang and C. Zhai. *Latent Aspect Rating Analysis without Aspect Keyword Supervision.* 2011.

[9] M. Hu and B. Liu. *Mining and summarizing customer reviews.* In KDD, pages 168–177. ACM, 2004, 2004.

[10] F. Jing L. Zhuang and X. Zhu. *Movie review mining and summarization.* 2006.

[11] Hsu W. Ma Liu, B. *Integrating Classification and Association Rule Mining.* KDD 1998, 1998.

[12] Giuseppe Di Fabbrizio Narendra Gupta and Patrick Haffner. *Capturing the stars: predicting ratings for service and product reviews.* 2010.

[13] Hen Su Choi Ortiz and Rajat Shah. *Predicting Amazon ratings using Neural Networks.* 2015.

[14] Bo Pang and Lillian Lee. *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales.* 2005.

[15] Andrea Esuli Stefano Baccianella and Fabrizio Sebastiani. *Multi-facet Rating of Product Reviews.* 2009.

[16] I. Titov and R. McDonald. *A joint model of text and aspect ratings for sentiment summarization.* In ACL '08, pages 308–316., 2008.

[17] Hao Wang and Martin Ester. *A Sentiment-aligned Topic Model for Product Aspect Rating Prediction.* 2015.

[18] C. Zhai Y. Lu and N. Sundaresan. *Rated aspect summarization of short comments.* In Proceedings of WWW'09, pages 131–140,, 2009.