# The economical values of review text mining

To be added

# Background of Text Mining

## 2.1 Terminology: words, document, corpus, dictionary

The subjects of language processing and text mining are different levels of texts. The basic level is character, punctuation and words, which form a written representation of thoughts, a document. Document could consist of sentences, paragraphs and chapters. In current study, one customer’s review is one document, which could have multiple sentences or paragraphs. The term corpus or text corpus is used to describe a large collection of documents [1], which could be a collection of review for a particular restaurant or reviews of restaurants across US. The collection of unique words in a corpus forms a dictionary.

## 2.2 Text Representation: Bag of words and N-grams models

The bag-of-words model is a simplifying way to represent the text, where text is converted/tokenized into a vector of distinct words or punctuations and their frequency. For example, a review ‘*First time eating there tonight, and it was a great experience. The service is terrific and the food is delicious*.’ could be converted to a vector of

[(‘first’, 1), (‘time’, 1), (‘eating’, 1), (‘there’, 1), (‘tonight’,1), (‘,’,1), (‘and’, 2), (‘it’, 1), (‘was’,1), (‘a’, 1), (‘great’, 1), (‘experience’, 1), (‘.’, 2), (‘the’, 2), (‘service’, 1), (‘is’, 2), (‘terrific’, 1), (‘food’, 1), (‘delicious’, 1)]

where the word ‘first’ appear once while ‘the’ show up twice. In the bag-of-words model, the order of words can be changed without impacting the outcome of analysis. The syntactical structure of a sentence or paragraph is intentionally ignored in order to efficiently handle the text.

Here the sequence of words does not matter, while n-gram model provides alternative to consider the local word order, which breaks down the text into a contiguous sequence of n words. Bag-of-words is actually a 1-gram or unigram model. The most commonly used n-gram models are bigram and trigram. The bigram representation of the first sentence of above review is

[(‘first time’, 1), (‘time eating’, 1), (‘eating there’, 1), (‘there tonight’, 1), (‘tonight,’, 1), (‘, and’, 1), (‘and it’, 1), (‘it was’,1), (‘was a’, 1), (‘a great’, 1), (‘great experience’, 1)]

And the trigram representation consists of 3 words/punctuation phrases

[(‘first time eating’, 1), (‘time eating there’, 1), (‘eating there tonight, 1), (‘there tonight,’ 1), (‘tonight, and’, 1), (‘, and it’, 1) (‘and it was’,1), (‘was a great’, 1), (‘a great experience’, 1)]

Generally, a document in a large corpus could be represented as a vector consisting of all the words/grams in the corpus dictionary and their frequency, which could be a sparse vector with lots of zero values.

## Topic modeling

In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "service" and "waiter" will appear more often in documents about service, "food" and "delicious" will appear in documents about foods, and "the" and "is" will appear equally in both. A document typically concerns multiple topics in different proportions; thus, in a document that is 10% about service and 90% about foods, there would probably be about 9 times more food words than service words. The "topics" produced by topic modeling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is.

### Generative Statistical Models

Generative Statistical Model is one important way to extract latent sematic topics from document. A generative model is a model for randomly generating observed data values, typically given some hidden parameters. Given a dictionary over a corpus, the probability of generating a particular document could be modeled as a joint probability of all word observed.

### Unigram Model

The simplest generative model used in text mining is unigram model, which is also the basis of other advanced generative models. It assumes that each document is formed by generating words independently, based on an assumption that the bag-of-words representation of language should satisfy a consistency rule called exchangeability. If we imagine generating a new language from our generative model, producing a large and ever-growing corpus of text, all statistical properties of the text should be homogeneous: the probability of finding a particular word at a given location in the stream of text should be the same everywhere in the stream [2].

For example, under the unigram model, the probability of generating a five words sentence of ‘it was a great experience’ is calculated as p(‘it’, ‘was’, ‘a’, ‘great’, ‘experience’)= p(‘it’) p(‘a’) p(‘great’)p(‘experience’), where the probability of each word is evaluated based on a large corpus as the frequency of words divided by the total word frequency. Here the latent topic could be considered as the word distribution to generate the sentences. Some documents with a ‘service’ topic could have a word distribution where ‘table’ and ‘waiter’ could have a high distribution.

### LDA (Latent Dirichlet Allocation) Model

LDA (latent Dirichlet allocation) is one of widely used generative statistical model for topic modeling, which was first introduced in 2003[1]. It assumes that one document could consist of multiple topics, which is a better conceptualization of real life problems than the unigram model. More specifically, LDA is a three level Bayesian hierarchical model.

1. Assume there are *K* latent topics in a *M*-document corpus and each of topics is a word distribution over all the *V* words in dictionary. We assume each word was generated by one topic, as shown in figure 1.

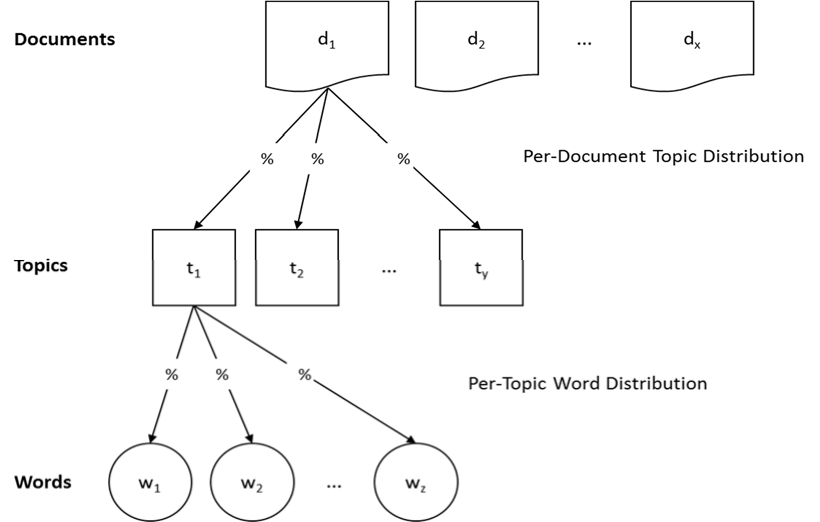


Figure 1. Schematic Overview of LDA [3]

1. For each word in position of document , it followed a multinomial distribution with only one draw, the parameter of which is the word distribution of its topic

where word distribution over the whole dictionary and, which means only one word is generated here

1. The assignment of topic also follows a multinomial distribution with one draw, the parameter of which is the topic distribution of document d,

where , which means only one topic is assigned here

Here it can be found the statistical generation of a word in a document is actually controlled by the topic distribution of that document and the word distribution in each topic

Then LDA use the Bayesian approach to model these two distributions by assigning priors of Dirichlet distribution, which is a conjugate prior of multinomial distribution.

1. Assume has a prior Dirichlet distribution, which is a conjugate prior for multinomial distribution, with a parameter
2. Assume the word distribution of one topic has a prior Dirichlet distribution, with a parameter

Then according to Bayesian rules, the total probability of generating a particular document is

The parameter could be estimated by EM algorithm, the details of which are not explained here.

After are estimated, the word distribution of k topics could be generated from the distribution in equation (1), and the coverage of each topic in one document can be evaluated. Here one topic is a word distribution over the entire dictionary. The word distribution of a topic extracted from a corpus with 19175 words dictionary is shown in Figure 2. This topic consists of 384 words with non-zero probability. Generally, for convenience a topic will be displayed as the first 10 probable words, as shown in Figure 3. In this case, the first 10 words accounts for 26.5% of total probability.

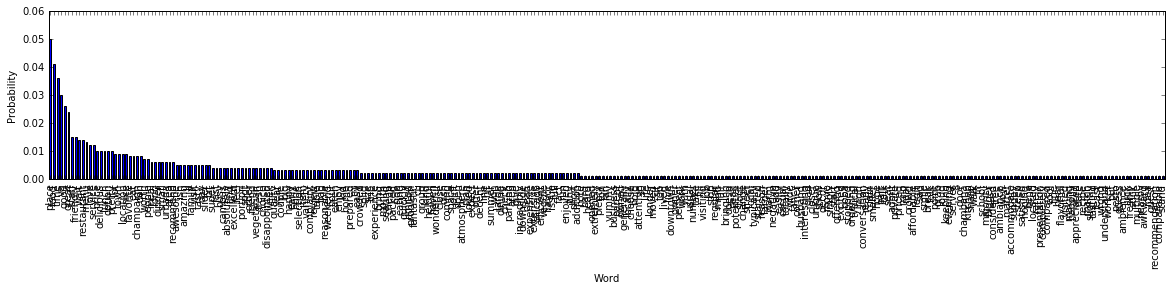


Figure 2. A topic of 384 words extracted from a corpus with 19175 words

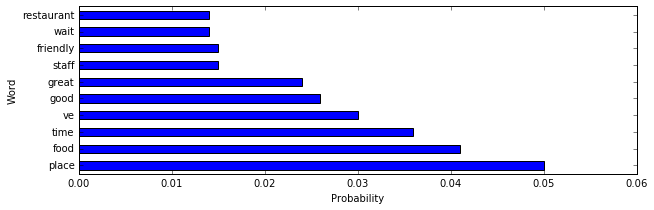


Figure 3. The 10 probable words in the topic of 384 words

## 2.4 Sentimental Analysis:

While the goal of topic modeling is to unveil the patent sematic structure in the documents, sentiment analysis focuses on analyzing people’s subjectivity and sentimental polarity. Obviously the most important indicators of sentiments are sentiment words. These are words that are commonly used to express positive or negative sentiments. For example, ‘good’, ‘awesome’ and ‘delicious’ are positive sentiment words, and ‘bad’, ‘awful’, and ‘terrible’ are negative sentiment words. Apart from individual words, there are also phrases and idioms, e.g., cost someone an arm and a leg. Sentiment words and phrases are instrumental to sentiment analysis for obvious reasons. A list of such words and phrases is called a sentiment lexicon. Over the years, researchers have designed numerous algorithms to compile such lexicons. [4]

VADER (Valence Aware Dictionary for Sentiment Reasoning) lexicon was introduced in 2014 based on existing well-established and human validated sentiment lexicons with additional lexical features in social media [5]. It collected about 7500 lexical features, including regular single words, abbreviation like ‘plz’, emoticons like “:)” and “:-(“, and social media morphology of words like ‘sux’ as another type of ‘sucks’[6]. It provided a sentiment valence score on a scale of [-4, 4] for each lexical feature based on statistical results of human manual rating under careful training and strict quality control procedure [5]. It also considered the booting effect of particular grammatical and syntactical factors, for example, ‘!’ and ‘extremely’ could increase the sentimental intensity of the words in the same sentence, which was evaluated by human expert and validated by a team of raters.

VADER Lexicon provided a convenient and fast approach to quantitatively evaluate sentimental polarity of social media data like customer reviews. Extracting sentimental words, looking up their valence scores, aggregating the scores to sentence, paragraph or document allow the sentiment polarity to be evaluated in multiple text levels.

# 3. Data Description

The data in this project was taken from Yelp data challenge round 7, which includes 67584 US businesses and 9858 business in other countries. The type of business includes restaurants, shopping centers, super markets, Home Services, etc, while restaurant is the largest group and accounts for around 25% of population. The data for restaurants around university of Illinois, Champaign and Urbane (UIUC) were studied in this research, around 300 business and 11205 reviews.

Besides customer views, Yelp also provides comprehensive business attributes such as open hours, neighborhood, categories.

**yelp\_academic\_dataset\_business.json**

{

"business\_id":"encrypted business id",

"name":"business name",

"neighborhood":"hood name",

"address":"full address",

"city":"city",

"state":"state -- if applicable --",

"postal code":"postal code",

"latitude":latitude,

"longitude":longitude,

"stars":star rating, rounded to half-stars,

"review\_count":number of reviews,

"is\_open":0/1 (closed/open),

"attributes":["an array of strings: each array element is an attribute"],

"categories":["an array of strings of business categories"],

"hours":["an array of strings of business hours"],

"type": "business"

}

**yelp\_academic\_dataset\_review.json**

{

"review\_id":"encrypted review id",

"user\_id":"encrypted user id",

"business\_id":"encrypted business id",

"stars":star rating, rounded to half-stars,

"date":"date formatted like 2009-12-19",

"text":"review text",

"useful":number of useful votes received,

"funny":number of funny votes received,

"cool": number of cool review votes received,

"type": "review"

}

# 4. Exploratory Analysis

The geographic distribution of restaurants near UIUC can be shown by plotting the latitude and longitude on the google map, as shown in Figure 1, where size of dot indicates the rating stars. It could be found that restaurants are located around the UIUC campus.

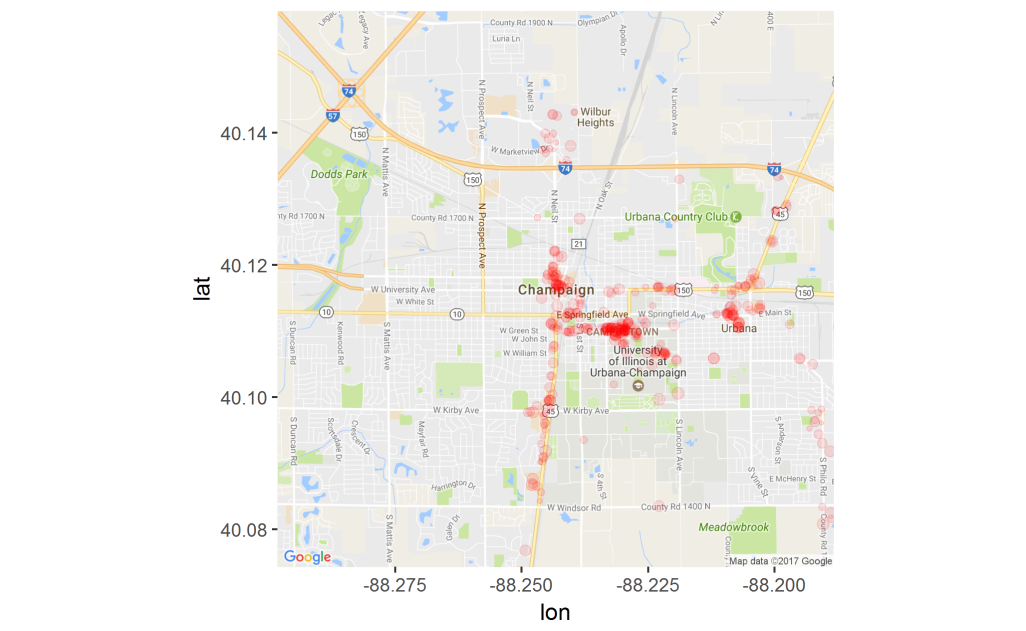


Figure 1 geographic distribution of restaurants around UIUC in Yelp Dataset

The reviews were rated from 1 to 5, while the overall rating for a business is rounded to every 0.5. The rating distribution among reviews and restaurants are shown in Figure 2, which indicates that customers generally leave positive reviews.

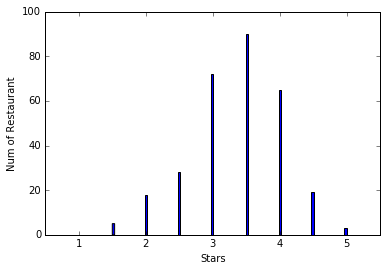
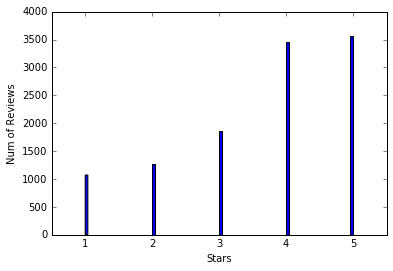


Figure 2 Rating star distributions of restaurants in the studied reviews a) rating of individual reviews b) average review ratings of restaurants

Figure 3 plots the relationship between the lengths of the review text in words, binned by an interval of 20, against the average rating for reviews of that particular length. The trend that is observed indicates that there is a steady decrease in the average rating as the number of words in the review increases. Moreover, the variation in average rating for lengths that are approximately the same is seen to be extremely high as the length increases, especially beyond 500, when compared to lower to middle lengths. It seems that reviewer giving extremely high or low ratings tend to write longer reviews to justify their ratings

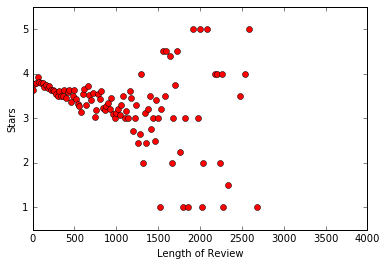
`

Figure 3 relationships between length of review and average rating

# 5. Preprocessing

## 5.1 Tokenizing sentences into bag of words:

In text mining, documents are simply represented as bags of words. Each sentence can be converted into a vector of distinct terms and terms frequency.

## 5.2 Removing stop words and punctuations:

Stop words refer to the most common words in the language which does not provide too much information for the document, such as ‘we’, ‘the’ and ‘there’; Stop words were removed against the NLTK stop words dictionary, which includes 153 words for English language. Common punctuations were also removed during this process.

## 5.3 Lemmatization

Words have different forms, such as ‘do’, ‘did’ and ‘doing’, and different derivations, such as ‘memory’ and ‘memorize’. Lemmatization refers to converting different forms or derivation of terms to the base or dictionary form of a word, which is known as the lemma.

Since VADER lexicon is able to evaluate the stop words and punctuation, the review texts are only tokenized before sentimental analysis.

# 6. Feature Extraction

Two approaches, bigram/trigram phrase extraction and Latent Dirichlet allocation (LDA) were used for topic modeling.

## 6.1 Bigram/Trigram Topic modeling

In text mining, bigram is a sequence of two adjacent words, such as ‘machine learning’ and ‘Chinese food’ while trigram is a sequence of three. Extracting bigram or trigram is a relative convenient way to fetch topics within words which is easy to interpret.

The GENISM wordtovec function, the frequent bigram/ trigram phrases were extracted from the review as table x.

Table 1 the 50 most common bigram and trigram phrase in studied reviews

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bigram Phrases | Frequency |  | Bigram Phrases | Frequency |
| pretty\_good | 629 |  | pulled\_pork | 240 |
| food\_good | 572 |  | love\_place | 232 |
| good\_food | 361 |  | fried\_rice | 226 |
| great\_place | 333 |  | service\_good | 216 |
| black\_dog | 326 |  | beer\_selection | 215 |
| champaign\_urbana | 308 |  | staff\_friendly | 214 |
| 5\_star | 299 |  | highly\_recommend | 210 |
| chinese\_food | 284 |  | good\_service | 209 |
| mexican\_food | 276 |  | service\_great | 208 |
| food\_great | 272 |  | good\_place | 206 |
| great\_food | 267 |  | fast\_food | 200 |
| burnt\_end | 262 |  | great\_service | 189 |
| deep\_dish | 244 |  | quality\_food | 183 |
| sweet\_potato | 244 |  | chip\_salsa | 180 |
| pad\_thai | 243 |  | customer\_service | 173 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Trigram Phrases | Frequency |  | Trigram Phrases | Frequency |
| sweet\_potato\_fry | 162 |  | big\_grove\_tavern | 28 |
| deep\_dish\_pizza | 87 |  | highly\_recommend\_place | 27 |
| food\_pretty\_good | 75 |  | hot\_sour\_soup | 26 |
| champaign\_urbana\_area | 74 |  | thai\_iced\_tea | 25 |
| thin\_crust\_pizza | 69 |  | 3\_5\_star | 25 |
| chicago\_style\_pizza | 52 |  | pad\_kee\_mao | 25 |
| baked\_potato\_casserole | 47 |  | chicken\_pad\_thai | 24 |
| pulled\_pork\_sandwich | 41 |  | 4\_5\_star | 24 |
| authentic\_mexican\_food | 35 |  | beer\_battered\_bacon | 23 |
| general\_tso\_chicken | 34 |  | burnt\_end\_sandwich | 22 |
| bi\_bim\_bap | 34 |  | sweet\_sour\_chicken | 22 |
| place\_champaign\_urbana | 34 |  | sweet\_potato\_chip | 22 |
| great\_beer\_selection | 32 |  | service\_bit\_slow | 21 |
| restaurant\_champaign\_urbana | 30 |  | free\_chip\_salsa | 21 |
| authentic\_chinese\_food | 30 |  | give\_5\_star | 21 |

It can be found that these phrases provided comprehensive information about the content of customer reviews. The phrase like ‘food\_good’, ‘great place’,‘give\_5\_star’ actually directly provide the customers’ feedback.

These bigram/trigram phrases could be used as self–generated labels when new reviews are added. For example, if a business has matched phrases, it can be represented by those labels. They also could be used as feature for predictive models.

## 6.2 LDA (Latent Dirichlet allocation) Topic Modeling

LDA is a generative statistical model for text mining, which considers each document as a mixture of topics. It uses a three-level of hierarchical Bayesian model to fit the data. A fixed number of topics could extract from all the reviews, and then a particular review could be represented as a mixture of topics. Take one review for example:

“I think this one is actually a 3.5 star, if that were allowed...An unpretentious little neighborhood spot, this bar and grill has a 'sports bar' theme but a very relaxed, although almost too clean for, near-dive-y feel. Not crammed with student crowds and a good mix of ages although many boomer types were represented. On U of I basketball nights, they serve free chili. Our dinner plans canceled abruptly, two housemates and I headed over to check it out. Expecting a watery, bean heavy chili soup; I was pleasantly shocked by the meaty and rich actuality. It was sooo good that we all decided to stop by sometime to try menu items we'd actually pay for in cash! Drinks were good, and strong after they realized we'd tip. The bar was crowded, but not unpleasantly. Only complaint was that the fresh diced onion and grated cheese toppers were not refilled after they ran out.”

It could be represented as a mixture for following topics out of 50 extracted topics:

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Topic (10 most probable words out of 19175 dictionary) | Proportion | Manual Interpretation |
| 29 | 0.130\*beer + 0.088\*bar + 0.040\*drink + 0.038\*selection + 0.026\*place + 0.021\*good + 0.021\*bartender + 0.020\*game + 0.018\*night + 0.018\*tap | 13.9% | Bar |
| 11 | 0.033\*italian + 0.014\*walking + 0.013\*chili + 0.011\*turkey + 0.010\*fault + 0.010\*husband + 0.010\*7 + 0.009\*lunch + 0.009\*pasta + 0.009\*opted | 12.1% | Food-Chili |
| 39 | 0.045\*place + 0.038\*nice + 0.024\*table + 0.023\*seating + 0.022\*good + 0.018\*great + 0.018\*staff + 0.017\*area + 0.016\*spot + 0.015\*sit | 13.1% | Service  (Seating) |
| 31 | 0.031\*table + 0.024\*time + 0.024\*waitress + 0.022\*asked + 0.021\*food + 0.019\*server + 0.018\*didn + 0.018\*service + 0.017\*drink + 0.015\*ordered' | 11.3% | Service  (waiting) |
| 45 | 0.066\*cheese + 0.049\*salad + 0.034\*bread + 0.028\*good + 0.024\*sandwich + 0.018\*tomato + 0.016\*onion + 0.015\*soup + 0.015\*chicken + 0.013\*lunch | 9.8% | Food-cheese |

It could be found from this model, as LDA is a generative statistical model and a topic is presented as distribution of 10 most probable terms from a large dictionary, the topics are not as straight forward to interpret as the bigram/trigram phrases. It should be noticed that sometimes the 10 most probable words in a topic may not appear in the text, which means the contribution of that topic actually come from the words from lower probability.

An important parameter of LDA is the number of topics to be derived. When the number is small, information from different topics could combine into one, while large number of topics could cause one topic to be divided into multiple subtopics. A topic number 50 was chosen after trials with 5, 20, 50 and 100.

## 6.3 Sentimental Analysis

Sentimental polarity of customers was evaluated based on a [-1,1] scale; where -1 represent absolute negative, 0 as neutral and 1 means positive. The polarity score of each sentence was derived using VADER approach. The VADER (Valence Aware Dictionary for Sentiment Reasoning) algorithm provides a polarity score lookup table of sentiment related words, which was derived from training on social media text(tweets, movie review and amazon reviews) . Then the polarity score of each review is derived by averaging of polarity scores of sentences, the distribution of which is shown in Figure 4.

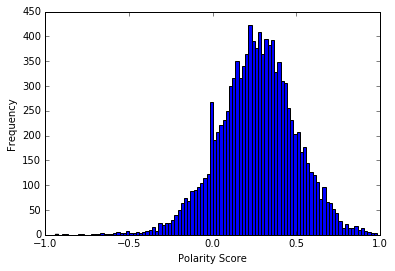


Figure 4 distributions of derived sentiment polarity/intensity scores

# 7. Predictive Analysis

Review ratings in a range of 1 to 5 can be considered either numerical or categorical, as there is no distinct difference when a 1 star review is misclassified as 2, while it is a terrible error if it is labeled as 5. Both approaches were implemented and compared. Regression models were built for numerical ratings; According to methodology of ‘Net Promotor Score’, an industrial system for measuring customer experience, Ratings 1 to 3 were relabeled as ‘detractor’, 4 as ‘indifference’ and 5 as ‘promoter’ for building the classification models.

**7.1 Regression Analysis of Review Ratings**

Linear regression, lasso, support vector machine, and random forest methods are used to build the regression models. Table 2 compares the model performance using 5 folder cross validation in term of R-squared,

where is the observed value, is predicted value and is the mean of observed data. Generally R-squared is less than 1. However, it can be negative when the performance of model is too poor so that the mean of observed data actually provides provide a better fit than the predicted values.

Simple linear regression yielded a large negative R2 which and support vector machine, while the random forest achieved highest performance.

The negative R-squared indicates that linear regression and support vector machine regression show strong overfitting when applied to this data It leads to severe deviation when the model was built on 80% of data and then tested on the rest 20%.

Table 2 Comparison of regression model performance based 5 fold cross validation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | 5 fold R2 | | | | | Average R2 |
| Linear Regression | 0.3896 | 0.3083 | 0.3504 | -6.104 | 0.4229 | -0.92656 |
| Lasso | 0.2170 | 0.2299 | 0.2151 | 0.1537 | 0.2430 | 0.21174 |
| SVR | -0.0712 | -0.0027 | 0.0584 | 0.0281 | 0.0321 | 0.0089 |
| Random Forest | 0.3813 | 0.3189 | 0.3761 | 0.3217 | 0.4460 | 0.3688 |

Random Forest model allows evaluating the relative importance of variables in the model building. Figure 5 shows the 50 most important features. It could be found that most important attribute is the derived sentiment polarity score, while the length of review is the second. The rest of import features are the topics derived from LDA approaches. Although LDA derived topics are harder to interpret, obvious they capture more useful information than the bigram and trigram phrases.

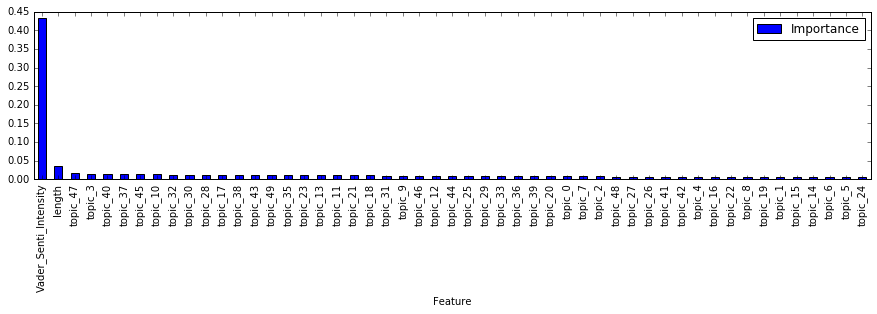


Figure 5 distributions of derived sentiment polarity scores

More model tuning and feature selection should be done in the near future

**7.2 Classification Analysis of Review Ratings**

Logistic regression, support vector machine classification (SVC), random forest classification and K Nearest Neighbors are used to build the classification models to predict the NPS category of rating, ‘Promotor’, ‘Indifference’ and ‘Detractor’. The performance of model is evaluated using the metric of accuracy (ACC):

Surprisingly, logistic regression actually achieved the highest accuracy. We could calculate the accuracy of regression model by converting the numerical prediction from into NPS category, which is 0.5331 for random forest regression, which is still slightly lower than the accuracy of Logistic regression.

Table 3 Comparison of classification model performance based 5 fold cross validation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | 5 fold Accuracy | | | | | Average |
| Logistic regression  (one vs rest) | 0.5660 | 0.5368 | 0.4908 | 0.5443 | 0.5924 | 0.5460 |
| SVC | 0.3898 | 0.3922 | 0.3837 | 0.3908 | 0.4107 | 0.3934 |
| Random Forest | 0.4924 | 0.4859 | 0.5227 | 0.4796 | 0.4859 | 0.4859 |
| k-nearest Neighbors | 0.4054 | 0.4234 | 0.4011 | 0.4042 | 0.4142 | 0.4097 |

More model tuning and feature selection should be done in the near future

Reference:

[2] David J. C. MacKay, Information Theory, Inference and Learning Algorithms, Cambridge University Press; 1 edition (October 6, 2003)

[3] Oliver Müller, Iris Junglas, Jan vom Brocke, Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial, Communications of the Association for Information Systems, vol 39, pp110-135, 2016

[4] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012

[5] C.J. Hutto, Eric Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, 2016

[6] <https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vader_lexicon.txt>