2019

Text Mining - Sentiment Analysis

DATAMINING FOR BUSINESS – IDS 572

DEEKSHA SRIDHAR 658798129
ASHNA MANOJ 672737275

(a) Explore the data. How are star ratings distributed? How will you use the star ratings to obtain a label indicating 'positive' or 'negative' – explain using the data, graphs, etc.? Does star ratings have any relation to 'funny', 'cool', 'useful'? (Is this what you expected?)

Th yelpreviews are a collection of reviews and accompanying star ratings from Yelp. A sample of the original dataset (over 4 million review by over a million users for 144K businesses) will be used here, to keep the assignment task manageable. The dataset we used contains around 50K user review of various restaurants. We primarily focused on id, cool, date, funny, stars, text and found no missing values.

On loading the dataset onto RapidMiner and applying descriptive statistics, we found the distribution as shown below:

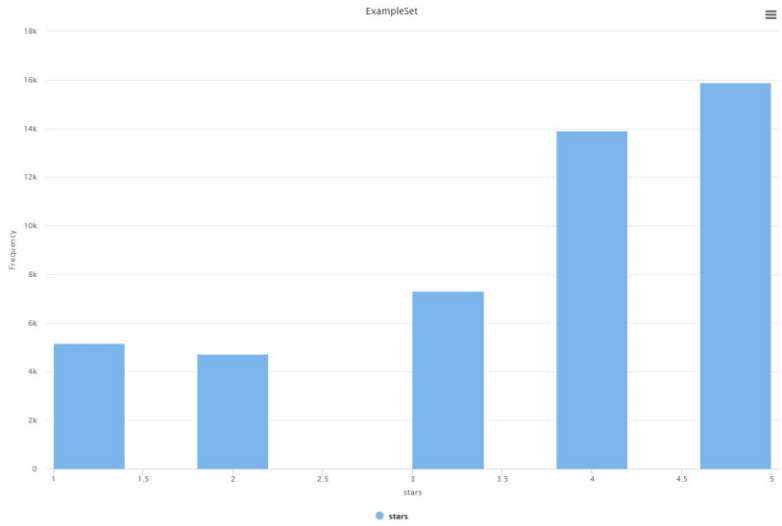


Figure 1: Star Rating Distribution

As seen in the 'star' ratings distribution above, almost more than 5000 reviewers have given a Rating of 1 and Rating of 2 were given by less than 5000 users. Around 8000 users have given a Rating of 3 and approximately 14,000 users have given a Rating of 4 and 16,000 users have given a Rating of 5. We have the star ratings have values from 1-5 . To use the star ratings to obtain a label indicating 'positive' or 'negative' the labelling is done by looking at the values and frequency of the ratings given in the histogram, we took star rating 1 and 2 as negative, 3 as neutral and 4 and 5 as positive rating.

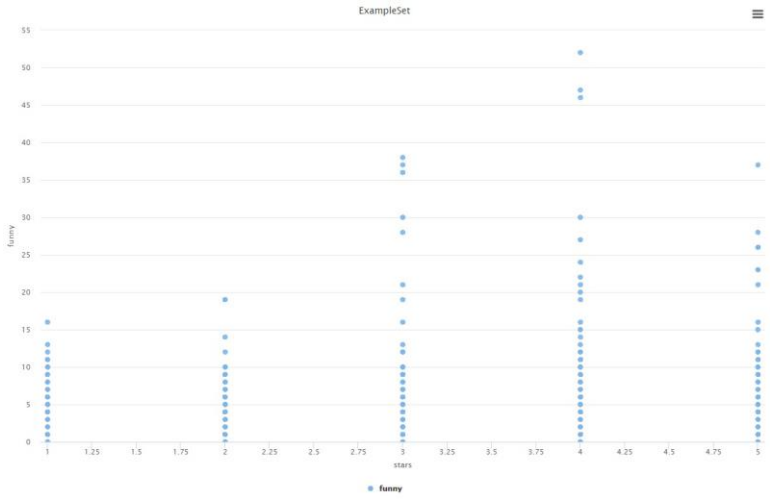
Moving forward to identify if the key words funny, cool and useful:

useful	Integer	0	Min	Max	Average	Deviation
			0	68	0.779	1.733

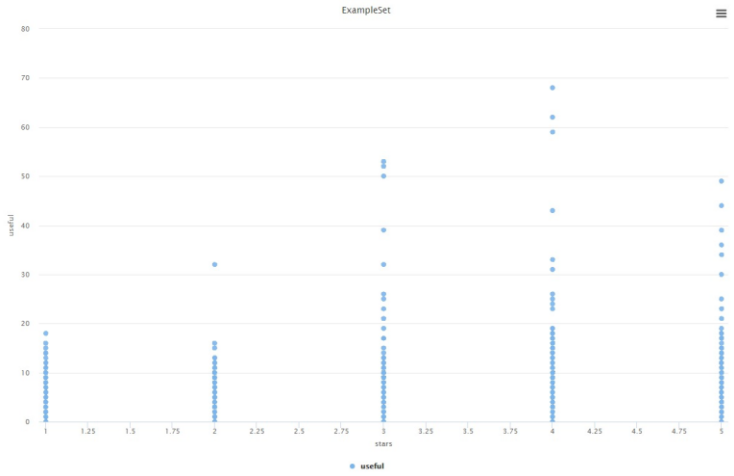
Open visualizations

We used a scatter plot across all star ratings to see the spread of the words- funny, cool and useful.

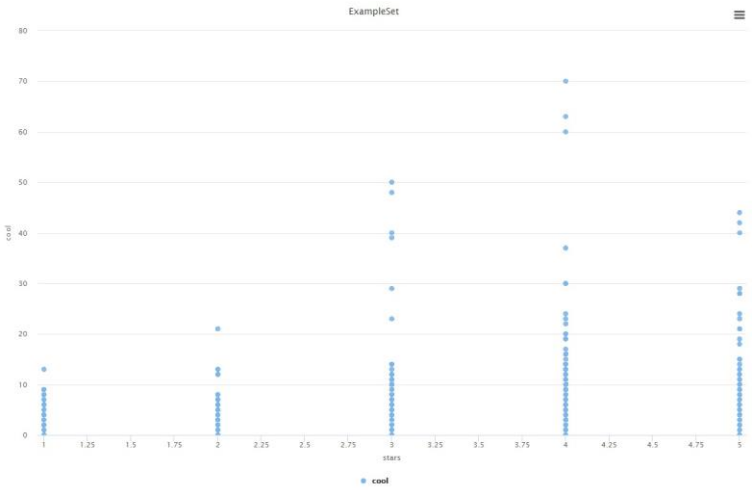
a. Funny Vs Star ratings

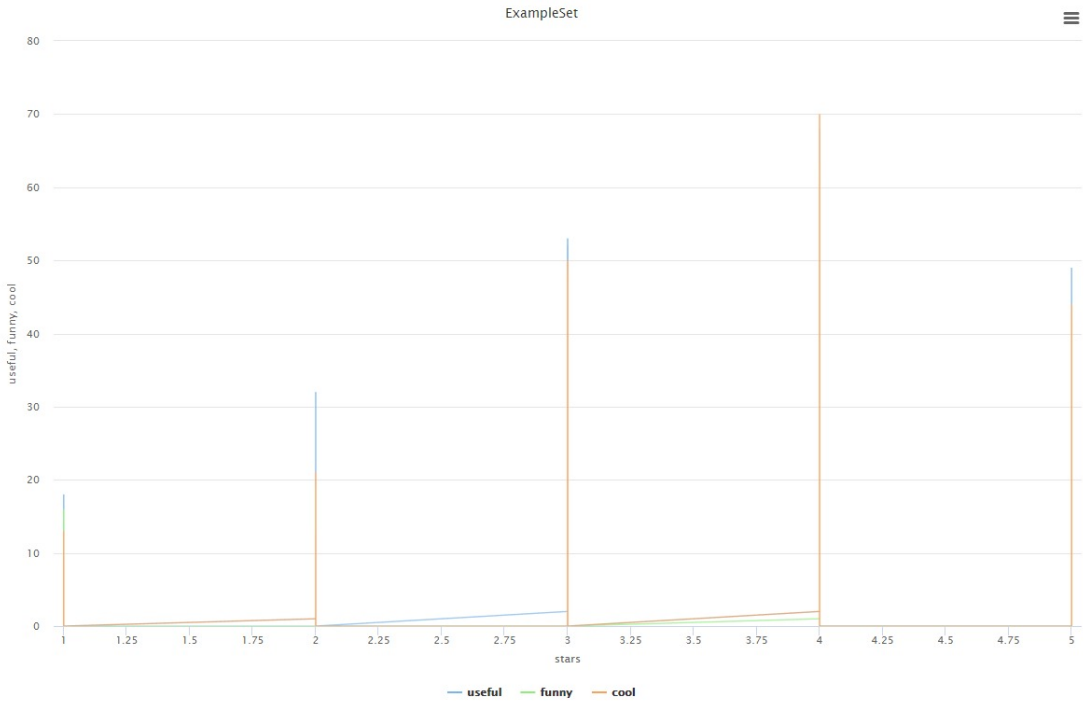


b. Useful Vs Star Ratings



c. Cool Vs Star Ratings





In conclusion, we can see that Ratings 1-2 have more values of Funny and Ratings 4-5 have more of the word Cool. However, the keyword Useful is balanced for all the ratings.

(a) What are some words indicative of positive and negative sentiment? (One approach is to determine the average star rating for a word based on star ratings of documents where the word occurs). Do these 'positive' and 'negative' words make sense in the context of user reviews? (For this, since we wish to get a general sense of positive/negative terms, you may like to consider a pruned set of terms -- say, those which occur in a certain minimum and maximum number of documents).

We found the average star ratings for each word by multiplying the star ratings of a document to the tf of the word for that document and taking average for all the documents where the word was present.

The in document tells the number of documents in which a word was used, while the total column tells the number of times multiple occurrences of the word was used in all the documents.

Positive Words: Using the generate attribute that we have created we can find relatively positive and negative words.

Below are the list of positive words:

Row No.	word	in documents	total	att_1 ↓
580	stars	2578	2966	11
123	de	660	1655	1.834
444	pho	847	1791	1.675
648	und	1111	3727	1.517
294	indian	626	965	1.374
252	greek	557	839	1.350
51	brisket	475	678	1.277
616	thai	1145	2084	1.269
596	sushi	1814	3622	1.260
233	gem	571	579	1.260
312	la	886	1493	1.259
703	yum	625	764	1.241
293	incredible	548	584	1.237
422	outstanding	609	658	1.234
535	sehr	701	1334	1.187
687	wings	1081	1756	1.180
445	phoenix	641	721	1.160
449	pizza	3495	7371	1.146
617	thank	514	543	1.143
298	ist	819	1835	1.140
196	favourite	518	564	1.128

Looking at a few of the positive words below, we can observe

- Words with good scores were related to food items
- The corresponding star ratings in the reviews where These words were used, the star ratings were in the range 3 to 5

Below are the list of Negative words:

- Words which are naturally associated with a negative or unpleasant experience have the lowest att_1.
  Example: Worst, horrible, terrible etc.
- Similarly, The corresponding star ratings in the reviews where these negative words were used, the star ratings were low.

Row No.	word	in documents	total	att_1 ↑
212	food	22876	34433	0.245
451	place	18448	26671	0.303
247	good	18865	28791	0.308
542	service	14436	16732	0.330
694	worst	957	1036	0.372
418	ordered	7375	9856	0.374
235	get	9520	12399	0.380
244	go	10037	12394	0.383
624	time	9670	12618	0.391
67	came	5668	7155	0.411
17	asked	2145	2622	0.417
417	order	6603	8771	0.418
386	nni	5204	6492	0.423
277	horrible	823	909	0.431
629	told	1612	1934	0.431
392	nnthe	5255	6914	0.433
248	got	5905	7707	0.434
495	restaurant	8267	11061	0.436
614	terrible	855	939	0.436
682	went	4947	5620	0.443
366	minutes	2617	3397	0.445

In conclusion, the positive and the negative words do relate in context to the project and can be associated with the resulting star rating of an establishment.

(c) We will consider three dictionaries – the Harvard IV dictionary of positive and negative terms, the extended sentiment lexicon developed by Prof Bing Liu of UIC-CS, and the AFINN dictionary which includes words commonly used in user-generated content in the web. Details on these are given below. As discussed in class, the first two provide lists of positive and negative words, while the third gives a list of words with each word being associated with a positivity score from -5 to +5. How many matching terms are there for each of the dictionaries? (i) Consider using the dictionary based positive and negative terms to predict sentiment (positive or negative based on star rating) of a movie. One approach for this is: using each dictionary, obtain an aggregated positiveScore and a negativeScore for each review; for the AFINN dictionary, an aggregate positivity score can be obtained for each review. Are you able to predict review sentiment based on these aggregated scores, and how do they perform? Does any dictionary perform better? (ii) Compare this approach with use of SentiWordNet. Describe how you use SentiWordNet.

i)

Our text mining analysis was carried out using 3 dictionaries that is, The Harvard Dictionary, The AFINN Dictionary and The Bing Liu Dictionary. We have tabulated the results from our observation in the table below:

Dictionary	Positive Words	Negative Words	Total Matching Words
Harvard IV	46	27	73
AFINN	68	20	88
Bing Liu	72	28	100

The following methodology was used to predict sentiment reviews for the three dictionaries mentioned above.

1. The positive and negative weights are summed up for all the reviews.
2. Positive and negative summary values were created by aggregating the individual scores which was calculated using Term Frequency of the matching positive and negative words and merging them.
3. If the positive_sum value is greater than the negative_sum value implies that the user has more positive words which is indicative of the fact that the user has given a positive rating and vice versa.
4. If positive_sum is greater thsn negative_sum implies that the star rating is higher than 3.5, which means that the user is positive.
5. If the negative_sum is greater than the positive_sum implies that the rating given by the user is negative, which is less than 2.5

The prediction and performance of the three dictionaries is given below:

1. The Harvard IV Dictionary

   The confusion matrix for the Harvard IV Dictionary obtained in RapidMinner is as follows:

accuracy: 57.98%

	true 1	true neutral	true 0	class precision
pred. 1	23158	4121	2971	76.56%
pred. neutral	0	0	0	0.00%
pred. 0	11349	3826	7565	33.27%
class recall	67.11%	0.00%	71.80%	

   The Harvard IV Dictionary has 73 matching words and an accuracy of 57.98%

2. The AFINN Dictionary

   The confusion matrix for the AFINN Dictionary obtained in RapidMiner is as follows:

accuracy: 68.59%

	true 1	true neutral	true 0	class precision
pred. 1	31893	6756	6082	71.30%
pred. neutral	0	0	0	0.00%
pred. 0	2614	1191	4454	53.93%
class recall	92.42%	0.00%	42.27%	

   The AFINN Dictionary has 88 matching words and an accuracy of 68.59%

3. The Bing Liu Dictionary

   The confusion matrix for the Bing Liu Dictionary obtained in RapidMiner is as follows:

accuracy: 70.63%

	true 1	true neutral	true 0	class precision
pred. 1	32164	6550	5271	73.12%
pred. neutral	0	0	0	0.00%
pred. 0	2343	1397	5265	58.47%
class recall	93.21%	0.00%	49.97%	

The Bing Liu Dictionary has 100 matching words and an accuracy of 70.63%

Conclusion:

From the above analysis it can be concluded that the Bing Liu Dictionary gives the best accuracy of 70.63% compared to Harvard IV and AFINN Dictionaries and is therefore our best model.

ii)

SentiWordNet is a lexical resource in which each WordNet synset is associated to three numerical scores Obj(s), Pos(s) and Neg(s), describing how objective, positive, and negative the terms contained in the synset are. A typical use of SentiWordNet is to enrich the text representation in opinion mining (OM) applications, adding information on the sentiment-related properties of the terms in text.

Although a generalized dictionary like WordNet may be used, the accuracy of the classifier get affected due to issues like negation, synonyms, sarcasm, etc.

The issues are:

1. Negation words are the words which reverse the polarity of sentence. These are dealt with under negation handling. For example, in the text "this smart phone is not good", the negation word "not" reverses the polarity of sentence.

2. Word sense disambiguation is the lexicon ambiguity that may be syntactic or semantic. It refers to the words with more than one meaning that completely different. For example, "I love this movie" and "This is the love movie". In this the word "love" has different meanings.

(d) Develop models to predict review sentiment. For this, split the data randomly into training and test sets. To make run times manageable, you may take a smaller sample of reviews (minimum should be 10,000). One may seek a model built using only the terms matching any or all of the sentiment dictionaries, or by using a broader list of terms (the idea here being, maybe words other than only the dictionary terms can be useful). You should develop at least three different types of models (Naïve Bayes, and two others of your choice ….Lasso logistic regression (why Lasso?), knn, SVM, random forest,…?) (i) Develop models using only the sentiment dictionary terms (you can try individual dictionaries or combine all dictionary terms). Do you use term frequency, tfidf, or other measures? What is the size of the document-term matrix?

i)

KNN	K=3	81.69%	67.78%
	K=5	76.56%	69.87%
	K=7	77.87%	70.09%
Naïve Bayes	Laplace Correction	74.03%	73.56%
	Without Laplace Correction	74.03%	73.56%

Harvard Dictionary Model	**Parameters**	**Training Accuracy**	**Testing accuracy**
Logistic regression	Lambda = 0.1, Alpha = 0.1	71.29%	73.46%
	Lambda = 0.1, Alpha = 0.5	72.56%	73.49%
	Lambda = 0.5, Alpha = 0.1	72.56%	73.49%
KNN	K=3	77.62%	70.43%
	K=5	76.98%	71.90%
	K=7	74.57%	73.26%
Naïve Bayes	Laplace Correction	66.57%	64.58%
	Without Laplace Correction	66.57%	64.58%

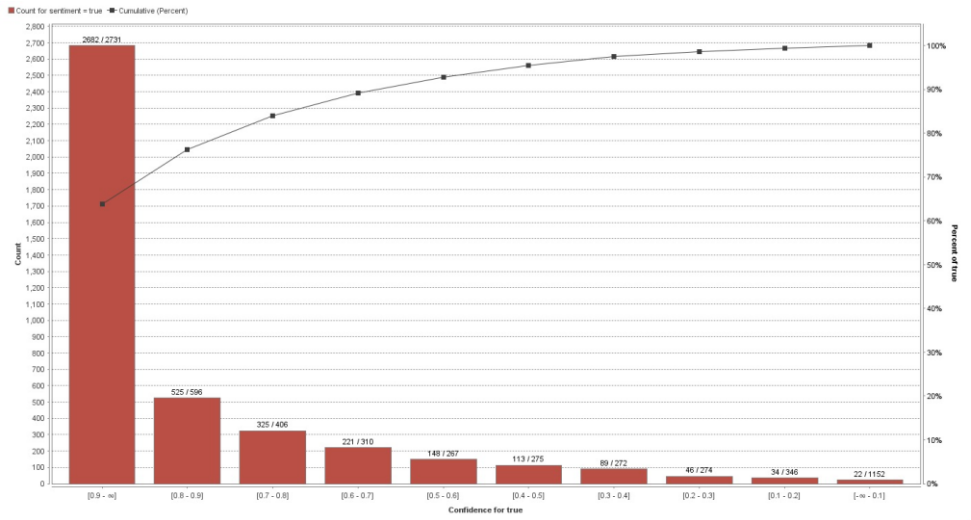


Figure: Logistic Lift Chart

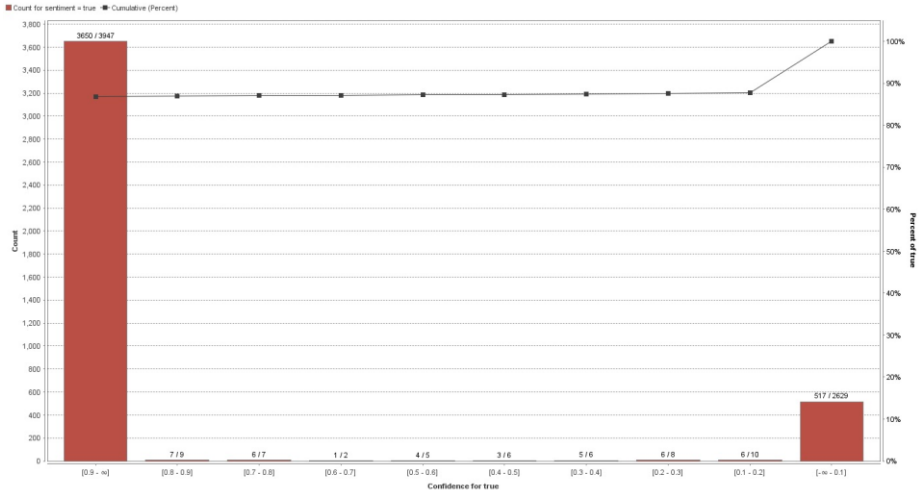


Figure: Naïve Bayes Lift Chart

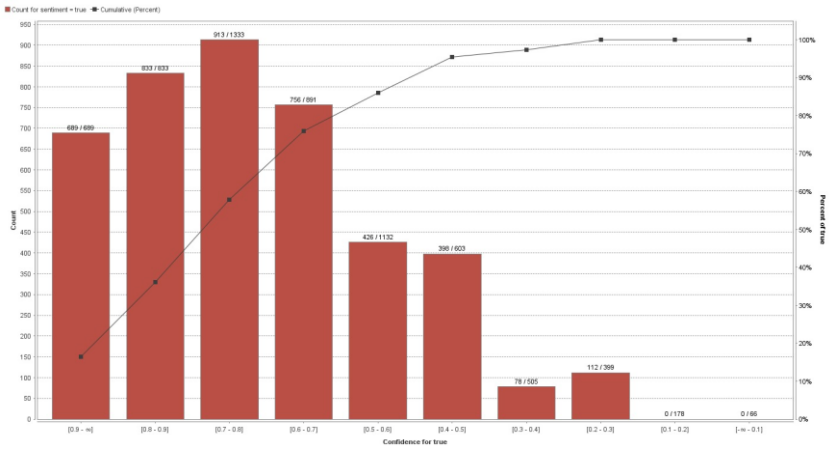


Figure: KNN Lift Chart

(ii) Develop models using a broader list of terms – how do you obtain these terms? Will you use stemming here? Report on performance of the models. Compare performance with that in part (c) above. For models in (i) and(ii): Do you use term frequency, tf-idf, or other measures, and why? Do you prune terms, and how (also, why?). What is the size of the document-term matrix?

Since we are using dictionaries, we will not be stemming as it will lead to inaccurate results.

We used the prune method with low 5% and 90%. We have calculated the accuracies using both TF-IDF and TF .

Model Used	Parameters		Train dataset Accuracy	Test dataset Accuracy
Logistic regression	Alpha = 1 (LASSO)		100%	66.46%
Naive Bayes			100%	73.21%
KNN	K= 5	TF-IDF	76.42%	69.71%

		TF	75.91%	65.72%

If we compare these above given results with part(c), we see that all the models -Logistic, KNN and Naïve Bayes performs better and they give better training and testing accuracies. In the performance, the model based approach should be used as compared to dictionary based to predict sentiment.