International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-3 Issue-11 September-2013

Feature based Opinion mining - towards Performance Measure

Ravikiran Kalava[1], G.Anil Kumar[2], Ch.Vasavi[3]

Abstract

In present era people are more depend on web for many activities like purchasing, investment, business makings etc. Because of this trend public are more dependent on other's opinion like booking a movie tickets or for purchasing electronic goods and many more things. Their final decision is totally influenced upon the blogs based on ratings, tags or comments obtained to that specific entity. Ideas and opinions of others always affect once own opinions. Opinion mining is a process in which it deals with opinions, sentiments and subjectivity of text .In this paper a research study is made upon a specific online shopping blog service and identified some specific features which are the main functional activities for the blog towards its performance(http://www.mouthshut.com/productreviews/Jabong-com-reviews-925660222) for classifying the performance measure based on opinions from the users are calculated using Naive Baye's Classifier and this approach gave a good experimental results.

Keywords

Opinion mining, sentiment, entity, features

1. Introduction

Decision-making is always a tough task in business perspective view because it depends on customer's activities, interest etc. In general customers have great impact of others opinions ,feelings ,talks in purchasing any item .This aspect has been became a major part for us today in performing any activity and much influenced in buying products, watching movie and in all types of investments. Opinions and its related concepts such as sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining, with the sudden growth of social media (e.g., reviews, forum discussions, blogs, comments, and postings in social network sites) on the Web, individuals and organizations are increasingly using the content in these media for decision making .Now a day's many companies, organizations are not even conducting any surveys, opinion polls or taking feedback from the customers, simply they are depending on these views and coming to a conclusion. Opinion mining extended its importance in the field of Electronic media in such a manner that decision makings are simply depend on the opinions got from the viewers ,analysts and from common people .

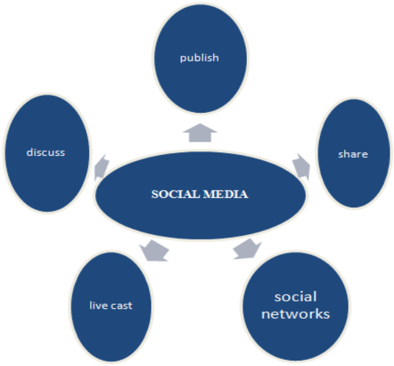


Figure 1: Social media landscape

In general solving an Opinion Mining Problem involves the following major stages:

- Defining the problem precisely.
- Selection and forming a proper training dataset.
- Transforming data to a specific format.
- Pre-processing data to increase the data quality.

Select an appropriate mining - method, includes

- Choosing a model or an algorithm
- Choosing the training parameters.
- Choosing a Proper Training/Testing data Validating and integrating the model.

2. Proposed Method

Naive-Baye's Classification

Ravikiran Kalava, Department of Computer Science and Engineering, R.V.R. Institute of Engineering and Technology, Hyderabad, Andhrapradesh, INDIA.
G.AnilKumar,Department of Computer Science and Engineering, R.V.R. Institute of Engineering and Technology, Hyderabad, Andhrapradesh, INDIA.
Ch.Vasavi , Department of Computer Science and Engineering, R.V.R. Institute of Engineering and Technology, Hyderabad, Andhrapradesh, INDIA.

We used Naive-Baye's classifier as it selects the mostly likely classification. It is also used for performing class membership probabilities prediction. The main advantage for Naive-Baye's is easy to construct and is mainly used for large dataset. Calculating of target class with respective of updating prior probability is done in posterior probability.

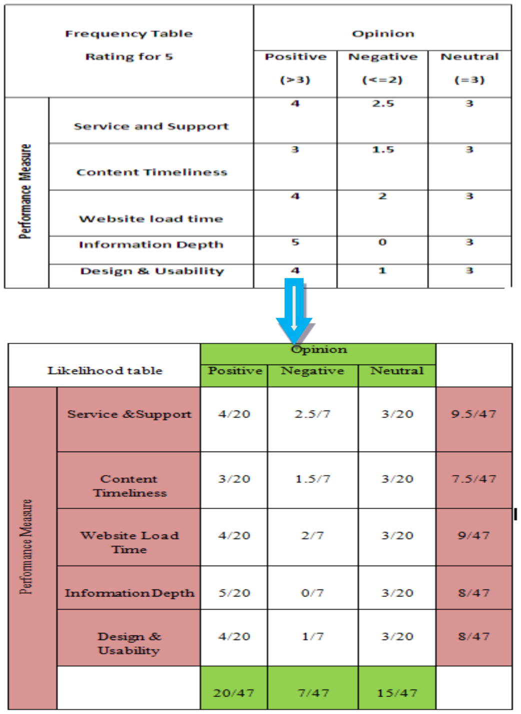
$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Where

$P(x|c) = P(x_1|c)*P(x_2 | c)* . . . P(x_n | c) * P(c)$

- P(c | x) - probability of instance 'c' (target class) being in 'x' (predicator).
- P(c) - Probability of occurrence of target class
- P(x) – prior probability of the predicator, same for all classes.
- P (x|c)–the likelihood which is the probability of predicator given class.

Posterior probability can be calculated by constructing a frequency table for each attribute against target class. Changing the table in to likelihood tables by calculating individual class mean values.



Frequency Table Rating for 5		Opinion		
		Positive (>3)	Negative (<=2)	Neutral (=3)
Performance Measure	Service and Support	4	2.5	3
	Content Timeliness	3	1.5	3
	Website load time	4	2	3
	Information Depth	5	0	3
	Design & Usability	4	1	3

Likelihood table		Opinion			
		Positive	Negative	Neutral	
Performance Measure	Service &Support	4/20	2.5/7	3/20	9.5/47
	Content Timeliness	3/20	1.5/7	3/20	7.5/47
	Website Load Time	4/20	2/7	3/20	9/47
	InformationDepth	5/20	0/7	3/20	8/47
	Design & Usability	4/20	1/7	3/20	8/47
		20/47	7/47	15/47	

Figure 2: Frequency table to Likelihood table

Some values with respective classes are taken from the above table for representation of posterior probability

P (x|c) = P (Service support | Positive) = 4/20=0.2

P (c) = P (Positive) =20/47=0.42

P (x) = P (service support)=9.5/47=0.20

Posterior probability:

P (c |x) =P(positive| service support) =0.2*0.42/0.20=0.42

The above value gives the positive opinion for one of the feature same can be calculated for all classes with respective of opinions.

Feature extraction

Feature based opinion mining mainly concentrates on specific features related to that particular entity taken in to consideration, based up on those features the entire performance capabilities are dependable. In our survey we went with an online shopping blog for performance measure, were specific features of the blog includes:-

1) Service and support.
2) Content timeliness.
3) Website load time.
4) Information depth.
5) Design and usability

In our research we felt that the above mentioned features help in measuring performance of a site as a good (positive), bad (negative) or none (neutral) which mean the opinion much better. In order to go with experimental we took some working principles in to consideration which helps for classifying. Opinion mining is a concept, in which it relates to the semantic of the web content available with respective of experiences, feelings, and actions and time. We are also considering some level of acceptance by identifying the semantic of some words in text related to opinion for experimental analysis. We mainly focused in identifying the grades of the users related to the each feature by allotting a grade value between (1-5) and assigning an opinion for each value as 1-2 as negative, =3 as neutral, >3 as positive. Calculated The result based on these criteria for each feature and based on the mean value the final opinion is calculated with respective of classifier.

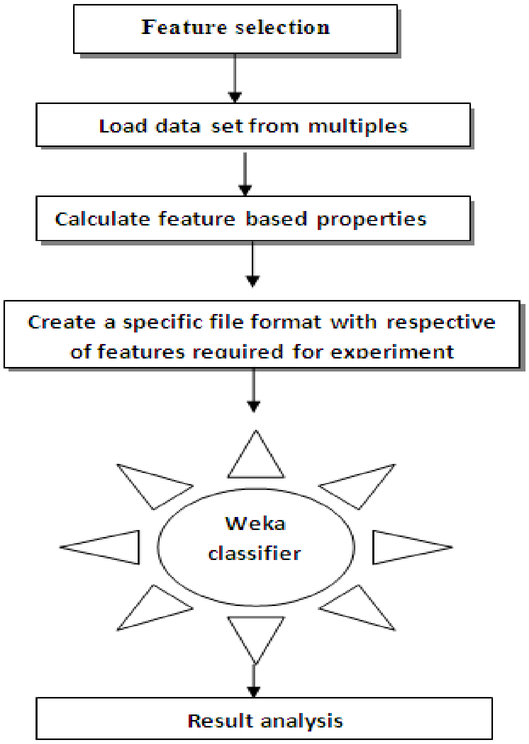

Feature selection
Load data set from multiples
Calculate feature based properties
Create a specific file format with respective of features required for experiment
Weka classifier
Result analysis

Figure 3: Steps involved in a process

3. Experimental Results

The main problem involved in opinion mining is to classify the opinion of the users with respective of three different measures positive, negative, neutral also termed in our experiments as good, bad, neutral . In order to implement the classification tool we used a data mining tool. The tool is called as Weka and it is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. ARFF (Attribute Relation File Format) is the specific format used in weka classifier for solving classification problems.

Sample ARFF File

@relation <relation name>
@attribute <Attribute name> <data type>
@attribute <Attribute name> <data type>
.   .   .
.   .   .
.   .   .
@data

Example
@relation rating
@attribute work performance {nominal or real etc}
@data

good or 4 [based on data type]

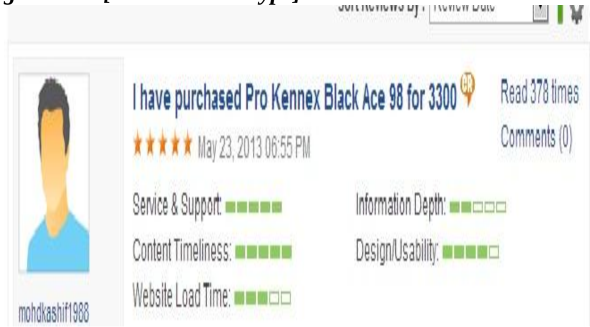
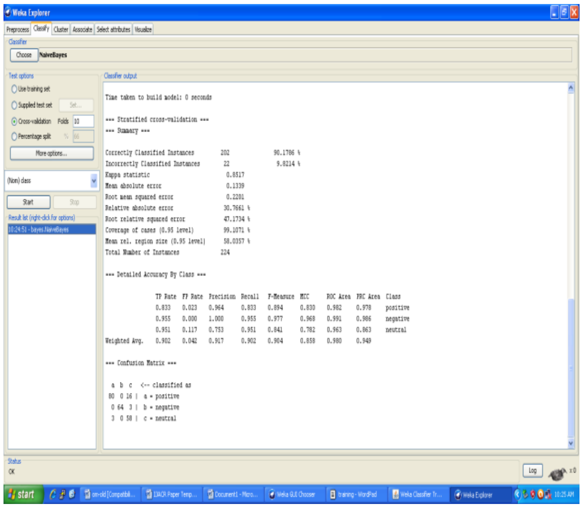

I have purchased Pro Kennex Black Ace 98 for 3300
★★★★★ May 23, 2013 06:55 PM
Read 378 times
Comments (0)
Service & Support: ■■■■■
Information Depth: ■■□□□
Content Timeliness: ■■■■■
Design/Usability: ■■■■□
Website Load Time: ■■■□□
mohdkashif1988

Figure 4: Individual feature rating for the blog

The above rating values for each individual feature is collected and maintained as a dataset from multiple reviewers, based on this data the opinion for each feature and final opinion can be calculated. In our experiment we used 220 reviewers' opinions and applied a Naive Baye's Classifier model for the dataset in order to find the class membership probabilities. Due to its high accuracy and very low error rate used in large data and the obtained results gave 90% accuracy.



Classifier: Choose NaiveBayes
Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66
(Nom) class
Result list (right-click for options): 10:24:51 - bayes.NaiveBayes
Status: OK

```
Classifier output

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         202               90.1786 %
Incorrectly Classified Instances        22                9.8214 %
Kappa statistic                          0.8517
Mean absolute error                      0.1339
Root mean squared error                  0.2201
Relative absolute error                 30.7661 %
Root relative squared error             47.1734 %
Coverage of cases (0.95 level)          99.1071 %
Mean rel. region size (0.95 level)      58.0357 %
Total Number of Instances              224

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.833    0.023    0.964      0.833    0.894      0.830    0.982     0.978     positive
                 0.955    0.000    1.000      0.955    0.977      0.968    0.991     0.986     negative
                 0.951    0.117    0.753      0.951    0.841      0.782    0.963     0.863     neutral
Weighted Avg.    0.902    0.042    0.917      0.902    0.904      0.858    0.980     0.949

=== Confusion Matrix ===

  a  b  c   <-- classified as
 80  0 16 |  a = positive
  0 64  3 |  b = negative
  3  0 58 |  c = neutral
```

Figure 5: Resultant output of the data for classifying opinion using Naive Baye's Classifier

Class	Precision	Recall	F-Measure	TP Rate	FP Rate
Positive	0.964	0.833	0.894	0.833	0.023
Negative	1.000	0.955	0.977	0.955	0.000
Neutral	0.753	0.951	0.841	0.951	0.117

Figure 6: Accuracy of Samples

Precision and Recall are the two terms related to information retrieval concepts. Precision gives the probability of similarity or relevant from retrieved documents. Recall is the ratio of relevant documents found in the search result to the total of all relevant documents. From our experiment with respective of data set we obtained the values as Precision of 0.964 means 96.4% of the documents were relevant. Recall of 0.833 means 83.3% relevant documents retrieved from all documents.

Experiment with J48 Classifier:

Tree representation is used in showing a data object in hierarchical manner J48 model is used as decision tree classifier. The general approach of the algorithm chooses an attribute of the data that most effectively splits its sample of data into subsets enriched in one class or the other. The criteria used to choose an attribute for splitting is information gain, which is the difference in the entropy values resulting from choosing an attribute for splitting the data. We got the correctly classified instance rate as 98.3% and incorrectly classified instance rate as 1.7%.

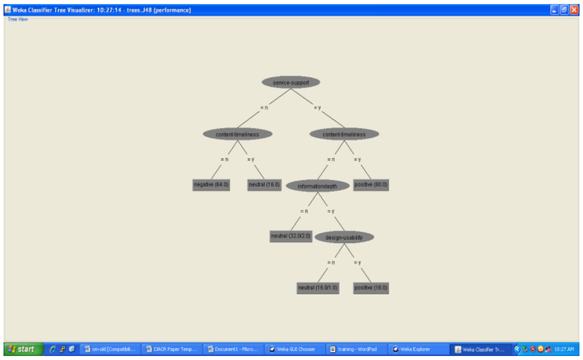


Figure 7: Decision tree generated by the classifier

4. Conclusion

Performance measure based on the opinions is a task purely related to the semantic of the words and value based. Business oriented many blogs maintaining and following reviews of the customers for the better improvement of their services, but the problem associated is how reviews are helping customers for their decision making towards any business deal .just depending on the reviews is not enough because of review spam .Automated or dynamic performance measure tool with respective of comments, opinions ,ratings all put together will drop down the problems of present manual work.

References

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

[2] Alekh Agarwal and Pushpak Bhattacharyya, Sentiment Analysis: A New Approach for Effective Use of Linguistic Knowledge and Exploiting Similarities in a Set of Documents to be Classified, International Conference on Natural Language Processing (ICON 05),IIT Kanpur, India, December, 2005.

[3] Alec, G.; Lei, H.; and Richa, B. Twitter sentiment classification using distant supervision. Technical report, Standford University. 2009.

[4] B. Liu. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Second Edition, Springer, July 2011.

[5] Liu, Bing, Sentiment Analysis and Opinion Mining, 5th Text Analytics Summit, Boston, June 1-2, 2009.

[6] Eamonn Keogh: Pattern Recognition and Machine Learning,Christopher Bishop, Springer-Verlag, 2006.

Ravikiran Kalava -Completed Masters in Software Engineering and Bachelors in Information Technology from JNTU Hyderabad associated institution. Presently working as Assistant Professor Department of Computer science and Engineering. He is a member in International Association of Computer science and Information Technology, International Association of Engineers. His research areas include Data mining, Cloud computing, semantic web.

G.AnilKumar-Completed Masters in Computer science and Technology from Andhra University, Bachelors in Computer science and Engineering from JNTU. Pursuing PhD from Andhra University. Presently working as Professor, Head Department of Computer science and Engineering. He is a member in International Association of Computer science and Information Technology, International Association of Engineers. His research areas include Data mining applications in image processing , Medical image processing.

Ch.Vasavi -Completed Masters in Software Engineering and Bachelors in Information Technology from JNTU-Hyderabad associated institution .she has 3 years of teaching experience presently working as Assistant Professor in CSE Department. Her area of interest includes Data mining, Data security, Cloud computing.