SENTIMENT CLASSIFICATION OF MOVIE REVIEWS BY SUPERVISED MACHINE LEARNING APPROACHES

P.Kalaivani
Research Scholar, Sathyabama university,
Chennai,India
vaniraja2001@yahoo.com

Dr. K.L.Shunmuganathan
Professor & Head
R.M.K Engineering College,
Chennai,India

Abstract

Large volumes of data are available in the web. The discussion forum, review sites, blogs and news corpora are some of the opinion rich resources. The opinions obtained from those can be classified and used for gathering online customer's preferences. Techniques are being applied to design a system that identifies and classify opinions spread largely in the internet. Few different problems such as sentiment classification, feature based classification and handling negotiations are predominating this research community. This paper studies online movie reviews using sentiment analysising approaches. In this study, sentiment classification techniques were applied to movie reviews. Specifically, we compared three supervised machine learning approaches SVM, Navie Bayes and kNN for Sentiment Classification of Reviews. Empirical results states that SVM approach outperformed the Navie Bayes and kNN approaches, and the training dataset had a large number of reviews, SVM approach reached accuracies of atleast 80%.

Keywords: opinions, sentiment classification, online reviews, supervised machine learning algorithm.

I. INTRODUCTION

In recent years, the problem of "sentiment classification" has been increasing attention [3] .Large volumes of reviews, rating, recommendations, and news corpora, online opinion could provide important information for business to market their product. Sentiment analysis helps in tracking the mood of the public about a particular product or object. Sentiment analysis also called opinion mining involves building a system that gathers opinion and examines it. This would prove useful in judging the success of a new product launched, how a new version of a product is received by the customer and the likes and dislikes constrained to a particular area.

The difficulties in sentiment analysis are an opinion word which is treated as denoting a positive side may be considered as negative in another situation. Second, people's way of expressing a situation varies. The traditional text processing considers that a small change in two pieces of text has no change in the meaning. But in sentiment analysis, "the picture is good" is different from "the picture is not good". Most statement contains both positive as well as negative opinions. The system processes it by analyzing one sentence at a time. However, blogs and twitter contains more informal statements which are easy to understand by the user and not by the system. For example," that movie was as good as its last movie" is dependent on another object whose description is not available.

The traditional text mining concentrates on analysis of facts whereas opinion mining deals with the attitudes [3]. The main fields of research are sentiment classification, feature based sentiment classification and opinion summarizing. Sentiment classification analyses the opinions on a certain object. Feature based classification focuses on analysing and classifying based on the features of the object [4]. Opinion summarizing is different from the classic text summarizing by the fact that only the feature of the product that customers have expressed their opinions are mined rather than considering a subset of a review and rewriting some of the original statements to capture the main idea.

The rest of the paper is organized as follows. Section 2 discusses related work of our study. Section 3 presents data source and data set used in this study. Section 4 presents models and methodology. Section 5 presents Experiments. Empirical results and discussing are given in section 6. Finally, section 7 concludes the paper.

II.RELATED WORK

Several techniques were used for opinion mining tasks in history. The following few works are related to this technique. Janice M.Wiebi [1] uses the data of review from automobiles, bank, movies, and travel destinations. He classified words into two classes (positive or negative) and counts on overall positive or negative score for the text. If the documents contain more positive than negative terms, it is assumed as positive document otherwise it is negative. These classifications are based on document and sentence level classification. These classifications are useful and improve the effectiveness of a sentiment classification but cannot find what the opinion holder liked and disliked on each feature.

To extract opinions, machine learning method and lexical pattern extraction methods are used by many researchers. Turney [2] introduced the results of review classification by considering the algebraic sum of the orientation of terms as respective of the orientation of the documents. He determined the similarity between two words by counting the number of results returned by web searches. The relationship between a polarity unknown word and a set of manually-selected seeds was used to classify the polarity-unknown word into a positive or negative class.

Pang et al [3], Mukras R.J [4] use the data of movie review, customer feedback review and product review. They use the several statistical feature selection methods and directly apply the machine learning techniques. These experience show that machine learning algorithm only is not well perform on sentiment classification. They show that the present or absent of a word seems to be more indicative of the content rather than the frequency of a word.

Morie Rimon [5] used the keyword based approach to classify the sentiment. In this approach, terms, mainly adjectives (e.g. awesome, awful) are used as sentiment indicators. The list of indicators can be prepared manually, composed semi automatically using sources such as WordNet or acquired by machine learning algorithms that infer the best indicators from tagged samples in the domain of interest.

Alec co [6] used the different machine learning classifiers and feature extractors to classify sentiment. The machine learning classifiers are Naive Bayes, Maximum Entropy and Support Vector Machines (SVM). The feature extractors are unigrams, bigrams, unigrams and bigrams, and unigrams with part of speech tags.

Changli Zhang [7], use the data of customer feedback review and product review. They use Decision learning method for sentiment classification. Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Learned trees can also be re-represented as sets of if-then rules to improve human readability. These learning methods are among the most popular of inductive inference algorithms and have been successfully applied to a broad range of tasks from learning to diagnose medical cases to learning to assess credit risk of loan applicants.

Kudo et al. [8] used sub trees of word dependency trees as features for sentence wise sentiment polarity classification. He used boosting algorithm with the sub tree-based decision stamps as weak learners.

Dave et al. [9] used machine learning methods to classify reviews on several kinds of products. Unlike Pang's [3] research, they obtained the best accuracy rate with word bigram-based classifier on their dataset. This result indicates that the unigram-based model does not always perform the best and that the best settings of the classifier are dependent on the data.

To use the prior knowledge besides a document, Mullen and Collier [10] at tempted to use the semantic orientation of words defined by Turney [2] and several kinds of information from Internet and thesaurus. They evaluated on the same dataset used in Pang et [3] study and achieved 75% accuracy with the lemmatized word unigram and the semantic orientation of words.

Yan Dang , Yulei Zhang and Hsinchun Chen [11] proposed a lexicon enhanced method for sentiment classification by combining Machine learning and semantic orientation approaches into one framework. Specifically, they used the words with semantic orientations as an additional dimension of features for the machine learning classifiers.

III. DATA SOURCE & DATA SET

User’s opinions are the valuable sources of data which helps to improve the quality of service rendered. Blogs, review sites and micro blogs are some of the platforms where user expresses his/her opinions.

To conduct the research, Movie reviews are considered here. Some of them are available at www.cs.cornell.edu/People/pabo/movie-review- data, multi-domain reviews are available at www.cs.jhu.edu/mdredze/datasets/sentiment. the mdr contains reviews on books, DVDs, electronics and kitchen appliances with 1000 reviews each for positive and negative.

IV. METHODOLOGY

Sentiment analysis is conducted at any of the three levels: the document level, sentence level or the attribute level. In this study, we applied three supervised machine learning models for sentiment classification of reviews for the selected movie reviews. These models are Naive Bayes (NB), and support vector machines and K-Nearest neighbourhood.

Naïve Bayes

In Naïve Bayes technique , the basic idea to find the probabilities of categories given a text document by using the joint probabilities of words and categories. It is based on the assumption of word independence.

The starting point is the Bayes theorem for conditional probability, stating that, for a given data point x and class C:

$$P(C/x)=\frac{P(x/C)\cdot P(C)}{P(x)}$$

Furthermore, by making the assumption that for a data point x = {x1,x2,…xj}, the probability of each of its attributes occurring in a given class is independent, we can estimate the probability of x as follows

$$P(C/x)=P(C)\cdot\prod P(x_j/C)$$

Training a Naïve Bayes classifier therefore requires calculating the conditional probabilities of each attributes occurring on the predicted classes, which can be estimated from the training data set. Naïve Bayes classifiers often provide good results, and benefit from the easy probabilistic interpretation of results. To provide sentiment classification of online reviews about movie was selected as feature selection technique in this studt.

Support vector machine

Support vector machine is a method for classification of linear data. It uses a non linear mapping to transform the training data into a higher dimension. Within the higher dimension it searches for the linear optimal separating hyper plane. Data from two classes can always be separated by a hyper plane. The SVM finds hyper plane using support vectors. This approach was developed by Vladimir vapnik, Bernhard Boser and Isabelle Guyan in 1992. A support vector machine (SVM) is considered the highly effective at traditional text classification method.

In the simple case where a linear function divides the two classes, a resulting hyperplane partitions the solution space. The following graph illustrates dividing hyperplanes for a sample of points belonging to 2 classes:

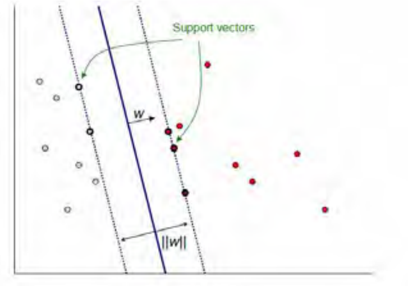

Figure 1 - Hyperplane Separating Two Classes

In the above example, there is a potentially unlimited number of separating hyper planes dividing the two classes. In choosing the best possible one, an intuitive idea would be to choose a hyper plane that has the largest distance between any points from either class, thus creating the widest possible margin between points from the 68 classes. The intuition behind this method is that a hyper plane with a large margin would be a “safer” classification boundary, less likely to make prediction errors by being to close to the boundaries of one of the classes. Finding such hyper plane is the objective of the Support Vector Machine algorithm, to achieve this, the problem can be formalised as finding the vector w, such that:

Let c_j ε {1, -1} (corresponding to positive and negative) be the correct class of documents d_j.For classes C1 and C2, and feature data vectors X = { x_t, r_t } where:

$R_t = +1$ *if* $x_t * C1$ *and*

$R_t = -1$ *if* $x_t * C2$

Find w and constant w0, such that the dot product of w and a given data vector is as follows:

$$wt * x_t + w_0 >= +1 \ if \ x_t \ \text{€}C1, \text{ and}$$
$$wt * x_t + w_0 >= -1 \ if \ x_t \ \text{€}C2$$

The above equations state that points belonging to class C1 and C2 are on separate sides of the orthogonal hyper plane defined by the vector w, and by maximizing the vector length ||w||, we obtain a dividing hyper plane with maximum distance between points from either class. It is demonstrated in (Boser et al, 1992) that finding this optimal hyper plane translates to a quadratic optimization problem, and whose complexity depends on the number of training vectors N, but not on the dimensionality of the data set. An interesting result is that the model obtained from training support vector machines takes into account only the data points close to the dividing hyper plane for predictions: these are called the support vectors, and are expected to be in much smaller number than the entire data set, thus providing an algorithm with good performance during execution time.

Nearest Neighbour Method

Nearest neighbour methods are considered one of the simplest and most yet effective classes of classification algorithms in use. Their principle is based on the assumption that, for a given set of instances in a training set, the class of a new yet unseen occurrence is likely to be that of the majority of its closest "neighbour" instances from the training set. Thus the k-Nearest Neighbour algorithm works by inspecting the k closest instances in the data set to a new occurrence that needs to be classified, and making a prediction based on what classes the majority of the k neighbours belong to. The notion of closeness is formally given by a distance function between two points in the attribute space, specified a priori as a parameter to the algorithm. An example of distance function typically used is the standard Euclidean distance between two points in an n-dimensional space, where n is the number of attributes in the data set

V .EXPERIMENTS

We conducted K-fold-cross-validation (Kohavi, 1995) in the experiments. In this research, K = 3 was adopted. The 1000 positive and 1000 negative reviews were applied to make a 3-fold cross validation in the data experiments. On each round of experiment, two folds were used a training data set, and the remaining fold was used as the testing data set. As stated, an objective of this study was to examine how a classifier works with various sizes of training data set. It was, therefore, important to create small subsets from a given large training set. Let T and Ts, respectively, be the training and testing datasets of each round. We further split training data set A into 10 disjoint sets (T1,T2, T3, . . ., T10), not necessarily of equal size, and then 10 new training sets (TT1,TT2,TT3,. . .,TT10) are constructed, where TT1 = T1, TTi = TTi_1 + Ti (i = 2,. . ., 10). The performance of a classifier could be assessed based on the results of 10 experiments conducted on 10 train-test pairs (TTi,Ts). We conducted each round of experiment by increasing the number of training examples with each experiment; this was repeated 3-fold. Table 1. shows the number of training examples of the categories in each round.

Table 1. Numbers of reviews in training data sets

No of experiments	Corpus Numbers of reviews in each round of training		
	Positive	Negative	All
1.	50	50	100
2.	100	100	200
3.	150	150	300
4.	200	200	400
5.	400	400	800
6.	550	550	1100
7.	650	650	1300
8.	800	800	1600
9.	900	900	1800
10.	1000	1000	2000

VI. PERFORMANCE EVALUATIONS

Accuracy, Precision and recall are method used for evaluating the performance of opinion mining. Here accuracy is the overall accuracy of certain sentiment models. Recall (Pos) and Precision (Pos) are the ratio and precision ratio for true positive reviews. Recall (Neg) and Precision (Neg) are the ratio and precision ratio for true negative reviews. In an ideal scenario, all the experimental results are measured according to the Table 2 and accuracy, Precision and recall as explained below.

$$Accuracy = \frac{a+d}{a+b+c+d}$$

$$\text{Re}\,call\,(Pos) = \frac{a}{a+c},\ \text{Re}\,call\,(Neg) = \frac{d}{b+d}$$

$$\text{Pr}\,ecision(Pos) = \frac{a}{a+b},\ \text{Pr}\,ecision(Neg) = \frac{d}{c+d}$$

Among all, movie review mining is the challenging task because of the fact that movie reviews are written with mixed real-life review data and ironic words. Product reviews are entirely different from movie reviews by two ways. Firstly, product reviews mostly contribute to the features which may be liked by some and disliked by some. Thus a review is either positive or negative depending on user's preference thereby making it difficult to categorize. Secondly, there also exists comparative statements i.e to compare other products and state it.

Table 2.Contingency table for performance evaluations

	True positive reviews	True negative reviews
Predict positive	a	b
Predict negative	c	d

The overall accuracies of the three algorithms in 10 rounds of experiments are indicated in Table 3 and Fig. 3. The result indicated that the SVM approach and N-gram approach had better accuracies than the Naïve Bayes approach, when the training data set had 150 or less reviews.

The precision for positive corpus in the testing dataset were showed in the table 4 and Figure 4.

The precision for Negative corpus in the testing dataset were showed in the table 5 and Figure 5.

Table 3. Accuracies in testing data set

No of experiments	Numbers of reviews in training dataset	Accuracy		
		SVM Classifier (%)	Navie Bayes (%)	k-NN (%)
1.	50	60.07	56.03	64.02
2.	100	61.53	55.01	53.97
3.	150	67.00	56.00	58.00
4.	200	70.50	61.27	57.77
5.	400	77.50	65.63	62.12
6.	550	77.73	67.82	62.36
7.	650	79.93	64.86	65.46
8.	800	81.71	68.80	65.44
9	900	81.61	71.33	67.44
10.	1000	81.45	72.55	68.70

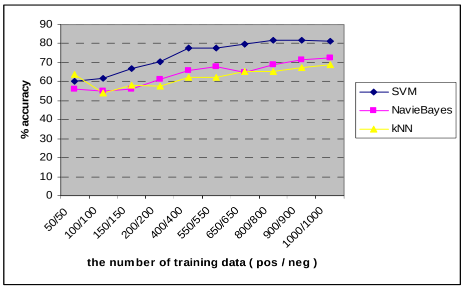

Fig. 3. Diagrammatic presentation of accuracies in the experiments

While comparing the accuracy of classification with respect to feature selection method, Information Gain (IG), SVM approach gave the best result. Generally speaking the SVM approach and NavieBayes approach had achieved better performance than k-NN approach. When the number of feature selected about 500 the difference among the algorithms (SVM & NB) was extremely significant.

Table 4.Precisions for positive corpus in testing data set

No of experiments	Numbers of reviews in training dataset	Precision on positive corpus		
		SVM Classifier (%)	Navie Bayes (%)	k-NN (%)
1.	50	60.87	54.84	61.29
2.	100	59.35	54.17	53.85
3.	150	64.41	54.84	57.50
4.	200	68.47	60.27	57.56
5.	400	73.61	64.64	61.74
6.	550	73.21	68.22	62.36
7.	650	75.29	68.52	64.76
8.	800	78.94	70.66	64.91
9	900	78.77	72.17	66.12
10.	1000	79.48	73.27	67.35

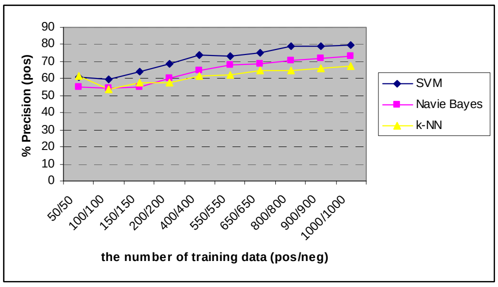

Fig. 4. Diagrammatic presentation of precisions for positive corpus.

Table 5 Precisions for negative corpus in testing data set

No of experiments	Numbers of reviews in training dataset	Precision on negative corpus		
		SVM Classifier (%)	Navie Bayes (%)	k-NN (%)
1.	50	59.26	57.89	68.42
2.	100	64.94	56.25	54.17
3.	150	70.73	56.16	58.57
4.	200	73.03	62.43	57.95
5.	400	82.93	66.76	62.53
6.	550	84.42	67.44	62.36
7.	650	86.63	68.40	66.24
8.	800	85.22	67.14	66.07
9	900	85.08	70.56	69.01
10.	1000	83.71	71.87	70.28

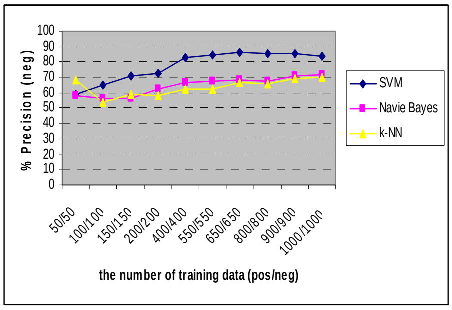

Fig. 5. Diagrammatic presentation of precisions for negative corpus.

The Recall for positive corpus in the testing dataset was showed in the table 6 and Figure 6.

While comparing the precision and recall for positive corpus with respect to Information Gain feature selection model, gave the best result when the number of reviews in training dataset more than 200.

Table 6.Recall for positive corpus in testing data set

No of experiments	Numbers of reviews in training dataset	Recall on positive corpus		
		SVM Classifier (%)	Navie Bayes (%)	k-NN (%)
1.	50	56.00	68.00	76.00
2.	100	73.00	65.00	56.00
3.	150	76.00	57.33	61.33
4.	200	76.00	66.00	59.00
5.	400	85.75	69.00	63.75
6.	550	87.45	66.73	62.36
7.	650	89.08	68.31	67.85
8.	800	87.12	65.62	69.12
9	900	86.56	69.44	71.56
10.	1000	84.80	71.00	72.60

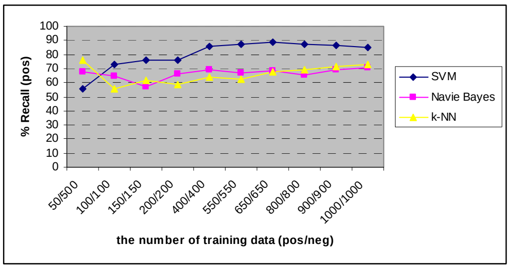

Fig. 6. Diagrammatic presentation of recalls for positive corpus

The Recall for Negative corpus in the testing dataset were showed in the table 7 and Figure 7.

Table 7.Recall for Negative corpus in testing data set

No of experiments	Numbers of reviews in training dataset	Recall on Negative corpus		
		SVM Classifier (%)	Navie Bayes (%)	k-NN (%)
1.	50	64.00	44.00	52.00
2.	100	50.00	45.00	52.00
3.	150	58.00	54.67	54.67
4.	200	65.00	56.50	56.50
5.	400	69.25	62.25	60.50
6.	550	68.00	68.91	62.36
7.	650	70.77	68.62	63.08
8.	800	76.15	72.05	61.67
9	900	76.67	73.22	63.33
10.	1000	78.10	74.10	64.80

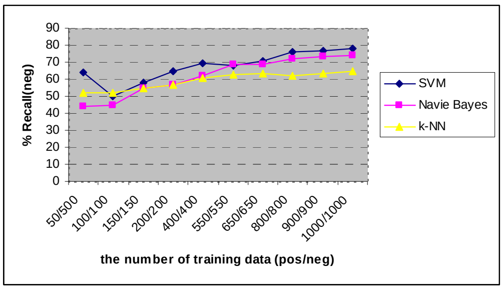

Fig. 7. Diagrammatic presentation of recalls for negative corpus

This study has applied three supervised machine learning algorithms of SVM, NavieBayes and kNN to the online movie reviews. We observed that well trained machine learning algorithms can perform very good classifications on the sentiment polarities of reviews about movies. In terms of accuracy, SVM algorithm can reach more than 80 % of the classification correctly. When the training data set was as small as 50,150 or 200 reviews, the difference among the NavieBayes and kNN was extremely significant. A large training dataset with 800 to 1000 reviews will perform better in sentiment classification for all three algorithms for the reviews about movie reviews.

VII. CONCLUSIONS

The aim of study is to evaluate the performance for sentiment classification in terms of accuracy, precision and recall in this study, In this paper, we compared three supervised machine learning algorithms of SVM, NavieBayes and kNN for sentiment classification of the movie reviews that contains 1000 positive and 1000 negative processed reviews. The experimental results show that the SVM approach outperformed than the NavieBayies and k-NN approaches and the training dataset had a large number of reviews, the SVM approach reached accuracies of more than 80%.

REFERENCES

[1] J. Wiebe, et al (2004)., “Learning Subjective Language,” The Association for Computational Linguistics, vol. 30, no. 3, pp. 277-308.

[2] P.D. Turney (July 2002) “Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” Proc. of the 40th Annual Meetings of the Association for Computational Linguistics, ACL Press, pp 417-424.

[3] B. Pang, et al (July 2002)., “Thumbs up? Sentiment Classification Using Machine Learning Techniques,” Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL Press, pp 79-86.

[4] R. Mukras, J. Carroll (2004). A comparison of machine learning techniques applied to sentiment classification, pp 200-204.

[5] Mori Rimon (2004), “Sentiment Classification: Linguistic and Non-linguistic Issues”, pp 444-446.

[6] Alec Go (2005), “Twitter Sentiment Classification using Distant Supervision”, Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, CA.

[7] Changli Zhang, Wanli Zuo, Tao Peng,Fengling He (2008), “Sentiment Classification for Chinese Reviews Using Machine Learning Methods Based on String Kernel”, Third 2008 International Conference on Convergence and Hybrid Information Technology.

[8] Kudo et al (2001). “An operational system for detecting and tracking opinions in on-line discussion”. In SIGIR Workshop on Operational Text Classification, pp 449-454.

[9] Dave K Lawrence, Pennock (2003), D. M. Mining, “opinion extraction and semantic classification of product reviews”, In Proceedings of the 12th international WWW conference, pp. 519-528, Hungary.

[10] Tony Mullen and Nigel Collier (July 2004) , “Sentiment analysis using support vector machines with diverse information sources”. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 412-418, Barcelona, Spain.

[11] Yan Dang, Yulei Zhang, Hsinchun ChenA (July 2010), “Lexicon Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews”, Department of Management Information Systems, vol. 25, no. 4, pp. 46-53.

BIOGRAPHY

P. Kalaivani**,** B.E., M.E., works as Associate Professor in St.Joseph’s College of Engineering, Chennai, and Tamilnadu, India. She has more than Eight years of teaching experience. Her areas of specializations are Data structures, Data mining, Web mining and Artificial Intelligence.

Dr. K. L. Shunmuganathan**,** B.E, M.E., M.S., Ph.D., works as the Professor and Head of the CSE Department of RMK Engineering College, Chennai, and Tamilnadu, India. He has more than 18 years of teaching experience, and his areas of specialization are Artificial Intelligence, Computer networks and DBMS.