Middle-East Journal of Scientific Research 24 (Techniques and Algorithms in Emerging Technologies): 341-345, 2016
ISSN 1990-9233;

DOI: 10.5829/idosi.mejsr.2016.24.TAET23337

Performance Analysis of Supervised Machine Learning Algorithms for Opinion Mining in E-Commerce Websites

[1]N. Yuvaraj and [2]A. Sabari

[1]CSE, KPR Institute of Engineering and technology, Coimbatore, India
[2]IT, KS Rangasamy College of Technology, Tiruchengode, India

Abstract: Online shopping websites have become popular as they allow customers to purchase the products easily from home. These websites requests the customers to rate the quality of the products. To enhance the quality of the products and services these reviews provides different features of the products. In order to buy a product a customer has to go through large number of reviews, which is quite hard for the customers as well as the marketers to maintain, keep track and understand the views of the customers for the products. This paper explains about various machine learning techniques available for Feature Extraction and also to carry out Sentiment analysis in order to enhance the product's quality and Customer services.

Key words: Supervised methods · Support vector machine · Naive Bayes · Nearest neighbor

INTRODUCTION

Data analytics is the process of examining a large dataset for the purpose of getting a conclusion about that information. Data Analytics in used in various fields to make better decisions and to verify or to disprove existing models or theories. It also refers to the quantitative and qualitative techniques and process which is used to enhance the productivity and business gain. Data is being extracted and categorized to identify and analyze its behavioral data and also the patterns according to the organization's requirements. Sentiment Analysis or Opinion mining is one of the emerging research fields in Data Analytics. It is one of the popular techniques which is used for analyzing and summarizing the customer's reviews about the products and the services. Some of the online shopping websites such as Amazon, Flipkart, EBay, etc contains the product reviews which enable the customers to go through the reviews while purchasing the product. Sentiments are extracted from the feedbacks of the products given by the users from the shopping websites. Extraction of opinion includes identification of opinion holder, subject of the review and the responses such as positive or negative.

Supervised Machine Learning Algorithms: Supervised machine learning algorithms takes the known set of data as an input and produces the output based on that data. In supervised learning technique the data is divided into two categories namely the training set and the test set. The machine is trained using the historical records or events and produces the prediction results according to the trained data. Then the quality of the training set is evaluated using the test set [1]. Some of the supervised learning algorithms used are SVM, Naïve Bayes, Decision Trees, Nearest Neighbor algorithm [2]. In particular the types of supervised machine learning algorithms used only for sentiment analysis are:

- Support vector machine
- Naïve Bayes
- Nearest neighbour

In this paper the above machine learning algorithm's working and performance are discussed and the accuracy of each algorithm is compared to find out the best suitable algorithm for Opinion mining.

Support Vector Machine: Support vector machine has the predefined input and the output format where the input is a vector space and the output will be either 0 for negative or 1 for positive [3, 4]. Using the pre-processing techniques the documents which are in original format is converted into machine learning input. The best classification of data techniques is supported by SVM [5]. SVM is one of the learning algorithms which provides

Corresponding Author: N. Yuvaraj, CSE, KPR Institute of Engineering and technology, Coimbatore, India.

best learning algorithm for text categorization. For 1000 features SVM provides High Dimensional Input Space by applying over fitting protection that does not depends on the number of features, which provides high ability to handle the features in large numbers [3]. SVM provides sparse Document Vector Space since the document vectors contains about very few non-zero elements [6]. SVM's task is to generalize the mapping of input-output.

SVM Performance Evaluation: For the text categorization, the input dataset will be the documents and the generated output will be the class of the respective documents. For example, In spam filtering the input will be the e-mail and the output will be 0's and 1's which indicates the negative as Spam and the positive as Not-Spam. The basic objective behind the SVM classification is to find out the hyper plane [7] that has maximum margin which separates the document vector of one class from the document vector of another class with maximum margin.

Linear classification is applied for 100000 iterations to classify the reviews either as positive or negative [8]. On experimenting the results shows that the Term-Frequency and the Inverse-Document-Frequency (TF-IDF) gave the maximum accuracy for the linear SVM classification [9]. The accuracy results for linear SVM classifier are tabulated.

	Accuracy	Precision	Recall
Term- Occurance	75.25	90.40	56.50
Term-Frequency	84.00	86.17	81.00
Binary-Term occurence	76.50	90.15	59.50
TF-IDF	84.75	82.63	88.00

On experimenting the Linear SVM with the data set containing 10001 positive and negative reviews it gave the best prediction results for about 9767 negative reviews correctly with an accuracy of about 97.66%.

SVM constructs a set of hyper planes or a set of hyper-planes with a high or infinite-dimensional space. A hyperplane having largest distance to the nearest training data points provides the good separation ie., larget the margin the lower will be the generalization error of the classifier. For the dataset containing 4000 positive and 4000 negative reviews and the dataset containing 1000 positive and 1000 negative reviews SVM works as follows [10, 8],

For a training set D, a set n points can be written as,

$$D = \{(x_i, c_i) \mid X_i \varepsilon R^p, C_i \varepsilon \{-1,1\}\}_{i-1}^{n} \quad (1)$$

where xi is a p-dimensional real vector. Then find the maximum-margin hyper plane for the points ci = 1 and ci = -1 where {1,-1} corresponds to positive or negative data [11], which can be given by,

$$w.x-b = 1 \text{ and } w.x-b = -1 \quad (2)$$

the distance between the planes is given by, $\frac{a}{\|w\|}$ in which ||w|| has to be minimized.

Using Lagrange's multipliers the optimization problem can be expressed as,

$$\min_{w,\ b} \max_{\alpha} \{\{\{\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [C_i (w.x_i - b) - 1]\}\}\} \quad (3)$$

From the equation 2 & 3, the ái should be greater than zero because they are the document vectors corresponding to w. Using this method SVM determines the hyperplane's location corresponding to the w's side ie.,{1,-1}[12]. SVM gave the accuracy of about 82.9% and also it provides the low dataset dimensionality.

Naive Bayes Classifier: Naive Bayes is a probabilistic classifier which is based on the Bayes Theorem. In the document classification process, the performance of work can be classified into two stages namely the learning stage and the classification stage. In a document the words that express the sentiments are choosen from the sample document which is done as a classification process in the learning stage [8]. This classification is used to determine the polarity of a document ie., positive or negative using the previous evidences [10, 13]. The conditional independence of Naïve Bayes classifier makes the data to train faster [14, 15, 7]. It assumes all the vectors in the feature vectors as independent and applies the Bayes rule in the sentence. The prior probability frequency for each label in the training set is also calculated [16]. This is used to find whether the event will occur. For an event X and the evidence Y, the probability can be written as

$$P(X|Y) = P(X)P(Y|X)/P(Y).$$

Naive Bayes Classifier's Performance Evaluation: On experimenting Naive Bayes classification with 100000 iterations to classify the reviews either as positive or negative [8]. It is found that the term-occurrence and the binary term occurrence gave the maximum accuracy [9]. It is found that the same results were obtained when the classifier is tested for 1000 positive and negative reviews.

	Accuracy	Precision	Recall
Term-occurance	70.00	70.41	69.00
Term-Frequency	68.50	68.69	68.00
Binary term occurance	70.00	70.41	69.00
TDF-IDF	67.50	66.51	70.50

The conditional probability of a sentiment will be given as,

$$P(sentiment \mid sentence) = \frac{P(sentiment)P(sentence \mid sentiment)}{P(sentence)}$$

Bayes classifier is tested with another set of data containing 4000 positive and negative reviews and 1000 positive and negative reviews. The algorithm works as follows:

Algorithm :

S1 : Initialize P(positive) ← num – propozitii (positive) / num_total_propozitii

S2: Initialize P(negative) ← num – propozitii (negative) / num_total_propozitii

S3: Convert sentences into words for each class of {positive, negative}:

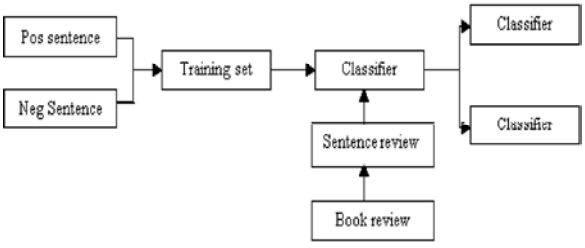
For each word in {phrase}

P(word|class) < num_apartii(word|class) | num_cuv(class) + num_total_cuvinte

P(class) ← P(class)* P(word|class)

Returns max{P(pos),P(neg)}

The algorithm can be represented as:



On experimenting 5000 sentences using Bayes classifier the results obtained gave an accuracy of about 0.79 or 0.8(approx). The calculation of conditional probabilities of every attributes in the predicted class, which can be estimated from the training set of data is the requirement for training a Naive Bayes classifier [17,18].

K-Nearest Neighbor: K-Nearest Neighbour is another supervised machine learning algorithm used for sentiment analysis [19, 20]. The principle of KNN is for a given instance in a training dataset, the class of a new occurrence is similar to that of the majority of its closest neighbor instances. KNN works by monitoring the k-closest instances in the set of data to a new occurrence which is needed to be classified and the prediction is done based on the classes from where the majority of k-neighbours belongs [21, 22]. Based on the majority of the neighbors the objects are classified. The closeness is given by a distance function between the two points in the space of the attribute. This is specified in prior as a parameter to the algorithm.

KNN's Performance Evaluation: On experimenting the KNN for 1000 positive and 1000 negative reviews using 3-fold cross validation where K=3 [22]. In this two folds were used as training dataset and one fold was used as a testing dataset. Inorder to create small subsets from a large training set, it should be divided into 10 disjoint set(T1,T2,,,,,,T10) and also 10 new training sets (TT1, TT2,,,,,TT10). Where,

TT1 = T1

Tti = TTi_1 + Ti (I=2,,,,10)

The performance of the KNN classifer is based on the results of 10 experiments which is conducted on 10 train set pairs. In addition the accuracy, precision and recall methodologies are also used for evaluating the performance of the opinion mining [22]. Accuracy defines the overall accuracy of the sentiment models. Recall (pos) and precision (pos) are the ratio of the recall and precision ratio for the positive reviews. Similarly Recall (neg) and Precision (neg) are the ratio of the recall and precision ratio for the negative reviews

	True Positive Reviews	True Negative Reviews
Predict positive	a	b
Predict Negative	c	d

The accuracy of the performance evaluation is calculated by,

$$Accuracy = \frac{a+d}{a+b+c+d} \tag{4}$$

The recall (pos) and recall (neg) is calculated by,

$$\text{Re}call(pos) = \frac{a}{a+c}$$
$$\text{Re}call(neg) = \frac{d}{b+d} \tag{5}$$

The precision (pos) and precision(neg) is calculated by,

$$\Pr ecision(pos) = \frac{a}{a+b}$$
$$\Pr ecision(neg) = \frac{d}{c+d} \quad (6)$$

From 4, the accuracy increases when increases. The recall and the Precision values are based upon the values of b and c values.

RESULTS

On experimenting 1000 positive and 1000 negative reviews the algorithms gave different types of accuracies in which the SVM classifier achieves the highest. The accuracies of the three algorithms are tabulated as:

Total no. of Experiments	Number of reviews in training dataset	Accuracy in%		
		SVM Classifier	Naive 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

The characteristics of the supervised learning algorithm such as speed, memory, flexibility and interpretability are tabulated as:

Classifier	Multi Class Support	Categorical Predictor Support	Speed of Predicti-on	Memory Usage	Interpret-ablty
SVM	It doesnot support multiclass.	Categorical predictor is supported.	Medium for linear SVM and slow for other types.	Medium for linear and multiclass, large for binary.	Easily interpretable for linear SVM and hard for other types.
Naive Bayes	It supports multiclass	Support categorical predictor	Medium for simple distributionsSlow for high-dimensional data.	Small for simple distributions & medium for high-dimensional data	Easy to interpret
Nearest Neighbour	It supports multiclass	Supports categorial predictor	Slow for cubic and medium for others.	Medium memory usage	Hard to interpret.

The data type support of predictors for each classifier is tabulated as:

Classifier	All Predictors Numeriic	All Predictors Categorical	Some Numeric, Some Categorical
SVM	Yes	Yes	Yes
NAIVE BAYES	Yes	Yes	Yes
NEAREST NEIGHBOUR	Euclidan distance only	Hamming distance only	No

CONCLUSION

From the survey it is found that SVM classifier fits well for the Sentiment analysis [23] providing an accuracy of about more than 80% since they have the high-dimensional input space and Document vector space [7]. SVM uses a method known as over fitting protection which doesnot depends on the total number of features which increases the ability to handle large number of features. The only disadvantage of SVM is suppose if any of the categorical or missing values is found it needs to be preprocessed [10]. Low processing memory requirement and less time of execution is the main advantage of Naïve Baye's classifier [24]. It is advisable to use this algorithm when the training factor seems to be a crucial factor for in the system. The main disadvantage of this classifier is that the assumptions of the attributes being independent which may not be valid [10]. When comparing the Naïve Bayes and NN algorithms, the accuracy was extremely significant when the training dataset is small as 50, 150, 200. NN algorithm has the highest computation cost because it is necessary to compute the distance of each instance of the query for all the samples of the training dataset and also it has the poor run time performance when the dataset is very large. NN algorithm is not suitable for distance based learning like SVM because it is not clear to know about which attributes will match the particular distance to provide better results [25]. Finally on comparing the three algorithms SVM gives the best accuracy rate of about more than 80% for large training datasets.

REFERENCES

1. Kuat Yessenov and Sasa Misailovic, 2009. Sentiment Analysis of Movie Review Comments, Spring.

2. Pablo Gamallo and Marcos Garcia, 2014. Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets, Proceedings of the 8th International Workshop on Semantic Evaluation, 2014
3. Gaurangi Patil, Varsha Galande, Vedant Kekan and Kalpana Dange, 2014. Sentiment Analysis Using Support Vector Machine, International Journal of Innovative Research in Computer and Communication Engineering, 2(1).
4. Rohini S. Rahate and M. Emmanuel, 2013. Feature Selection for Sentiment Analysis by using SVM, International Journal of Computer Applications, 84(5).
5. Jayashri Khairnar and Mayura Kinikar, 2013. Machine Learning Algorithms for Opinion Mining and Sentiment Classification, International Journal of Scientific and Research Publications, 3(6).
6. Ankush Sharma Aakanksha, 2014. A Comparative Study of Sentiments Analysis using Rule Based and Support Vector Machines, International Journal of Advanced Research in Computer and Communication Engineering, 3(3).
7. Safa Ben Hamoud and Jalel Akaich, 2013. Social Networks' Text Mining for Sentiment Classification: The case of Facebook' statuses updates in the "Arabic Spring" Era, International Journal of Application or Innovation in Engineering & Management (IJAIEM), 2(5).
8. Frits Gerit John Rupilele, Danny Manongga, Wiranto Herry Utom, 2013. Sentiment Analysis of Nation Exam Publication Policy With Naïve Bayes classifier Method, Journal of Theortical and applied Information Technology, 58(1).
9. Gautami Tripathi, Naganna S., 2015. Feature Selection And Classification Approach For Sentiment Analysis, Machine Learning And Application: An International Journal, 2(2).
10. Pravesh Kumar Singh, Mohd Shahid Husain, 2014. Methodological Study of Opinion mining and Sentiment Analysis Techniques, International Journal on Soft Computing, pp: 5.
11. Bo Pang and Lillian Lee, Shivakumar Vaithyanathan, 2012. Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP.
12. Chandra Kala, S. and C. Sindhu, 2012. Opinion Mining And Sentiment Classification:A survey, CTACT Journal On Soft Computing, vol. No, Issue 0, October 2012.
13. Dhiraj Gurkhe, Niraj Pal and Rishit Bhatia, 2014. Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification, International Journal of Computer Applications, 99(13).
14. Lina L. Dhande and Girish K. Patnaik, 2014. Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier , International Journal of Emerging Trends & Technology in Computer Science, 3(4).
15. Vivek Narayanan, Ishan Arora and Arjun Bhatia, Fast and accurate sentiment classification using an enhanced Naïve Bayes Model.
16. Govindarajan, M., 2013. Sentiment Analysis of Movie Reviews Using Hybrid Method of Naive Bayes and Genetic Algorithm, International Journal of Advanced Computer Research, 3(4): 13.
17. Christos Troussas, Maria Virvou, Kurt Junshean Espinosa, Kevin Llaguno and Jaime Caro, 2013. Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning, IEEE.
18. Alexandre Trilla, Francesc Alias, Sentiment Analysis of Twitter messages based on Multinomial Naive Bayes.
19. Vohra, S.M. and J.B. Teraiya, 2013. A Comparative Study Of Sentiment Analysis Techniques, Journal Of Information, Knowledge And Research In Computer Engineering, 2(2).
20. Ahmed Alsaffar and Nazlia Omar, 2015. Integrating a Lexicon Based Approach and K Nearest Neighbour for Malay Sentiment Analysis, Center for AI Technology, FTSM University Kebangsaan Malaysia, UKM 43000 Bangi Selangor, Malaysia.
21. Kalaivani, P. and K.L. Shunmuganathan, 2013. Sentiment Classification of Movie Reviews By Supervised Machine Learning Approaches, Indian Journal of Computer Science and Engineering, 4(4).
22. Aamera Z.H. Khan, Mohammad Atique and V.M. Thakare, 2015. Sentiment Analysis Using Support Vector Machine, 5(4).
23. Juan Diego Rodriguez, Learndro Alzate, Manuel Lucania, Inaki Inza and Jose Antonio Lozano, 2011. Approaching Sentiment Analysis by Using Semi-Supervised Learning of Multi-dimensional Classifiers , University of the Basque Country.
24. Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam, 2014. Comparative Study of Classification Algorithms used in Sentiment Analysis, International Journal of Computer Science and Information Technologies, 5(5).
25. http://people.revoledu.com/kardi/tutorial/KNN/Strength%20and%20Weakness.htm.