

Sentiment Analysis with Machine Learning Techniques and Improved J48 Decision Tree Technique

Sakshi koli

Department of Computer Science, Tula's Institute, Dehradun, India

Author's Mail Id: kolisakshi84@gmail.com

DOI: <https://doi.org/10.26438/ijcse/v9i6.7782> | Available online at: www.ijcseonline.org

Received: 09/Jun/2021, Accepted: 13/Jun/2021, Published: 30/Jun/2021

Abstract— Last few years the area of social media , e- commerce, social field has seen a large increase in the web world. The product view became the basic need of today's world . The product reviews channel the customers and help them in making decisions regarding various available products which otherwise would bemuse them. This circumstances opened a new area of research called Opinion Mining and Sentiment Analysis. sentiment analysis is the process of determining the emotion ,feeling, and views of the people towards the piece of text, that comes under the area of blog view, article review , product review, social media buzzing etc. This research paper presents machine learning methods for detecting the sentiment expressed by movie reviews. The semantic point of reference of a review can be positive or negative.

Keywords— Sentiment analysis, sentiment analysis techniques, Experimental result, comparative analysis, conclusion

I. INTRODUCTION

Sentiment analysis is a natural language processing task that deals with the identification and extraction of subjective information or the opinion from the given text documents.

It determines the attitude or the contextual polarity of the document. Sentiment analysis conveys the basic task of classification of the expressed opinion in a document into "positive", "negative", or "neutral" class. Beyond polarity, sentiment classification can be used with the emotional states such as "happy", "sad", and "angry." There are many researches in the area of sentiment analysis of user reviews. The performance of sentiment analyzer is mostly dependent on the topic. As a result, we cannot determine which classifier is the finest. Recently, sentiment analysis has taken enormous interests as the rise of social media such as blogs and social networks .With the eruption of reviews, ratings, recommendations and other forms of online exterior, online opinion has turned into a type of virtual currency for businesses looking to market their products, identify new opportunities and manage their reputations. It can also be used to make decisions to purchase or to use services by individuals. Advertisement market can use sentiment analysis to place ads on praised sites. And, sentiment analysis can also be used for opinion retrievals^[1]. Sentiment analysis is the process of determining the opinion feeling, emotions in the piece of text. With the ever-increasing popularity of social networking, micro-blogging and blogging websites, a large amount of data is engendered every day. These social websites depend largely on the user content that are generated rapidly.

Typically, when people destine to purchase a product, they browse online sites to gain some information about the product before they make their final purchase. They take into consideration the ratings of these products and available reviews on these websites before making purchases. Thus, in order to make this process resourceful and to automate it, several sentiment analysis techniques are used. Sentiment analysis is by and large conducted at different levels varying from coarse-level to fine-level. Coarse-level analysis is primarily concerned with finding the sentiment score of the whole document whereas fine-level deals with attribute level. Sentence-level sentiment analysis is sandwiched between these two.

II. RELATED WORK

Domingos et al.[1] discovered that Naive Bayes works well for certain problems with highly dependent features. This is astounding as the basic assumption of Naive Bayes is that the features are independent.

Zhen Niu et al. acquainted a new model in which efficient approaches are used for feature selection, weight computation and classification. The new model is established on Bayesian algorithm. Here, weights of the classifier are arranged by making use of representative feature and unique feature. 'Representative feature' is the information that represents a class and 'Unique feature' is the information that benefit in distinguishing classes. Using those weights, they determined the probability of each classification and thus improved the Bayesian algorithm.

Barbosa et al.[2] implemented a 2-step automatic sentiment analysis method for classifying tweets. They used a noisy training set to trim down the labelling effort in developing

classifiers. Initially, they classified tweets into subjective and objective tweets. Later than subjective tweets are classified as positive and negative tweets. Pak et al.[3] created a twitter corpus by automatically collecting tweets using Twitter API and automatically annotating those using emoticons. Using that corpus, they made a sentiment classifier based on the multinomial Naive Bayes classifier that uses N-gram and POS-tags as features. In that method, there is a possibility of error since emotions of tweets in training set are labeled exclusively based on the polarity of emotions. The training set is also less efficient since it contains only tweets having emotions.

Xia et al. [4] used an ensemble framework for sentiment classification. Ensemble framework is acquired by combining various feature sets and classification techniques. In their work, they considered two types of feature sets and three base classifiers to form the ensemble framework. Two types of feature sets are implemented using Part of speech information and Word-relations. Naive Bayes, Maximum Entropy and Support Vector Machines are considered as base classifiers. They applied different ensemble methods like Fixed combination, Weighted combination and Meta-classifier combination for sentiment classification and found better accuracy. Various attempts are made by some researches to identify the public opinion about movies, news etc. from Twitter posts. Melville[5], Rui[6], Ziqiong [7], Songho [8], Qiang[9] and Smeureanu [10] used naïve bayes for classification of text. Naïve bayes is one of the most trendy method in text classification. It is acknowledged as one of the most simple and efficient approaches in NLP. It works by manipulating the probability of an element being in a category. First the earlier probability is calculated which is afterwards multiplied with the likelihood to calculate the final probability. The method accepted every word in the sentence to be independent which makes it easier to implement but less accurate. This is the reason that this method is given the name 'naïve'

III. METHODOLOGY

1)Data collection: The dataset consider for training and testing of model. In this work we have used labelled polarity movie dataset. For training the model we have used polarity-dataset v 2.0. For testing dataset we have used polarity – dataset v 3.0.We have used labelled dataset for the classification.

Proposed Work

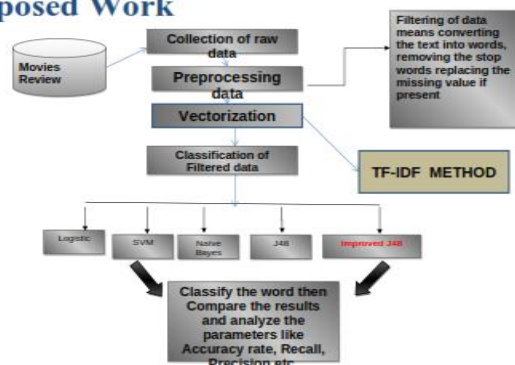


Figure 1

2) Pre-processing: Pre-processing the data is the procedure of cleaning and preparing the text for classification. Online texts surround usually lots of noise and uninformative parts such as HTML tags, scripts and advertisements. In addition, on words level, many words in the text do not have force on the general orientation of it. Keeping those words makes the dimensionality of the difficulty high and hence the classification more difficult since each word in the text is treated as one dimension. Interpretations of having the data correctly pre-processed are to reduce the noise in the text should help progress the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis[12].The whole process has several steps: online text cleaning, white space removal, expanding abbreviation, stemming ,stop words removal, negation handling and finally feature selection. All of the steps but the last are called transformations, while the last step applying some functions to select the required patterns is called filtering[12].

3) Vectorization: Vectorization is the process of convert dataset into numerical vectors where each vector represents review and entries of each vector represent the presence of feature in the review.

Term frequency - Inverse document frequency (TF-IDF): IF-IDF is the process to find important words in the document file. TF finds .number of times a term occurs in a document and IDF shows how relevant and non relevant documents are.

$$IF = \frac{\text{Number of times word appear in the documents}}{\text{Total Number of terms in the document}}$$

$$IDF: \text{Log} \frac{\text{Total number of documents}}{\text{Number of document with specific word}}$$

With following values:

F1: Total occurrences of the specific word in the document.

F2: Total number of the specific word occurring in document.

F3: Total number of documents in which the term specific word occurs.

$$F = \frac{F1}{F2} \quad IDF = \log_e \left(\frac{F2}{F3} \right)$$

Count Vectorizer (CV): Count Vector based on the number of a feature in the given review dataset. In count vector a sparse matrix is generated. In sparse matrix each word is feature. In count vectorizer each sentence in a document is covert into the token and according to the number of occurrence of each token a spare matrix is generated.

4) Machine learning algorithm mclassifier:

A. Existing Classification Methods

Naïve bayes classifier: The naive Bayesian Classification can be represents as supervised learning method as well as a statistical method for classification. It is a probabilistic classifier. It uses the properties of bayes theorem. It assumes strong independence between the features . Consider an underlying probabilistic model and it allows

us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. This Classification is assigned after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Bayesian classification present practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification utilize a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. It is a generative model.

For a review document 'd' and for a class 'c' could be positive and negative ,the conditional probability for each class a review is $P(c|d)$., where that assigns the class c to a given document.

$$P(c|d) = \frac{p(d|c) * P(C)}{P(d)}$$

To compute the term $P(d|c)$, it is decomposed by assuming that f_i 's are conditionally independent given d's class. This decomposition of $P(d|c)$ is expressed in following equation:

$$P(c|d) = \operatorname{argmax}_c \left(\frac{p(c) \prod_i p(f_i|c)}{p(d)} \right)$$

Logistic regression classifier :

Logistic regression models, Also known as maximum entropy, Gibbs, exponential and multinomial logic models, offer a general purpose machine learning technique for classification and prediction which has been successfully applied to fields of sentiment analysis.

Logistic regression classification is an alternative technique which has proven effective in a number of natural language processing applications show that it sometimes, but not always, outperforms Naive Bayes at standard text classification. Logistic regression is flexible they allow stochastic rule with syntactic, semantic and pragmatic features. Logistic regression is discriminative All the feature in discriminative model is correlated . it does not allow the conditional independency between the features in a given field given observation sample document d and labelled class c, for each document sample d belongs to D ,its probability being assigned labelled class c estimate by max entropy model is :

$$P_{ME}(c | d) := \frac{1}{Z(d)} \exp \left(\sum_i \lambda_{i,c} F_{i,c}(d, c) \right) ,$$

where $Z(d)$ is a normalization function. $F_{i,c}$ is a feature/class function for feature f_i and class c , defined as follows:⁶

$$F_{i,c}(d, c') := \begin{cases} 1, & n_i(d) > 0 \text{ and } c' = c \\ 0 & \text{otherwise} \end{cases} .$$

Decision tree classifier: Decision tree work with hierarchical division of the underlying data space with use of different text feature. The hierarchical division of the data space is planned in order to create class partitions which are more twisted in terms of their class distribution. For a known text instance, we determine the partition that

it is most likely to belong to, and use it for the purposes of classification. J48 is an unwrap source Java implementation of the C4.5 algorithm in the Weka data mining tool. C4.5 is a program that creates a decision tree based on a set of labeled input data [46]. The C4.5 algorithm is (Quinlan, 1993; Han and Kamber, 2007; Dunham, 2004) extension of his own ID3 algorithm for generating decision trees. Bagging and Boosting are general strategies for improving classifier and predictor accuracy.

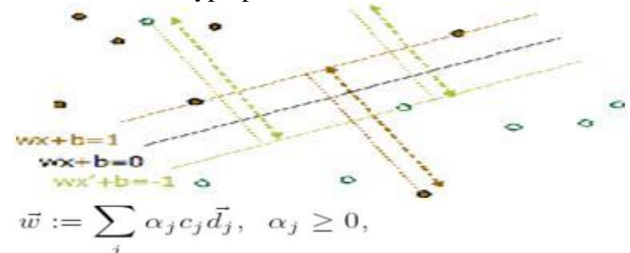
Suppose that we the symptoms. In place of asking one doctor, we may choose to ask several. If particular diagnosis occurs more than any others, we may choose this as the final or best diagnosis. That is the resulted diagnosis is made based on a majority vote where each doctor gets an equal vote. Now replace each doctor by a classifier, we have the basic idea behind bagging support vector machine (SVM) classifier: Support vector machine performs non-probabilistic binary linear classifier . At traditional text categorization effective svm have been shown to be highly. It is generally outperforming Naive Bayes. They are large-margin, rather than probabilistic, classifiers, in contrast to Naïve Bayes and MaxEnt. Support vector machine performs classification by finding the hyper plan that maximize the margin between two classes .The vector that define the hyper plan is the support vector. Support vector machine works in infinite and finite dimensional space.

In this case we want to find an hyper plane able to separate correctly all the documents assigned to a class c. Furthermore, the algorithm finds an hyper plane with a margin as higher as possible separated from other classes and the class c. We talk about hyper plane and not plane because is generalized to N dimensions.

The hyper plan can be describe by the following equation:

$$L_P(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^N \alpha_i [y_i (w^T x_i + b) - 1]$$

We need to compute derivations from Lagrangian Dual to find the alpha for each dimension to solve the hyperplane. Those alpha are known as support vectors that can be used to calculate the hyperplane in the formula:



$$\bar{w} := \sum_j \alpha_j c_j \bar{d}_j, \quad \alpha_j \geq 0,$$

Figure 2: SVM (Different boundary decisions are possible to separate two classes in two dimensions. Each boundary has an associated margin).

B. Proposed Bagged Ensemble Classifiers

Improved j48 decision tree: The improved j48 is based on bagging (bootstrap aggregation) model .Bagging is idea

of making various sample of the training set and the classifier generate for each sample.

The proposed algorithm creates the models (classifiers or predictors) for a learning scheme where each model gives an equally-weighted prediction.

Input:

D, a set of d training tuples; m, the number of models in the ensemble; a learning scheme J48

Output: A composite model, M*. Method:

- for $i = 1$ to m do $_ \rightarrow$ create m models:
- create bootstrap sample, D_i , via sampling D with replacement;
- use D_i to derive a model, M_i ;
- end for
- To apply the composite model on a tuple, X :
- if classification then
- for classifying X use each of the k models and return the majority vote;
- if prediction happens then let each of the k models predict a value for X and return the average predicted value;

IV PERFORMANCE EVALUATION MEASURES

A. Cross Validation Technique

- Cross-validation is a technique for retrieving how the results of a statistical analysis will generalize to an independent data set. It is mostly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. 10-fold cross validation is commonly used. In stratified K-fold cross-validation the folds are elected so that the mean response value is roughly equal in all the folds.

B. Criteria for Evaluation

- The key metric for evaluating classifier performance is classification Accuracy – the percentage of test samples that are correctly classified. The accuracy of a classifier states to the ability of a given classifier to correctly predict the label of new or previously unseen data (i.e. tuples without class label information). Likewise, the accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new or previously unseen data.

C. Confusion matrix: For finding the performance of classifier we are used confusion matrix. Confusion matrix shows the relation between correctly and wrongly predicted review. Confusion matrix is the one popular tool to evaluate the performance of a trained model in tasks of classification or prediction. It not focus on the how the fast model takes to perform the classification rather its focus on the predictive capability of a model. The instances of predicted class represents by each row and each the

instances of actual class column represents. One of the advantages of using this performance evaluation tool is that the data mining analyzer can easily see if the model is confusing two classes and mislabeling easily detects by confusion matrix. The matrix also shows the accuracy of the classifier as the percentage of correctly classified patterns in a given class divided by the total number of patterns in that class. The average accuracy of the classifier is also evaluated by using the confusion matrix.

Table 1

Confusion matrix	Predicted	Predicted
Actual	True positive	False negative
Actual	False positive	True negative

Confusion matrix is the combination of true positive, false negative, false positive, true negative. True positive represents the number of positive movie review that are correctly predicted positive. False negative represents the number of negative review that wrongly predicted negative. False negative represents the number of negative movie review that correctly predicted negative. True negative represents the number of negative movie review that correctly predicted negative.

b) Precision : It shows the exactness of the classifier. It is the ratio of number of correctly predicted positive movie reviews to the total number of movie reviews predicted as positive.

$$precision = \frac{TP}{TP + FP}$$

c) Recall : It represents the completeness of the classifier. It shows the ratio between the number of correctly predicted positive reviews and the actual number of positive reviews present in the corpus.

$$RECALL = \frac{TP}{TP + FN}$$

d) F-measure: F-measure also known as f –score provides the combination of precision and recall. It is the mean of precision and recall. F-measure consider best value as 1 and worst value as 0. The formula for calculating F-measure is presented as:

$$F - Measure = \frac{2 * precision * recall}{precision + recall}$$

e) Accuracy: It is calculated the ratio of number of correctly predicted reviews to the number of total number of reviews present in the dataset. The formula for calculating accuracy is given as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

IV. RESULTS AND DISCUSSION

A. Dataset Description

The basic data set consist of 128 movie reviews, 64 labelled positive and 64 labelled negative (so they have a uniform class distribution). The sewere downloaded from Bo Pang's web page: [http:// www. cs. cornell. edu/people/pabo/moviereview- data/](http://www.cs.cornell.edu/people/pabo/moviereview-data/).

Result of base machine learning classifier:

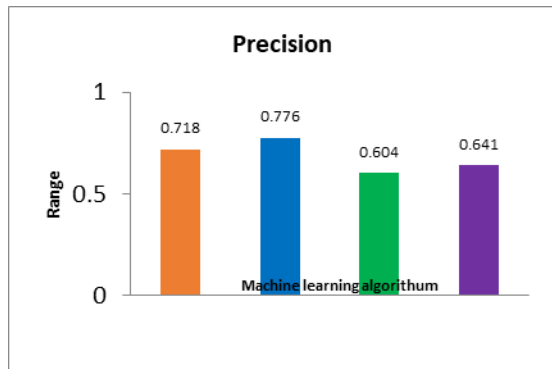


Figure 3

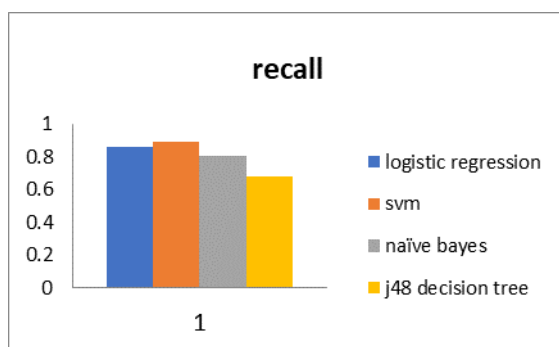


Figure 4



Figure 5

In this research work, new ensemble classification method is proposed using bagging classifier in conjunction with j48 decision tree as the base learner and the performance is analyzed in terms of accuracy, precision ,recall. Here, the base classifiers are constructed using logistic regression, j48 classifier , support vector machine , naïve bayes.

10-fold cross validation technique is applied to the base classifiers and proposed improved j48 and evaluated classification accuracy. Bagging is performed with j48

decision tree to obtain a very good classification performance. Table 1 to 3 shows classification performance for movie review using existing and proposed bagged j48 decision tree. The analysis of results shows that the proposed bagged j48 decion tree classifier are shown to be superior to individual approaches for movie review in terms of classification accuracy, precision, recall.

According to Fig. 1 to 3 proposed combined models show significantly larger improvement of classification accuracy than the base classifiers.

V. CONCLUSION AND FUTURE SCOPE

We have done sentiment analysis with different machine learning and improved j48 decision tree technique. In end of the work result of five prediction model are compared in order to find the best method for sentiment analysis. Support vector machine has showed the best result in recall and native bayes has showed the best result in precision. Improved j48 is the best classifier among four machine learning technique. Different classifier accuracy obtained as 86.6% for improved J48 decision tree, % for j48 decision tree. As the Comparison of machine learning technique, improved j48 showed up as a best technique for classification. Sentiment analysis system can be further enhanced by adding more testing and training dataset for predication. System can be tested by adding more feature selection method in proposed technique. Hybrid approach with machine learning algorithm and lexical method could be used for more improvement in the sentiment analysis system.

REFERENCES

- [1] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2-3, pp. 103–130, 1997.
- [2] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 36–44, Association for Computational Linguistics, 2010.
- [3] A .Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, 2010, pp.1320-1326.
- [4] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences: an International Journal*, vol. 181, no. 6, pp. 1138–1152, 2011
- [5] Melville, Wojciech Gryc, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification", *KDD'09*, June 28–July 1, 2009, Paris, France. Copyright 2009 ACM 978-1-60558-495-9/09/06.
- [6] Rui Xia, Chengqing Zong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", *Information Sciences* 181 (2011) 1138–1152
- [7] Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li, "Sentiment classification of Internet restaurant reviews written in Cantonese", *Expert Systems with Applications* xxx (2011) xxx–xxx

- [8] Songbo Tan, Jin Zhang, “An empirical study of sentiment analysis for chinese documents”, Expert Systems with Applications **34 (2008) 2622–2629**.
- [9] Qiang Ye, Ziqiong Zhang, Rob Law, “Sentiment classification of online reviews to travel destinations by supervised machine learning approaches”, Expert Systems with Applications **36 (2009) 6527– 6535**.
- [10] on SMEUREANU, Cristian BUCUR, “Applying Supervised Opinion Mining Techniques on Online User Reviews”, Informatica Economică **vol. 16, no. 2/2012**.
- [11] A. Mardin, T. Anwar, B. Anwer, “*Image Compression: Combination of Discrete Transformation and Matrix Reduction*”, International Journal of Computer Sciences and Engineering, **Vol.5, Issue.1, pp.1-6, 2017**.
- [12] A. Balahur, J. M. Hermida, and A. Montoyo, “Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model,” Affective Computing, IEEE Transactions on, **vol. 3, no. 1, pp. 88– 101, 2012**.
- [13] T. Peng, C. Shih, “An Unsupervised Snippet-Based Sentiment Classification Method for Chinese Unknown Phrases without Using Reference Word Pairs.” Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, **2010, pp.243-248**.

AUTHORS PROFILE

Ms Sakshi Koli pursued Bachelor of Science from DIT dehardun in 2014 and Master of Science from DIT University in year 2016. She is currently working as Assistant Professor in Department of Computational Sciences, Tulas Insitute dehardun.. Her main research work focuses on data mining, Machine Learning , Computational Intelligence based education.

