Sentiment Analysis on Movie Reviews

B. Lakshmi Devi, V. Varaswathi Bai, Somula Ramasubbareddy and K. Govinda

Abstract Movie reviews help users decide if the movie is worth their time. A summary of all reviews for a movie can help users make this decision by not wasting their time reading all reviews. Movie-rating websites are often used by critics to post comments and rate movies which help viewers decide if the movie is worth watching. Sentiment analysis can determine the attitude of critics depending on their reviews. Sentiment analysis of a movie review can rate how positive or negative a movie review is and hence the overall rating for a movie. Therefore, the process of understanding if a review is positive or negative can be automated as the machine learns through training and testing the data. This project aims to rate reviews using two classifiers and compare which gives better and more accurate results. Classification is a data mining methodology that assigns classes to a collection of data in order to help in more accurate predictions and analysis. Naïve Bayes and decision tree classifications will be used and the results of sentiment analysis compared.

Keywords Prediction · Movie reviews · Naive Bayes · Decision tree · SLIQ

1 Introduction

Sentiment analysis of movie-rating sites can be of various applications like for other online movie-rating sites and for sites regarding opinions on books, products and various other things. Sentiment analysis or opinion mining is a method to

B. Lakshmi Devi · V. Varaswathi Bai
SOET, SPMVV University, Tirupati, Andhra Pradesh, India

S. Ramasubbareddy (✉)
Information Technology, VNRVJIET, Vignana Jyothi Nagar, Nizampet Rd, Pragathi Nagar, Hyderabad, Telangana 500090, India
e-mail: svramasubbareddy1219@gmail.com

K. Govinda
SCOPE, VIT University, Vellore, Tamil Nadu, India

P. Venkata Krishna and M. S. Obaidat (eds.), *Emerging Research in Data Engineering Systems and Computer Communications*, Advances in Intelligent Systems and Computing 1054, https://doi.org/10.1007/978-981-15-0135-7_31

systematically extract and identify affective state and the subjective information. In a bunch of text, we can use sentiment analysis for separating words which denote 'happy' emotion from words which denote 'sad' emotion, as a simple example. Opinion mining (also called emotion AI or sentiment analysis) is the utilization of computational linguistics, natural language processing, biometrics and text analysis to consistently establish, study, quantify and extract subjective data and emotional states. Sentiment analysis is extensively useful to voice of the client materials like survey rejoinders and reviews, online and social mass media, and healthcare materials for requests that vary from customer service to promoting to clinical medicine. Naive Bayes classifiers, in the field of machine learning, are a class of simple 'probabilistic classifiers' based on the application of Bayes' theorem with robust unconventional norms amongst the features. Naive Bayes has been examined broadly since 1950. The community involved with the retrieval of texts was introduced to it in the early 1960s and remained a preferred methodology for categorization of text, the issue of deciding texts as fitting into one class or the opposite with frequency of words as the characteristics. With applicable pre-processing, it is competitive in this area with additional cutting-edge methods as well as support vector machines. Naïve Bayes additionally finds use in automated clinical analysis. Decision tree-based learning makes use of a decision tree to travel from remarks concerning an item (denoted within the branches) to deductions concerning the element's goal worth (signified within the leaves). It is one amongst the predictive modelling tactics employed in machine learning, data mining and statistics. In classification trees, the tree model can take a distinct set of values for the target variable. Class labels are represented by leaves, and the combination of features which come to end with these class labels is represented by branches. Regression trees are those decision trees where the targeted variable can take uninterrupted values like in the case of real numbers. Decision trees are used to explicitly and visually signify decisions and higher cognitive process in decision analysis. A decision tree labels data in data mining, but the ensuing classification tree can be used for decision-making. Decision tree shapes regression or classification models in the kind of a tree assembly. It disrupts a data set into smaller subsets, whereas at a similar time a linked decision tree is established incrementally. The concluding outcome is a tree with leaf and decision nodes. Decision trees can manage numerical and categorical information. Advantages of such a system are that it permits for machine-controlled film scoring arrangement. Incorrect rating cannot be given because system computes on the basis of clients' remarks. It eliminates human blunders that normally arise throughout physical scrutiny. An unbiased result is provided by the system. Therefore, the system dismisses need of human labours and conserves resources and time. System disadvantages are that it should be provided with true inputs; else wrong outcomes are formed. Cloud hosting of the system needs to be done to process and receive results across the nation.

2 Background

Sentiment analysis could be a sort of language process for following the mood of the general public a couple of specific product or topic. Sentiment analysis, that is additionally referred to as opinion mining, involves in building a system to gather and examine opinions concerning the merchandise created in weblog posts, comments, reviews or tweets. Sentiment analysis is helpful in many ways that like in decision-making the success of a commercial campaign or new product launch confirms that versions of a product or service are standard and even determine that demographics like or dislike specific options. There are many challenges in sentiment analysis. The primary is associate degree opinion word that's thought-about to be positive in one state of affairs is also thought-about negative in another state of affairs. A second challenge is that individuals do not invariably express opinions in a same manner [2]. Each film review comment is going to be coordinated in contrast to the opinion terms in sentiment dictionary and classified into three categories: positive (+1), neutral (0) and negative (−1), reliant on the comparative number of negative and positive opinion terms in the remark, i.e., |neg(c)|and|pos(c)|, respectively. Explicitly, the sentiment polarization of comment c is outlined as: polarity(c) = sgn(|pos(c)| − |neg(c)|) [3]. To upsurge the precision of the classification, we should reject regular n-grams, i.e., n-grams that neither do not powerfully specify any objectivity nor show sentimentality of a sentence. In all data sets, such n-grams occur consistently [5]. Recommendation systems have become extremely important in the last decade with the rapid increase in the size and complexity of data provided over the World Wide Web. Internet users usually desire to be fed by simplified and customized procedures to access the information they require. A key step in the success of a satisfactory recommendation system is the effective prediction of the rating that will be potentially given by a user to a specific item. In this way, the system can recommend the users the items that they likely enjoy [1]. As we tend to all apprehend that the rating to a moving picture will mirror the favour degree of the user and also the same moving picture can receive totally different rates from individuals with different preferences. As a result of the user interest on movies that typically varies with the attributes of flicks, like director, actor and so on, the influence of user rating is often delineated by the attributes of the moving picture to some extent [4]. Despite its delusive unconventionality assumption, the Naive Bayes classifier is astonishingly effective in practice since its classification decision could usually be correct although its chance estimates are inaccurate [6]. Even though the discriminative logistical regression algorithmic rule features a lower asymptotic fault, the generative Naive Bayes classifier may additionally converge more quickly to its asymptotic fault. Thus, because the variety of training examples is exaggerated, one would expect generative Naive Bayes to at first do higher, except for discriminative logistical regression to eventually catch up to, and quite seemingly overtake, the performance of Naive Bayes [7]. Regardless of its straightforwardness, Naive Bayes can every so often beat more complex classification procedures. In a few fields, its functioning

has been revealed to be analogous to that of decision tree learning and neural network [8]. The group of accounts accessible for evolving classification approaches are commonly disintegrated into two disconnected subsets, test set and training set. Training set is used for originating the classifier, while the test set is used to assess the accurateness of the classifier. SPRINT and SLIQ are decision tree classifiers which have been shown to achieve decent efficiency, accuracy and compactness for very bulky data sets [9]. A great quantity of programmes has been fashioned by the machine learning community to generate decision trees for classification. Quite prominent midst these for data classification include ScalParc, SPRINT, SLIQ, CHAID, CART, C5, C4.5, ID3 [10].

3 Proposed Method

Sentiment analysis of the review data set will be done by two methods: Naïve Bayes classifier and decision tree classifier. Data set used is a set of 500 IMDB movie reviews (half positive and half negative reviews). First, the data set will be read using pandas and data frame will be used for further processing. Pre-processing part involves cleaning of data, i.e., tokenization, removal of stop words, removal of special characters, stemming, etc. In this way, the data will be cleaned. Then, the data can be used for further classification and hence sentiment analysis. For classification, tf-idf matrix of review column is used and labelled sentiment column is used. Naive Bayes is a straightforward method for building classifiers: representations that allocate class tags to problem examples, depicted as trajectories of feature standards, wherever the category tags are obtained from some restricted set. There is no one formula to train such classifiers, however a category of procedures supported by a standard opinion: the whole lot of Naive Bayes classifiers adopt that the significance of a specific attribute is detached from the usefulness of the other element, provided the group variable. Naive Bayes classifiers are a family of simple probability-based classifiers based on applying Bayes' theorem with robust non-aligned conventions amongst the traits. Using Bayes' theorem, the conditional probability will be decomposed as

$$p(C_k/x) = \frac{p(C_k)p(x/C_k)}{p(x)}$$

To create the classifier model, the possibility of a given set of entries is found for altogether probable values of the category variable y and the output with highest likelihood is picked up. The conforming classifier, a Bayes classifier, is the operation that allocates a class label for $\hat{y} = C_k$ some k as the following:

$$\hat{y} = \underset{k \in \{1,\ldots,k\}}{\arg\max}\, p(C_k) \prod_{i=1}^{n} p(x_i/C_k)$$

Feature vectors denote the occurrences with which specific outcomes have been produced by a multinomial distribution. This is the occurrence prototype characteristically used for article cataloguing.

Decision tree classifier

There are couple of algorithms to develop a decision tree.

- Classification and regression tree (CART) makes use of Gini index as metric.
- Iterative dichotomizer 3 (ID3) makes use of entropy function and information gain as metrics.

A decision tree is a tree-like structure in which each interior node characterizes a 'test' on an attribute, each branch signifies the consequence of the test, and each leaf node denotes a class label (decision taken subsequently figuring out all attributes). The paths from root to leaf signify classification guidelines.

Tree-generating processes is done using C5.0, C4.5 and ID3. The theory of entropy from information theory is the basis of information gain.

Entropy is expressed as the following:

$$H(T) = I_E(p_1, p_2, \ldots, p_J) = -\sum_{i=1}^{J} p_i \log_2 p_i$$

where $p_1, p_2 \ldots$ are fractions that sum up to 1 and denote the proportion of each class existing in the child node that outcomes from a fragment in the tree.

$$\overbrace{\mathrm{IG}(T,a)}^{\text{InformationGain}} = \overbrace{H(T)}^{\text{Entropy(parent)}} - \overbrace{H(T/a)}^{\text{WeightedSumofEntropy(children)}}$$
$$= -\sum_{i=1}^{J} p_i \log_2 p_i - \sum_{a} p(a) \sum_{i=1}^{J} -pr(i/a) \log_2 pr(i/a)$$

Information gain is employed to determine that attribute to divide on at every stage in constructing the tree. Uncomplicatedness is preferable; therefore, tree should be kept tiny. To do so, at every phase we should always opt for the rift that ends up in the deepest daughter nodes. A normally used degree of transparency is called information that is deliberated in bits. For each node of the tree, the data worth represents the expected amount of information that may be required to specify whether or not a new instance ought to be classified yes or no, provided that the example reached that node.

The results of both the classifiers will be compared, and a conclusion will be made as to which model is a better fit for the data set in use.

Table 1 Accuracies of proposed approach

	Naïve Bayes	Decision tree
Precision	0.86	0.65
Recall	0.80	0.65
F-measure	0.80	0.65
AUC	0.97	0.65

4 Results

For the purpose of sentiment analysis in movie reviews, a database of 500 movie reviews has been used. Both Naïve Bayes classification and decision tree classification have been implemented on the data set. To compare both classifiers, the comparison parameters that have been evaluated are confusion matrix, precision, recall, f-measure, receiver operating characteristic curve (ROC curve) and the area under the curve (AUC) (AUC gives the accuracy of the classifier) (Table 1).

Below is tabulated the various parameters that have been evaluated for both classifiers:

The data set used is of 500 IMDB movie reviews, half positive and half negative reviews. From the actual table of reviews [1], reviews from row 24,750 to 25,250

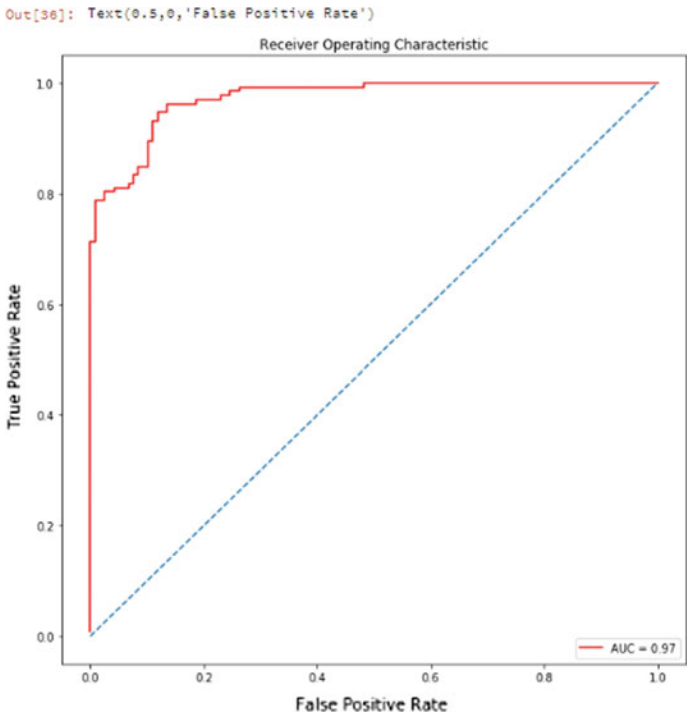

Fig. 1 Accuracy of Naïve Bayes

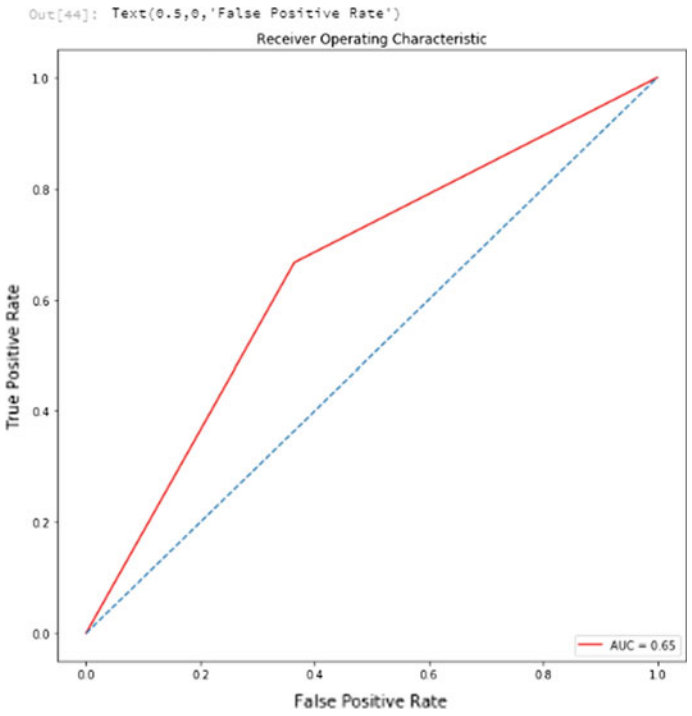

Fig. 2 Accuracy of decision tree

have been used, since the model of laptop used could not handle the memory load of using the previously intended 1 lakh reviews. From the data set, 50% reviews were randomly used as training data and the other 50% were randomly used as test data. After applying classification, the predicted values of sentiment for test data are, ROC obtained are [*Y*-axis: true positive rate (sensitivity), *X*-axis: false positive rate (1- specificity)].

Here, Fig. 1 shows the accuracy is said to be higher, since the curve is more inclined towards the top left corner of the graph.

Here, Fig. 2 shows the accuracy is said to be lower, since the curve is less inclined towards the top left corner of the graph.

5 Conclusion

Sentiment analysis can successfully rate the movie and deduce the emotion of the review almost accurately. After careful analysis of results obtained from both classifiers, I have concluded that when training set is used from the reviews itself, the accuracy of Naïve Bayes classifier (0.97) is more than that of decision tree

classifier (0.65). Therefore, Naïve Bayes classifier is a better fit for the movie review data set used. Naïve Bayes classification predicts accurate sentiment analysis for the given movie reviews.

References

1. Ogul, H., Ekmekciler, E.: Two-way collaborative filtering on semantically enhanced movie ratings. In: The Proceedings of the ITI, pp. 361–366 (2012)
2. Vinodhini, G., Chandrasekaran, R.: Sentiment analysis and opinion mining: a survey. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **2**(6), 282–292 (2012)
3. Wang, J., Liu, T.: Improving sentiment rating of movie review comments for recommendation. In: The IEEE International Conference on Consumer Electronics—Taiwan, ICCE-TW 2017, pp. 433–434 (2017)
4. Li, J., Xu, W., Wan, W., Sun, J.: Movie recommendation based on bridging movie feature and user interest. J. Comput. Sci. **26**, 128–134 (2018)
5. Pak, A., Paroubek, P.: Twitter as a Corpus for sentiment analysis and opinion mining in the computer, pp. 1320–1326 (2010)
6. Rish, I.: An empirical study of the Naive Bayes classifier. In: The Empirical Methods in Artificial Intelligence Workshop, IJCAI (2001) 22230, Jan. 2001, pp. 41–46 (2001)
7. Ng, A.Y., Jordan, M.I.: On discriminative versus generative classifiers: a comparison of logistic regression and Naive Bayes. In: Advances in Neural Information Processing Systems, pp. 841–848 (2002)
8. Islam, M., Wu, Q., Ahmadi, M., Sid-Ahmed, M.: Investigating the performance of Naïve-Bayes classifiers and K-Nearest neighbour classifiers. J. Converg. Inf. Technol. **5**(5), 133–137 (2010)
9. Alsabti, K., Ranka, S., Singh V.: CLOUDS: tree classifier for large datasets. In: The Proceedings of the Fourth Knowledge Discovery and Data Mining Conference, pp. 2–8 (1998)
10. Lavanya, D., Usha Rani, K.: Ensemble decision tree classifier for breast cancer data **2**(1), 17–24 (2012)