International Journal of Synthetic Emotions
Volume 11 • Issue 1 • January-June 2020

A Comparative Study of Different Classification Techniques for Sentiment Analysis

Soumadip Ghosh, Academy of Technology, Kolkata, India
https://orcid.org/0000-0003-4817-5363

Arnab Hazra, Academy of Technology, Kolkata, India

Abhishek Raj, Academy of Technology, Kolkata, India
https://orcid.org/0000-0002-8121-3440

ABSTRACT

Sentiment analysis denotes the analysis of emotions and opinions from text. The authors also refer to sentiment analysis as opinion mining. It finds and justifies the sentiment of the person with respect to a given source of content. Social media contain vast amounts of the sentiment data in the form of product reviews, tweets, blogs, and updates on the statuses, posts, etc. Sentiment analysis of this largely generated data is very useful to express the opinion of the mass in terms of product reviews. This work is proposing a highly accurate model of sentiment analysis for reviews of products, movies, and restaurants from Amazon, IMDB, and Yelp, respectively. With the help of classifiers such as logistic regression, support vector machine, and decision tree, the authors can classify these reviews as positive or negative with higher accuracy values.

KEYWORDS

Decision Tree, Logistic Regression, Machine Learning, Sentiment Analysis, Support Vector Machine

1. INTRODUCTION

Sentiment analysis or opinion mining (Neethu M. et al., 2013) refers to emotions and opinions by analysis of texts, processing of natural languages to methodically identify, extract, count, and study some interesting information. Sentiment analysis has gained popularity in the recent past. The idea of performing analysis on texts is important for marketing research, where analysts wish to find out some useful information from customer feedback. It is vastly applied to various forms of customer feedback such as reviews and survey responses found on the web and social media. Commercial websites such as Amazon, eBay, Yelp and IMDb provide users the platform required to express their opinions towards any specific product or subject. Individuals post reviews of movies they have watched on websites like IMDb.

DOI: 10.4018/IJSE.20200101.oa

Performing analysis of sentiments from various data sources found on the web is valuable for any organization to maintain quality control of their products. For instance, getting user feedback means requesting people with surveys on every aspect the organization is interested in. One of the sources of doing this is web blogs and another one is electronic discussion boards, where individuals can talk about different types of topics or can request other people's views. This approach is beneficial for numerous reasons. Primarily, the people who share their views usually have more noticeable opinions than the average, which are furthermore convincing others to read them. Secondarily, product and service reviews obtained from commercial web sites also help us to choose which products to buy and which services to use. Furthermore, the individual reviews obtained from personal blogging sites are mostly unbiased and have individual experience towards a specific product or service. Mining these opinions is thus carrying valuable information for the improvement of the business.

Opinion mining is a technique of categorizing opinions articulated in the text sentences (Manning et al., 2008) obtained from several data sources. Basically, text sentences carry personal review or attitude concerning any specific product or subject. Opinion mining of small texts is thought-provoking because they are contextually limited. Decisions are to be made based on the inadequate texts provided by the user. We refer to this method as a supervised learning technique as it can categorize each user review correctly (Pang, Lee, & Vaithyanathan, 2002).

Machine learning (ML) (Witten I. H. et al., 2011) based classification models are trained with data sets containing text sentences and their performances are evaluated as well. Classification techniques such as Logistic Regression (LR) (Cramer J. S., 2002), Support Vector Machine (SVM) (Cortes C., & Vapnik V., 1995) and Decision Tree (DT) (Quinlan J. R.,1987) from ML domain can be applied to text data for performing sentiment analysis. These research studies (Kamal S. et al, 2016, 2017, 2018) contributed some methods which we have applied in our work.

The different sections of the research paper are as follows. In the first section, we have introduced about sentiment analysis and described its importance in business. Section 2 provides literature reviews that worth mentioning in this domain. Section 3 presents the data set description which is followed by the proposed methodology in section 4. Section 5 describes and analyzes results with explanations. Finally, section 6 is attributed to the conclusion and future works.

2. LITERATURE REVIEW

There are several ML-based types of research available to classify sentiments from the text. Some of them are listed below.

ML consists of several classification models such as Artificial Neural Network (ANN), SVM, decision tree, Logistic Regression, etc. These techniques are employed to categorize reviews of products. The research study (Mejova Y. et al., 2009) showed that using the presence of every character, frequency of occurrences of every character, text sentence containing negation, etc. as the features to build feature vector. He also showed that using unigram and bigram approaches one could create feature vectors efficiently in Sentiment analysis.

The research work (Domingos P., 1997) proposed that the Naive Bayes classifier could do well using dependent features for a certain problem. This work (Niu Z. et al., 2012) developed a new classifier based on the Bayesian algorithm. The model employed some effective approaches for the selection of a feature, computation of weight and classification. The research study (Barbosa L., & Feng J., 2010) designed a two-step analysis method which was an automatic sentiment analysis for classifying tweets. In the first step, tweets were classified into subjective and objective tweets. Then, in the second step, subjective tweets were classified as positive and negative tweets.

The research work (Celikyilmaz A. et al., 2010) developed a word clustering method based on the pronunciation of words. This method is applicable for normalizing noisy tweets. There are some words with the similar pronunciation but dissimilar meanings. So, to eliminate this kind of conflict, methods were developed. In the stated method, words having the same pronunciation were clustered

and assigned with common tokens. This study (Wu Y., & Ren F., 2011) recommended a model to analyze the sentiments in tweets. In this study, if a user idea is found in the tweet, it took prompt action to help towards influence probability.

The research work (Pak A., & Paroubek P., 2010) established a method for sentiment analysis using some automatic twitter texts. This work designed a Naïve Bayes classifier for sentiment analysis which used emotions in the texts as a feature. Some researchers developed methods to recognize public opinion about movies, news, etc. from tweets. The research study (Peddinti V. et al., 2011) had taken the information from other publicly available databases such as IMDB and Blippr for review analysis.

3. ABOUT THE DATA SET

The data set is taken from the UCI (Kotzias D., 2015). It contains three data sets namely *yelp_labelled*, *imdb_labelled* and *amazon_cells_labelled* data set. In this data set, Score for review is measured either by value 1 (for positive review) or 0 (for negative review). The texts are originated from three different websites namely yelp.com, imdb.com, and amazon.com. For each of these websites, there are 500 positive and 500 negative texts. These are selected randomly from larger review data sets. We have selected sentences that have a positive or negative review. No neutral reviews are considered here. These attributes are essentially texts, extracted from reviews of products, movies, and restaurants from Amazon, IMDB and Yelp respectively.

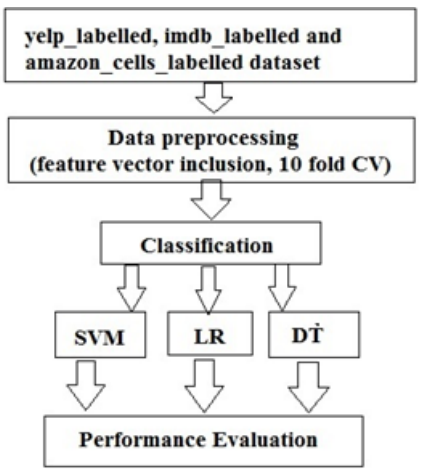
4. METHODOLOGY

Sentiment analysis is all about analyzing texts. These texts can be of books, reviews and all sorts of texts of some HTML webpages that we extract from web scrapping. By using Natural Language Processing (NLP) (Khurana D. et al., 2017) with regular expressions based operations we can perform predictive analysis on text. The workflow of our methodology is shown in Figure 1.

4.1. Step 1: Data Pre-Processing

Data pre-processing is a significant phase in this process. Data are stored in a plain text format along with positive and negative reviews. The data set contains lots of informal words or noises which should be taken care of before being suitable for a model. Therefore, data pre-processing is essential

Figure 1. Workflow of the proposed methodology



to extract all important or meaningful text that will be relevant for training the model. It also contains different forms of verbs of a particular word which has to be converted into one specific form. To deal with this problem, the concept of feature vector has been brought in. But, before using it, pre-processing is done on each review. Then, features are extracted in two phases: the first phase deals with the extraction of the review specific word. Then, they are removed from the given text. The extracted feature vector is then converted to normal text.

After that, features are extracted from the review which is the normal text without any informal words. These extracted features are then added to develop the feature vector.

4.2. Step 2: Separating Training and Testing Data Set

In ML, we generally split our original data set into two sub-sets namely training set and testing set, and then fit our model on the train data, to make predictions on the test data. For these data sets, we have used *k-fold cross-validation (CV)* (here k=10) for splitting the original data sets into training and testing sets.

The training set contains a known output and the model is trained on this data in order to be generalized to other data later on. Then, we have the testing data set to test the prediction capabilities of these models.

4.3. Step 3: Training the Model

Classification is a method to categorize our data into a desired and separate number of classes where we can assign a label to each class. We have used three different classification models namely Support Vector Machine (SVM), Decision Tree and Logistic Regression for our study.

The present work uses an *SVM classifier* using a *Gaussian RBF kernel* with kernel function *K* as:

$$K\left(x_i, x_j\right) = \phi\left(x_i\right) \cdot \phi\left(x_j\right) \tag{1}$$

Here $\varphi\left(x\right)$ is a mapping function applied on the training instances. The SVM classifier can be defined as:

$$K\left(x_i \cdot x_j\right) = e^{\frac{-\left\|x_i - x_j\right\|_2}{2\sigma^2}} \tag{2}$$

Here, we employ the *Classification and Regression Trees (CART)* algorithm which uses the *Gini index* for selection of attribute. This can be represented as:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{3}$$

Logistic regression forecasts the probability of an outcome that is having two values. This technique is used when the dependent variable (i.e. target variable) is categorical. It can be seen as:

Output = 1 or 0

Hypothesis such that: Z = WX + B (4)

$h^*\Theta(x) = \text{sigmoid}(Z)$

We have compared the performance of these algorithms considering the reviews of products, movies, and restaurants from Amazon, IMDB and Yelp respectively.

4.4. Step 4: Testing the Model

Finally, the model is applied to the testing phase. The results of this phase are evaluated against well-known metrics such as RMSE (Armstrong JS., & Collopy F., 1992), Kappa statistic (Carletta J., 1996), and Confusion matrix (Stehman S. V., 1997) based metrics namely Accuracy, Precision, Recall and F1-Measures for performance analysis.

5. RESULT AND DISCUSSION

We have applied three classifiers namely Support Vector Machine, Logistic Regression, and Decision tree to the given UCI data sets for reviews of products, movies, and restaurants from Amazon, IMDB and Yelp respectively. We have divided each of the data sets into two sub-sets namely training set and testing set. The results described here are based on the simulation experiment developed in Python. Several comparisons of these classifiers are done based on some performance measures like classification accuracy, root-mean-square error (RMSE), and kappa statistic values. We have also performed detailed accuracy checking for these classifiers using Precision, Recall and F1-Measure values derived from the confusion matrix of each classifier. Classifiers (SVM, LR, and DT) are applied to a test set for classification after completion of the training phase on each of these data sets.

5.1. YELP Labelled Data Set

Performance comparisons of the three classifiers are presented in Table 1.

Table 1. Performance evaluation based on predicted class level

Classifier	Classification Accuracy (%)	RMSE	Kappa Statistic
SVM	76.5	0.4847	0.5318
Logistic Regression	83.5	0.4062	0.6699
Decision Tree	77.0	0.4795	0.54

We have used classification accuracy, RMSE and kappa statistic values for each of the classifiers in Table 1 for the YELP Labelled Data set. By analyzing Table 1, we see that the Logistic Regression has the highest accuracy and Kappa statistic values and lowest RMSE value among these three classifiers.

Next, we have computed Precision, Recall and F1-measure from the confusion matrix. The result of each of the parameters for each classifier for the Yelp Labelled data set is shown in Table 2.

By analyzing Table 2, we can see that the Precision, Recall and F1-measure values of Logistictic Regression are highest. Values of Logistictic Regression are 84%, 83%, and 83% respectively.

5.2. IMDB Labelled Data Set

Performance comparisons of the three classifiers are described in Table 3.

We have calculated classification accuracy, RMSE and kappa statistic values for each of the classifiers in Table 3 for the IMDB Labelled Data set. By analyzing Table 3, we see that the Logistic

Table 2. Performance evaluation based on confusion matrix

Classifier	Precision	Recall	F1-Measure
SVM	77%	77%	76%
Logistic Regression	84%	83%	83%
Decision Tree	77%	77%	77%

Table 3. Performance evaluation based on predicted class level

Classifier	Classification Accuracy (%)	RMSE	Kappa Statistic
SVM	77.0	0.4795	0.5403
Logistic Regression	79.5	0.4527	0.59
Decision Tree	74.0	0.5099	0.48

Regression is having the highest accuracy and Kappa statistic values and lowest RMSE value among these three classifiers used.

Next, we have computed Precision, Recall and F1-measure values from the confusion matrix. The result of each of the parameters for each classifier for IMDB Labelled Data set is shown in Table 4.

By analyzing Table 4, we can see that the Precision, Recall and F1-measure values of Logistictic Regression are highest. These values for Logistictic Regression are 80%, 80%, and 79% respectively.

Table 4. Performance evaluation based on confusion matrix

Classifier	Precision	Recall	F1-Measure
SVM	77%	77%	77%
Logistic Regression	80%	80%	79%
Decision Tree	76%	74%	73%

5.3. AMAZON CELLS Labelled Data Set

Performance comparisons of the three classifiers are given in Table 5.

We have presented classification accuracy, RMSE and kappa statistic values for each of the classifiers in Table 5 for the AMAZON CELLS Labelled Data set. By analysing the Table 5, we can see that SVM has the highest accuracy and Kappa statistic values and lowest RMSE value among these three classifiers.

Table 5. Performance evaluation based on predicted class level

Classifier	Classification Accuracy (%)	RMSE	Kappa Statistic
SVM	83.0	0.4123	0.6599
Logistic Regression	82.5	0.4183	0.6492
Decision Tree	78.0	0.469	0.5578

Next, we have computed Precision, Recall and F1-measure from the confusion matrix. The result of each of the parameters for each classifier for AMAZON CELLS Labelled Data set is shown in Table 6.

By analyzing Table 6, we can see that the Precision, Recall and F1-measure values of SVM are highest. SVM is having the same value i.e. 83% for each of these evaluation metrics.

Table 6. Performance evaluation based on confusion matrix

Classifier	Precision	Recall	F1-Measure
SVM	83%	83%	83%
Logistic Regression	82%	82%	82%
Decision Tree	78%	78%	78%

6. CONCLUSION AND FUTURE WORKS

We can conclude that machine learning-based techniques can be applied to analyze reviews of products, movies, and restaurants from Amazon, IMDB and Yelp data sets. Sentiment analysis is thought-provoking as it is difficult to detect the words that reveal emotion form reviews and also due to the presence of informal words, hast tags, etc. To deal with this problem, the concept of feature vector has been brought in. Before introducing feature vector pre-processing is done on each review. Then features are extracted in two phases: First phase deals with the extraction of the review specific word. Then, they are removed from the given text. The extracted feature vector is then converted to normal text.

After that, features are extracted from the review which is the normal text without any informal words. These extracted features are then added to develop the feature vector. Finally, different ML-based classifiers are applied to the pre-processed data set for classifying the reviews. From our results, we have shown that LR, SVM, and DT based classifiers perform well and also provide higher accuracy. The result shows that the Logistic Regression classifier performs better than the other classifiers considering all scenarios. In the future, the Logistic Regression based classifier can be used for other kinds of sentiment analysis such as to stop spreading rumors against some sensitive issues or to prevent terrorism.

REFERENCES

Armstrong, J.S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting, 8*(6980).

Barbosa, L., & Feng, J. (2010). Robust Sentiment Detection on Twitter from Biased and Noisy data. *23rd International Conference on Computational Linguistics: Posters*, 3644.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic Computational Linguistics. *MIT Press*.

Celikyilmaz, A., Hakkani-Tur, D., & Feng, J. (2010). Probabilistic Model-Based Sentiment Analysis of Twitter Messages. *Spoken Language Technology Workshop (SLT)*, 7984. doi:10.1109/SLT.2010.5700826

Cortes, C., & Vapnik, V. (1995, September). Support-vector networks. *Machine Learning*, *20*(3), 273–297. doi:10.1007/BF00994018

Cramer, J. S. (2002). The origins of logistic regression (Technical report). *Tinbergen Institute.*, *119*, 167–178.

Domingos, P., & Pazzani, M. (1997). On the optimality of the Simple Bayesian classifier under zero-one loss. *Machine Learning, 29*, 2–3.

Kamal, M.S., Chowdhury, L., Khan, M.I., Ashour, A.S., Tavares, J.M.R.S., & Dey, N. (2017). Hidden Markov Model and Chapman Kolmogrov for Protein Structures Prediction from Images. Computational Biology and Chemistry, 68, 231–244. 10.1016/j.compbiolchem.2017.04.003

Kamal, S., Dey, N., Nimmy, S. F., Ripon, S. H., Ali, N. Y., Ashour, A. S., & Shi, F. (2018). Evolutionary framework for coding area selection from cancer data. *Neural Computing & Applications*, *29*(4), 1015–1037. doi:10.1007/s00521-016-2513-3

Kamal, S., Ripon, S. H., Dey, N., Ashour, A. S., & Santhi, V. (2016). A MapReduce approach to diminish imbalance parameters for big deoxyribonucleic acid data set. *Computer Methods and Programs in Biomedicine*, *131*, 191–206. doi:10.1016/j.cmpb.2016.04.005 PMID:27265059

Khurana D., Koli A., Khatter K., & Singh S. (2017, Mar. 25). *Natural Language Processing*. Natural Language Processing RSS.

KotziasD. (2015). *UCI*. https://archive.ics.uci.edu/ml/data sets/Sentiment+Labelled+Sentences

Mejova, Y. (2009). Sentiment analysis: An overview. Academic Press.

Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in Twitter using Machine Learning Techniques. *4th ICCCNT*.

Niu, Z., Yin, Z., & Kong, X. (2012). Sentiment classification for microblog by machine learning. *Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on*, 286–289. doi:10.1109/ICCIS.2012.276

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion mining. *Proceedings of LREC*.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up! Sentiment classification using machine learning techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 79-86.

Peddinti, V., Chintalapoodi, P., & Kiran, V. M. (2011). Domain adaptation in sentiment analysis of twitter. In *Analyzing Microtext Workshop*. AAAI.

Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, *27*(3), 221234. doi:10.1016/S0020-7373(87)80053-6

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge University Press.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 7789. doi:10.1016/S0034-4257(97)00083-7

Witten, I. H., Eibe, F., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Morgan Kaufmann.

Wu, Y., & Ren, F. (2011). Learning sentimental influence in twitter. *Future Computer Sciences and Application (ICFCSA), 2011 International Conference*, *119-122*.