JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

An Efficient Approach for Sentiment Analysis Using Regression Analysis Technique

Rajit Nair[1*], Vaibhav Jain[2], Amit Bhagat[3], Ratish Agarwal[4]

[1,2]Department of Computer Science, School of Engineering and Technology, Jagran Lakecity University, Bhopal, India
[3]Department of Computer Application, Maulana Azad National Institute of Technology, Bhopal, India
[4]Department of Information Technology, University Institute of Technology, RGPV, Bhopal, India

[*]*Corresponding Author: rajit.nair@jlu.edu.in, Tel.: +91-7000760748*

DOI: https://doi.org/10.26438/ijcse/v7i3.161165 | Available online at: www.ijcseonline.org

Accepted: 10/Mar/2019, Published: 31/Mar/2019

***Abstract*—** Sentiment Analysis is one of the major areas in text analytics. It primarily focuses on the recognition and categorization of opinions. Sentiment analysis is the way by which we mine the reviews given by people on different events, products, movies and many more. People rely on the reviews provided by the users of the product before shopping. Likewise people depend on the reviews of a movie before watching it. In this work, we have shown how regression algorithm work on the sentiment analysis of movie reviews and we also which regression algorithm is better for sentiment analysis. The regression algorithm which we have implemented is Random Forest, Ridge, Linear and ElasticNet. The dataset which we used for sentiment analysis is based on movie reviews also known as IMDB dataset and the parameters which we have used for analysis is mean square error and R squared error. From the result, it can be easily concluded that regression analysis with the best accuracy can be considered as a benchmark for all the other algorithms.

***Keywords*—**Sentiment Analysis, Regression, Naïve Bayes, Random forest, Features

I. INTRODUCTION

The continuous growth in the area of web technology has replaced the manner by which people can express their perspectives. People rely on the reviews provided by the users of the product before shopping. Likewise people depend on the reviews of a movie before watching it. There are various ways for connecting the users using web technology. Posts including hash tags on various social media platforms like facebook, instagram, twitter etc. help the users to connect together and to explore the information about various latest trends. The amount of data is expanding day by day. Hence it is difficult for the users to analyze it and to reach upon some conclusion.

Sentimental analysis primarily focuses on the recognition and categorization of opinions. It can be performed in two ways; using knowledge based approach and the other machine learning techniques [1]. In the first approach, there is a requirement of the huge database consist of already defined emotions and effective representation of the knowledge. The machine learning approach utilizes the trained and test datasets to design a classifier. Therefore it is easier than the knowledge based approach.

There are various types of challenges which were found in the field of sentimental analysis [2]. One of the major challenges is that the word which expresses the opinion may be positive or negative relying upon the circumstances.

There are various types of challenges which were found in the field of sentimental analysis. One of the major challenges is to classify the word which expresses the opinion may be positive or negative relying upon the circumstances. Another major challenge is that the way of expressing the opinion of people may not be the same. The relation between textual reviews and the consequences of those reviews can be obtained using opinion mining.

II. SENTIMENT ANALYSIS

Sentimental analysis can be performed to distinguish customers and followers on the basis of their perspective towards product or movie using their reviews. With the help of sentimental analysis it can be easily predicted that the user is satisfied with the product or not. The reviews may be positive or negative based on the experience of the user.

Major steps for performing sentimental analysis:

- Preprocessing of datasets
- Feature Extraction & Feature Selection

- Classification Model

2.1 Preprocessing of datasets

The database collected from the real world usually possesses a requirement of preprocessing as it is consists of noisy and incomplete data [3]. For improving the efficiency and accuracy, the data should be clean before any further processing. There are many techniques to perform the preprocessing of the data. For sentimental analysis, cleaning and preparing the data for classifications are two major techniques for pre-processing of the data. Other preprocessing techniques include integration [4], aggregation, discretization and transformation. One of the examples of preprocessing of the data can be taken as removing of URL's punctuations, symbols, stop words and emoticons from the tweets or posts [5].

2.2 Feature Extraction & Feature Selection

Dimensionality reduction is process of reducing the number of random variables in the data, thereby including only most significant information [6]. The data which has been collected for any specific problem can have a number of attributes. In other words it can have lots of dimensions. It is possible that, all of these attributes or dimensions may not influence the output equally. If large number of attributes or features have an impact on the computation, then it can cause the problem of over fitting which may lead to the poor result.

In Feature Extraction method of dimensionality reduction, a new set of features is inherited from the original features. This method often consist an irreversible transformations over the features as some of the loss in information may occur. Principal component analysis is one of the techniques of feature extraction. It performs a linear transformation over the features. It retrieves the attributes which maximize the variance.

In Feature Selection method of dimensionality reduction, a subset of the original feature is taken in consideration. Important features are selected on the basis of correlation. It is also called Correlation based Features Selection (CBFS) [7]. Furthermore, Best First Search method is used in this technique in order to reduce the dimensionality [8].

2.3 Classification Model

The task of classification is performed in two parts. Firstly, a classifier is formed explaining a set of data classes which are predetermined. This step is considered as one of the significant learning step and is performed using the training data, where learning on training set is performed and thus a classifier is developed. The training set includes of either database tuples and their associated class labels or simple text. This is also called as Supervised Learning [9]. Some of the techniques of classification in Machine Learning have been given below.

A. Naïve Bayes

Naïve Bayes is one of the simplest supervised machine learning algorithms which results in an effective classification [10]. Naïve Bayes theorem is based on one of the important mathematical theorem called Bayes Theorem. It is based on the facts that the features should not correlate to each other. In case of any dependency, the features or attributes should independently affect the probability. Therefore it is called Naïve Bayes. Various researchers have obtained an accuracy of around 84% with 10 most significant features from the sentimental analysis dataset with the help of recursive features elimination based SVM.

B. Support Vector Machine

This algorithm has been emerged as another famous supervised learning algorithm which can be used as predictor and also as a classifier. It consists of a predefined target variable. In case of classification, a hyper-plane is detected to distinguish between the different classes. In this technique, the points in feature space represent the training points while the test data points are mapped to the same space. In such a manner these points are classified on the basis of the area in which they fall. As per our study, the researchers have implemented SVM model on the sentimental analysis dataset and have achieved around 99% accuracy.[11]

C. K – Nearest Neighbour

It is also a very simple but effective machine learning algorithm. It has been widely used for classification purpose without making any assumption [12]. It is used for classification purpose when there is very less or no prior knowledge available about the distribution of the data. It maps the k nearest data points among the training set to the data points for which a target value is not known and assigning the average value of the obtained data point to it. Researchers have found an accuracy of around 87% in predicting the heart disease as per our studies using the value of k as 9. KNN has been optimized with the help of one of the famous optimization technique called Ant colony optimization. Researchers have achieved an accuracy of 87% in sentimental analysis with such combination which is very effective.

D. Decision Tree

Another supervise learning algorithm called Decision tree, which is widely used for classification purpose problems. It achieves very effective results using the attributes which are continuous and categorical in nature. In decision tree algorithm, the population is divided into two or more parts based on the most relevant predictors. Decision tree algorithm works in two recursive steps. In the first step it calculates the entropy of each and every attribute. In second step, the dataset is divided with the help of the variables or predictors. These predictors possess maximum information gain or minimum entropy. Ass per our study, in worst case the researchers have obtained an accuracy of 77% with the help of decision tree while in best case decision tree have achieved an accuracy of around 83% to predict sentimental features [13].

E. Random Forest

Random Forest is another famous supervised machine learning algorithm. It can be implemented for both regression and classification purpose but usually performs better in case of classification purpose. It is called random because there are multiple decision trees which are taken in consideration before providing the output. So, it can be viewed as group of decision trees. This technique is relying on the fact that a right decision will be obtained with more number of trees. In case of classification, it utilizes a voting system and then selects the class, while in case of regression it takes the mean of all the outputs of each of the decision trees. It is suitable for the problems where large datasets are used with high dimensionality. Random forest methods have achieved outstanding results in order to predict the heart diseases. The researchers have achieved around 97% accuracy in prediction of the sentimental analysis [14] .

In, Neethu M S and Rajasree R [15] authors have implemented four distinguish algorithms included Naive Bayes, SVM, Maximum Entropy and Ensemble classifiers. Almost similar accuracy has been achieved using these algorithms. Naive Bayes has provided comparatively improved precision with respect to other classifiers. It was slightly lower accuracy and recall. The accuracy, precision and recall achieved by using SVM were almost same as Maximum Entropy Classifier and they have provided 90% accuracy. The Naive Bayes has achieved an accuracy of 89.5%.

Akshay Amolik and Dr.M.Venkatesan [16] have implemented two different machine learning classifiers which are Naïve Bayes and Support Vector Machine. Both of them provided better accuracy. It can be observed from the result that the authors achieved a 5 % accuracy using SVM and 65% accuracy form Naïve Bayesian classifier. They concluded that the accuracy of classification can be increased by increasing the training data.

Nagamma P and Pruthvi H.R. [17] has described classification through clustering which provided better accuracy. The data required for prediction can be reduced by implementing clustering. Therefore, classification using or without using clustering produced the same result. It is because of less data which is used for prediction is small. They have summarized the effect of implementing clustering with classification model could be seen predominantly if the dataset used is large.

III. Algorithms used for Sentiment Analysis

The proposed technique involves the construction of a “Regression Algorithm” and the problem has been studied intensively. The regression algorithm is basically used for prediction and in this work prediction is related to movie reviews means we are predicting the sentiment based on movie reviews. Here we have implemented four approaches; they are Random Forest [14], Ridge [18], Linear [19] and ElasticNet regression algorithm [20]. The parametric measures are mean square error [21] and R square score [22].

Mean Square Error - In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated. It is a performance measure and it is mostly used for regression problems. In case of classification, accuracy is a more appropriate measure. A small mse means good prediction and large means bad model.

R Squared score - R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determinations for multiple regressions. 0% indicates that the model explains none of the variability of the response data around its mean.

IV. Proposed Work

Normally regression are used for prediction, previous work in regression are mostly done in the area of financial, weather and market analysis and it has given a significant result in these areas. So this time we have introduce the regression algorithm in the area of sentiment analysis and that is related to movies. In our work linear regression has show a better result than other algorithms which are already working in this area, we are saying better because the parameter which we have taken is mean square error and that should be less. The regression algorithm is said to be better when the mean square error is less. The regression algorithm is basically used for prediction and in this work prediction is related to movie reviews means we are predicting the sentiment based on movie reviews.

Proposed Algorithm

1. Import all the libraries including numpy, pandas and many more
2. Import sklearn library
3. Initilalize Imdb dataset

   ```
   IMDB=pd.read_csv('IMDB-Movie-Data.csv')
   ```
4. Initialize feature and prediction list

   ```
   Features_List=["Rank", "Year", "Runtime", "Rating", "Votes", "Revenue"]
   ```

   ```
   Predict_List=["Metascore"]
   ```
5. Adding the missing values

   ```
   imputer_features=Imputer(missing_values="NaN", strategy="mean", axis=0)
   ```

   ```
   imputer_predict=Imputer(missing_values="NaN", strategy="mean", axis=0)
   ```
6. Splitting the train and predict dataset

   ```
   Features_Train, Features_Test, Predict_Train, Predict_Test=train_test_split(Features_Scaled, Predict_Scaled, test_size=0.2, random_state=0)
   ```
7. Now initialize our predictor with our regression

   ```
   Predictor=LinearRegression()
   ```
8. Fitting the regression on training dataset

   ```
   Predictor.fit(Features_Train, Predict_Train)
   ```
9. Calulating the metascore

   ```
   Metascore_Predicted=Metascore_Predictor.predict(Features_Test)
   ```
10. Calculating mean square error and r score of the regression model

    ```
    Mean score error=mean_squared_error(Predict_Test, Metascore_Predicted)
    ```

    ```
    R squared Score  = r2_score(Predict_Test, Metascore_Predicted3)
    ```

    Print Mean score error and R squared score

V. Results

The files are in the form of comma separated files which are also known as csv files and these are taken from IMDB dataset [23]. This dataset contain the features as Rank, Title, Genre, Description, Director, Actors, Year, Runtime (Minutes), Rating, Votes, Revenue (Millions) and Metascore. Out of these features Metascore is our target variable and the regression algorithm is for metascore prediction. Given below results of regression algorithms which has been implemented using Python 3 [24]:

According to Random Forest Regression

Mean Square error of the Model: 0.7886989988718577

R squared Score of the Model: 0.2534653796813343

According to Ridge Regression

Mean Square error of the Model: 0.720183446017948

R squared Score of the Model: 0.31831804503133465

According to linear Regression

Mean Square error of the Model: 0.7200778731753033

R squared Score of the Model: 0.3184179738788576

According to Elastic Net Regression

Mean Square error of the Model: 0.9778489328180965

R squared Score of the Model: 0.07442752832887123

	Mean Square Error	***Models***	***R-Squared Score***
0	0.788699	Random Forest	0.253465
1	0.720183	Ridge	0.318318
2	0.720078	Linear	0.318418
3	0.977849	Elastic Net	0.074428

Table 1: Results Obtained

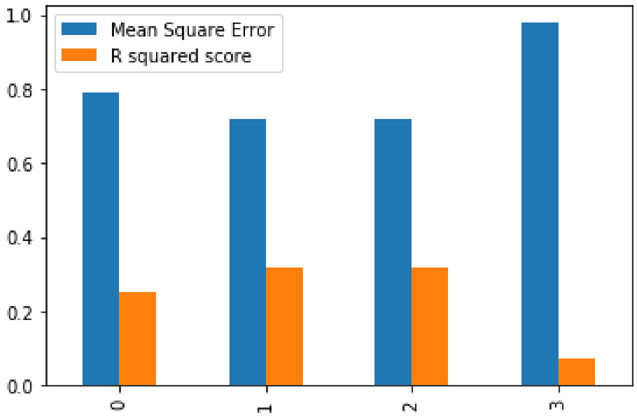

Figure 1: Results Obtained in Graphical Form

In this graph 0 is for Random Regression, 1 is for Ridge, 2 is for Linear and 3 is for elastic net. It has observed that Ridge regression perform better than other regression algorithms.

VI. Conclusion

The paper mainly addresses the comparative study of different machine learning algorithms that can be used to extract sentiments from text. Along with detailed studies of sentimental analysis techniques, a effective technique for sentimental analysis has been used using regression analysis techniques have been implemented which provided better accuracy. The proposed technique is simpler and efficient. It can be easily concluded that regression analysis with the best accuracy can be considered as a benchmark for all the other algorithms. It provides enough information that could help to improve the predictions in further research work. It can be concluded that cleaner the data, better the performance of an algorithm in predicting the success rate of the movies.

REFERENCES

[1] S. H. Huddleston and G. G. Brown, "Machine learning," in *Informs Analytics Body of Knowledge*, 2018.

[2] P. H. Shahana and B. Omman, "Evaluation of features on sentimental analysis," in *Procedia Computer Science*, 2015.

[3] R. Nair and A. Bhagat, "A Life Cycle on Processing Large Dataset - LCPL Rajit Nair," vol. 179, no. 53, pp. 27–34, 2018.

[4] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *J. Eng. Appl. Sci.*, 2017.

[5] H. Saif, M. Fernández, Y. He, and H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of Twitter," in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, 2014.

[6] L. van der Maaten, E. Postma, and J. van den Herik, "Dimensionality Reduction: A Comparative Review," 2009.

[7] M. Hall, "Correlation-based feature selection for machine learning," *Diss. Univ. Waikato*, 1999.

[8] R. Dechter and J. Pearl, "Generalized best-first search strategies and the optimality af A*," *J. ACM*, 2002.

[9] "Supervised learning," in *Springer Tracts in Advanced Robotics*, 2010.

[10] D. Jurafsky and J. Martin, "Naive Bayes and Sentiment Classification," *Speech Lang. Process.*, 2017.

[11] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, 1995.

[12] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers," *Mult. Classif. Syst.*, 2007.

[13] S. R. Safavian and D. Landgrebe, "A Survey of Decision Tree Classifier Methodology," *IEEE Trans. Syst. Man Cybern.*, 1991.

[14] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R. P. Sheridan, and B. P. Feuston, "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling," *J. Chem. Inf. Comput. Sci.*, 2003.

[15] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," in *2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2013*, 2013.

[16] A. Amolik, N. Jivane, M. Bhandari, and M. Venkatesan, "Twitter sentiment analysis of movie reviews using machine learning technique," *Int. J. Eng. Technol.*, 2016.

[17] P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha, "An improved sentiment analysis of online movie reviews based on clustering for box-office prediction," 2015.

[18] G. C. McDonald, "Ridge regression," *Wiley Interdiscip. Rev. Comput. Stat.*, 2009.

[19] D. J. Olive, *Linear regression*. 2017.

[20] C. Hans, "Elastic net regression modeling with the orthant normal prior," *J. Am. Stat. Assoc.*, 2011.

[21] M. D. Schluchter, "Mean Square Error," in *Wiley StatsRef: Statistics Reference Online*, 2014.

[22] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, 2014.

[23] D. Demir, O. Kapralova, and H. Lai, "Predicting IMDB movie ratings using Google Trends," *Dept. Elect. Eng, Stanford Univ., Calif.*, 2012.

[24] M. Pilgrim, *Dive into python 3*. 2009.