International Journal of Research in Engineering, Science and Management
Volume-3, Issue-6, June-2020
www.ijresm.com | ISSN (Online): 2581-5792

A Review On Sentiment Analysis Using Machine Learning

R. Radhu Krishna[1], Aliena Shaju[2], Arya Jeshkumar[3], Melvi Ann Vinod[4], B. Nandana[5*]
[1]*Assistant Professor, Dept. of Computer Science and Engineering, College of Engineering, Chengannur, India*
[2,3,4,5]*Student, Dept. of Computer Science and Engineering, College of Engineering, Chengannur, India*
*Corresponding author: nandanabindu@gmail.com

***Abstract*: The aim of this review is to describe various models that can be used to find the proper sentiment of data. The work that was done on examining the various techniques that can be used to find the proper sense of a data are described. The techniques include SVM, Naive Bayes, Decision tree, ELM and Logistic Regression. This review describes models along with its performance in the test phase. It was found that Logistic regression produced the best results out of all the models used.**

***Keywords*: Decision tree, Logistic regression, Sentiment analysis, SVM.**

1. Introduction

With the development of the internet, there are many online channels through which users can share their viewpoints and opinions regarding products, services and brands. They could share the opinion with their friends, family, and others as well. This data contains large amounts of product WOM documents. The WOM documents are of great importance at present. From this data, it is possible to identify the correct views and preferences of the consumer. So for businesses and commercial organizations, it is very important to identify the proper sense of this documents. This would help them provide suitable recommendations to the consumers. Also consumers could refer to previous customer experience. But Word of Mouth documents are often ambiguous. Different words have different senses in different context. So analyzing and finding the proper sentiment of this documents using machine learning techniques is of great significance.

2. Methodology

A. SVM

Real data is often messy and cannot be separated easily with a hyper plane. So data preprocessing must be done. It involves dividing the data into attributes and labels and dividing the data into training and testing sets. The Support Vector Machine algorithm represents the labeled training examples as points in the space and separates them with a hyper plane.

SVM is used to discover the best hyper plane that separates the data points that belongs to two different labels. The hyper plane thus acts as a decision boundary for any new example. The data points that are nearer to the hyperplane are called support vectors and it affects the position of the hyperplane. These support vectors, are used to maximize the margin of the classifier.

After training, new examples are mapped into the same space and classified based on the side of the hyper plane they belong to. SVM can be generalized to categorical output variables that can take more than two values.

B. Naive Bayes

The process of Naive Bayesian is that we assume big about how we can calculate the probability of occurrence of a document. It is equal to the product of the probabilities of each word within its occurrence. This specifies that there is no linkage between one word and another word. The probability of occurrence of a word can be found using a positive or negative sentiment and counting the number of times it occur in each class. Estimation is done by using: P(token | sentiment) as: count(this token in class) + 1/count(all tokens in class) + count(all tokens).

The classify function executes by calculating the prior probability based on the number of positive and negative examples. Then tokenization of the incoming document is done. For each class multiply together the likelihood of each word being seen in that class. Then we sort the final result and return the highest scoring class. We classify the polarity of the status update in a sentence level since it is more accurate.

C. Decision Tree

Decision Tree is a classification algorithm which represents a tree. It consists of root node, internal node and leaf node. Root and internal node indicate the feature of instance data while leaf node represents the class. Each split of node represent values of the feature presented by the split node.

Determination of which feature is being used as the root and internal node for each depth is a big problem. One method is by choosing feature with highest information gain. Information gain represent how important is a feature classifying the data ranging from 0 to 1. The information gain value close to 1 is more significant. Information gain can be calculated by:

$$InfoGain(C, f) = Entropy(C) - Entropy(C|f)$$

- C is class
- f is feature

VALUE OF ACCURACY, PRECISION AND RECALL AVERAGE USING RAPIDMINER

Method	Accuracy	Precision	Recall
Decision Tree	80 %	79.96 %	84 %

We got the accuracy of Decision Tree as 80%. Precision of Decision Tree amounted to 79.96%. The result for recall from the Decision Tree is 84%. So it can be seen that the Decision Tree classifier is average for use with social media datasets because it doesn't provide much accurate and precise predictions. Difference is there in the results of the previous research, showing the accuracy of Decision Tree Classifier of 64.42%. The difference is because of the characteristics of different datasets and processes.

D. ELM

Extreme Learning Machine (ELM) is a method proposed for single hidden layer feed-forward networks (SLFNs). The ELM employs a feed-forward neural network architecture and it works with randomly determined input weights. In this aspect, ELM depends on the principle that helps to determine the weights and biases in the network. In the first phase of ELM which can be called feature mapping, random values are used that distinguishes it from other methods such as Support Vector Machines (SVM) and Deep Neural Networks that uses a kernel function for this purpose.

The primary goal of the ELM after the feature mapping step, is to learn weights between hidden and output layers with the aid of minimizing the error. Therefore, the ELM has gained a great deal more popularity recently, and can be carried out for solving various problems like regression, classification and dimension reduction. To determine the performance of the ELM used for sentiment analysis, we compare it with SVM, one of the most successful machine learning algorithms used for sentiment analysis. From the results obtained, the ELM classifiers outperformed the SVM classifiers in the majority of cases, and even obtained the best accuracy for both datasets. This shows the importance of the ELM classifiers which can achieve a high performance, which is comparable and most often greater than most popular and cutting edge methods in the field of text classification.

Algorithm for ELM is as:

for a given training set ℵ =(xi, ti)|xi ε Rn, ti ε Rm, i = 1, ..., N, it's activation function f(x) and number of hidden nodes K:

1. Randomly assign input weights aj and hidden layer biases bj, j = 1, ..., K
2. Then Calculate the hidden layer output matrix H
3. Calculate the output weight β as , β = H+T Hence the ELM classifier produces m outputs say z = [z1, z2, ..., zm] for a pattern sample x. It can be represented as follows:

   z=Σ βjf(x,aj ,bj)

To which class label, the pattern x is belongs to can be determined by the output node which is having the largest output value. Mathematically, zˆ(x) = argmaxj=1,...,mzj (x) zˆ(x)zˆ(x) = argmaxj=1,...,mzj (x)

E. Logistic Regression

Logistic regression is a fundamental classification technique which belongs to the group of linear classifiers. It is widely used when the dependent variable is binary. It is used for predictive analysis to explain the relationship between one dependent binary variable and a set of independent predictor variables. It is a regression model but can be used for classification problems when thresholds are used on the probabilities predicted for each class.

Logistic regression being a linear classifier uses a linear function called logit.

$f(\mathbf{x}) = b_0 + b_1x_1 + \cdots + b_rx_r$,

The values b_0, b_1, …, b_r are the predicted weights of the regression model

The logistic regression model uses the sigmoid function of $f(\mathbf{x})$: $p(\mathbf{x}) = 1 / (1 + \exp(-f(\mathbf{x}))$. The sigmoid function has values very close to either 0 or 1 across most of its domain making it suitable for classification methods. The function $p(\mathbf{x})$ can be interpreted as the predicted probability that the output for a given $\mathbf{x}$ is 1 and $(1 - p(x))$ is the probability that the output is 0.

Logistic regression determines the best predicted weights such that the function $p(\mathbf{x})$ is as close as possible to all actual responses y_i, $i = 1, \ldots, n$, where n is the number of observations. In order to get the best weights, the log-likelihood function (LLF) for all observations $i = 1, \ldots, n$ are usually maximized using the maximum likelihood estimation.

The contributions of individual variable to the final fit can be better understood, and the outputs can be directly interpreted as probabilities which is a big advantage over models that can only provide the final classification. Tuning isn't necessary and outputs well-calibrated predicted probabilities. Logistic regression works better when the attributes unrelated to the

Table 1
Comparison table

Method	Regression	Accuracy	Training Time	Raw Implementation	Interpretability
Logistic Regression	Easy	High	Low	Easy	Better
SVM	Possible	High	High	Very difficult	Medium
Naive Bayes	Bad Performance	Low	Low	Medium	Good
Decision Tree	Inadequate to apply	Low	High	Difficult	Good
ELM	Difficult	Low	Low	Medium	Comparatively less

output variable as well as those that are very similar to each other are removed. It is a highly interpretable model.

3. Conclusion

Sentiment Analysis is a method to evaluate the sense in written or spoken language. The web based social platforms and pages like IMBD draws in a huge number of users that are online for imparting their insights in the form of reviews or comments. The reviews can be then classified into a positive or negative sense. The classifier used is Logistic regression classification with tf-idf for feature extraction. K fold cross-validation data mining technique is used for accuracy. From the analysis of the result, it is evident that the prediction given by Logistic regression model is more accurate than the other methods. The model gives an accuracy of 88.86.

References

[1] J. Kranjc, J. Smailovi C, Gr Car Podpe Can, M. Znidar si, N. Lavra, Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform, Inf. Process. Manage. 51 (2015) 187–203.

[2] A. Pak, P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010.

[3] S. Wahyuningsih, D. R. Utari, U. B. Luhur, D. Tree, and K. Validation, "Perbandingan Metode K-Nearest Neighbor, Naïve Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit," Konf. Nas. Sist. Inf. 2018, pp. 8–9, 2018.

[4] S. Poria, E. Cambria, A. Gelbukh, F. Bisio, A. Hussain, Sentiment data flow analysis by means of dynamic linguistic patterns, IEEE Comput. Intell. Mag. 10 (4), (2015) 26-36.