ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Sentiment Analysis for Movie Review

Shiva Gupta
M.Tech (CSE) Student
Faculty of Engineering and Technology,
Agra College, Agra.

1. ABSTRACT

Sentiment analysis is an emerging field of research to know peoples opinion about a particular service. Today's sentiment data is available in a large amount on social media in the form of blogs, updates, posts, tweets etc. Sentiment analysis can be performing on various machine learning techniques. Sentiment analysis refers to the emotions and the opinion of the user. in this paper we are proposing a sentiment analysis of latest movies with the help of random forest method .we are correctly classifying the comments as positive, negative and neutral. We are comparing our proposed method with the Support Vector Machine, Naïve Bayes classifier.

Keywords: ***Machine Learning, Support Vector Machine, Naïve Bayes, Random Forest.***

2. INTRODUCTION

Machine learning is a field of computer science that has evolved into one of the most powerful domain that has enabled computers to make decisions without being programmed explicitly. Machine learning and AI algorithms are based on mathematical optimization to understand and make accurate predictions on data. It is widely employed in search engines, recommendation systems, driverless cars, spam filtering, etc. The tasks handled by machine learning are typically classified under supervised learning (labeled data), unsupervised learning (unlabeled data) and reinforcement learning (computer interacts with a real time environment where it has to perform a goal without being taught on how to achieve it). Sentiment analysis (also called as Opinion mining) uses computational linguistic and natural language processing [1] models for text understanding. It faces challenges in short string texts, varying contexts and a myriad of opinions of individuals, which makes it extremely hard to analyze

3. DATA PRE-PROCESSING

There are some steps for data preprocessing these are as follows-

- Remove the HTML tags to get original text.
- Remove Punctuation marks and any other non-alphabetic symbols.
- Convert all data to lower case.

4. DATA CLASSIFICATION

During the classification and prediction stage, different classifiers can be used. In this paper, we have used three popular supervised learning classifiers namely, Naïve Bayes, Support vector machines and Random Forests.

A. Naïve Bayes Classifier:

Naïve Bayes [2] is a probabilistic classifier based on Bayes theorem, with the features being independent of each other. Each feature is considered to contribute to the probability of any given test instance to belong to a particular class. Consider ***n*** features to be represented as a vector:

$$\mathbf{x} = (x_1, \ldots, x_n) \quad \textbf{(1)}$$

The probabilities that the Naïve Bayes model [3] assigns to the ***k*** classes will be as follows:

$$p(C_k | x_1, \ldots, x_n) \quad \textbf{(2)}$$

Implementing Bayes theorem, we can determine the conditional probability of predicting the class given a feature.

$$b(C^k|\mathbf{x}) = \frac{b(\mathbf{x})}{b(C^k)\, b(\mathbf{x}|C^k)}. \quad \textbf{(3)}$$

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The Bayes classifier then designates a particular class label to a given test instance based on which is the most probable class. Since we have represented the data in terms of tf-idf vectors, we have used Multinomial variant of Naïve Bayes classifier. Let the distributed dataset have parameter vectors as:

$$\theta_y = (\theta_{y1}, \ldots, \theta_{yn}) \quad \textbf{(4)}$$

for every class ***y*** and ***n*** being the number of features (or the size of the vocabulary in our case).

In the above equation, θ_{yi} represents the probability of a feature ***i*** found in sample of class ***y*** given by P (**xi** | **y**). The parameter of the vector is optimized by a smoothing factor alpha (=1 in our case for Laplace smoothing) in the following equation:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \quad \textbf{(5)}$$

Where Nyiis count of occurrence of feature ***i*** found in a sample of class ***y*** and ***Ny*** is the total count of all the features occurring in a sample of class ***y***.

B. Support Vector Machine Classifier:

Support vector machines [4] are associated with learning algorithms which learn from data to decipher patterns in classification and regression analysis. SVM models aim to find a hyper plane that separates the data points lying in the different classes as wide as possible so that when a new sample comes in, it is classified based on which side of the gap they fall in.

The hyper plane equation for every class ***y*** and points ***x*** has the following constraints:

$$(\boldsymbol{w} \cdot \boldsymbol{xi} + b) \geq 1, \text{ if } \boldsymbol{yi} = 1 \quad \textbf{(6)}$$

$$(\boldsymbol{w} \cdot \boldsymbol{xi} + b) \leq -1, \text{ if } \boldsymbol{yi} = -1 \quad \textbf{(7)}$$

Where b is a constant, ***w*** is called the weight vector, and ||***w***|| is minimized to maximize the separation between the classes. While implementing SVM models, one must supply parameters such as C, gamma, etc. to obtain the highest accuracy keeping in mind the bias-variance tradeoff. Bias-variance [6] dilemma is frequently dealt with in supervised learning algorithms as we tend to generalize beyond the training set. Bias error leads to underfitting of the data as it misses important interaction between the features and classes. On the other hand, variance error leads to overfitting as it is highly sensitive to noise and fluctuations that may be present in the training set. The accuracy of a model is largely dependent on these parameters, and hence the optimum values are found using grid-search technique.

C. Random Forest Classifier:

Random Forests [5] is a powerful ensemble learning algorithm often used in classification tasks. It classifies based on the results obtained from the myriad of decision trees it generates while training, where the mode of the targeted outputs from each decision tree is the output of the forest. Since trees are known to overfit data as they have low bias and high variance, Random Forests tend to average out the multitude of decision trees.

Proposed approach focuses on a classifier model known as random forest. Few of the researchers have started using this model for sentiment classification but none of them focused on importance of hyper parameters. Random forest contains few set of hyper parameters which requires manual tuning.

5. RANDOM FOREST METHOD.

5.1 Introduction.

Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark [7]. The term came from random decision forests that were first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho[8] and Amit and Geman [9] in order to construct a collection of decision trees with controlled variation.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

5.2. How random forest works.

Each tree is grown as follows:

1. Random Record Selection**:** Each tree is trained on roughly 2/3rd of the total training data (exactly 63.2%) . Cases are drawn at random with replacement from the original data. This sample will be the training set for growing the tree.

2. Random Variable Selection**:**
Some predictor variables (say, m) are selected at random out of all the predictor variables and the best split on this m is used to split the node.

3. For each tree, using the leftover (36.8%) data, calculate the misclassification rate - out of bag (OOB) error rate**.** Aggregate error from all trees to determine overall OOB error rate for the classification.

4. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes over all the trees in the forest. For a binary dependent variable, the vote will be YES or NO, count up the YES votes. This is the RF score and the percent YES votes received is the predicted probability. In regression case, it is average of dependent variable.

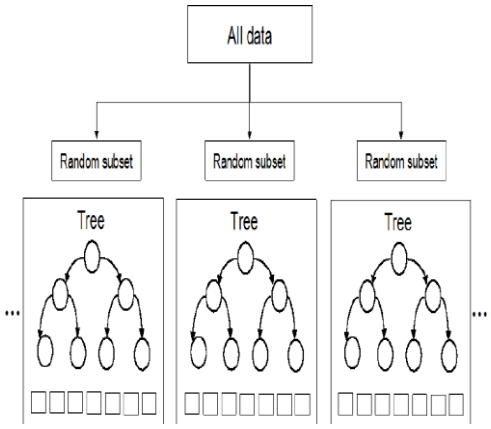

Figure: Working of Random Forest

6. COMPARISON AND RESULT.

6.1. Comparison.

Now at this stage we have compared various techniques Naive Bayes, Support Vector Machine with our proposed method. After comparison the conclusion is that our proposed method is best.

6.2. Result.

The previous results obtained for the Naïve Bayes, Support vector machine and Random Forest classifiers respectively when deployed on the testing set. As we can see, for positive polarity tweets Naïve Bayes classifier obtains the highest precision of 85% as compared to 82% and 81% for SVM and Random Forest classifiers, while in the case of negative polarity tweets we see that both Naïve Bayes as well as SVM attain precision >90% although SVM is slightly better at 92% precision with Random Forests achieving 89% precision [10].

Our proposed method is best as its result is best. Hence the **RANDOM FOREST** technique is best suitable as it give best result as compared to others.

TECHNIQUES	ACCURACY
Naïve bayes	91.2%
Support Vector Machine	87.9%
Random Forest	92.85%

Table1: Comparative Result

7. CONCLUSION

In this paper, we have analyzed the sentiment of social network comments. We provided a thorough comparison and performance analysis of the salient classification algorithms used in supervised learning namely Naïve Bayes, Support vector machine and Random Forests. For our standard movie reviews dataset, we can say that Random Forest was the most accurate with a score of 92.85% followed by SVM (87.9%) and Naïve Bayes (91.2%).Compared to previous works which have shown 89% accuracy in classification for Naïve Bayes and 88% for SVM and 85% for Random Forests.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

8. FUTURE SCOPE

We can even carry out the experiment using Latent sentiment analysis [11] techniques which employs Singular value decomposition (SVD) in word count matrix.

9. REFERENCES:

[1] Ralph Wieschdel, Jaime Carbonell, Barbara Grosz, Wendy Lehnert, Mitchell Marcus, Raymond Perrault, Robert Wilensky."White Paper on Natural Language Processing" Proc. October 1989 DARPA Speech and Natural Language Workshop, pp.481-493, October 1989.

[2] I. Rish, J. Hellerstein, and T. Jayram. An analysis of data characteristics that affect naive Bayes performance. TechnicalReport RC21993, IBM T.J. Watson Research Center, 2001.

[3] P. Langley, W. Iba, and K. Thompson. An analysis of Bayesianclassifiers. In Proceedings of the Tenth National Conference onArtificial Intelligence, pages 399–406, San Jose, CA, 1992. AAAI Press.

[4] Christopher J.C. Burges. A Tutorial on Support VectorMachines for Pattern Recognition. Data Mining and KnowledgeDiscovery 2, 121-167, 1998.

[5] Gerard Biau. Analysis of a Random Forests Model. Journal ofMachine Learning Research 13 (2012) 1063-1095

[6] Marina Sokolova, Nathalie Japkowicz, Stan Szpakowics.Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation. Association for the Advancement of artificial intelligence (2006)

[7] Breiman, Leo. "Random Forests" Machine Learning 45(1)32:doi:10.1023/A:1010933404324. 2001.

[8] Ho, Tin Kam. "The Random Subspace Method for Constructing Decision Forests". IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8): 832–844. doi:10.1109/34.709601. 1998

[9] R. Lippmann, D. Fried, I. Graf, J. Haines, K. Kendall, D. McClung, D. Weber, S. Webster, D. Wyschogrod, R. Cunningham and M. Zissman. Evaluating Intrusion Detection Systems: The 1998 DARPA Off-line Intrusion Detection Evaluation. In Proceedings of the DARPA Information Survivability Conference and Exposition, volume 2, pages 12–26. IEEE Computer Society Press, 2000.

[10] Anmol Nayak, Dr. S Natarajan , "Comparative study of Naïve Bayes, Support Vector Machine and Random Forest Classifiers in Sentiment Analysis of Twitter feeds", International Journal of Advanced Studies in Computer Science and Engineering, Volume 5, Issue 1, 2016.

[11] Lan Wang, Yuan Wan. Sentiment Classification of Documents Based on Latent Semantic Analysis. S.Lin and X. Huanh (Eds.): CESM 2011, Part II, CCIS 176, pp.356-361, 2011.