# Machine Learning Algorithms for movie reviews using Sentiment Analysis

M. Anil Shukla [1] ,M. Vatsalya [2], N. Teenu Anand [3] , N. Pavan Kumar [4]
Assoc. Prof. J. Madhu babu [5]

[1] Department Of Computer Science and Engineering, VVIT, Guntur

[2] Department Of Computer Science and Engineering, VVIT, Guntur

[3] Department Of Computer Science and Engineering, VVIT, Guntur

[4] Department Of Computer Science and Engineering, VVIT, Guntur

[5] Associate Professor, Department Of Computer Science and Engineering, VVIT, Guntur

----------*********----------

## ABSTRACT

Sentiment analysis is the process of determining whether a piece of writing is positive, negative or neutral. It is carried out to see the level of public sentiment or public opinion relating to goods or services and even a figure, both political and celebrity figures. The accuracy for Logistic Regression is more compared to Naive Bayes classifier. There are several steps taken to conduct this sentiment analysis, which is to collect data using libraries in python, text processing, testing training data, and text classification using logistic regression method.

From this classification or model it is possible to find whether the movie has either positive or negative or neutral reviews of the movie. Here, the dataset is taken from the movie dataset, which contains the review or the feedback of the movie. Then the combined data was tested from the training data used for each presidential candidate to get an accuracy.

**Keywords**: Sentiment analysis, logistic regression algorithm, text mining.

## 1. Introduction

As online marketplaces have been popular during the past decades, the online sellers

and merchants ask their purchasers to share their opinions about the products they have bought. Everyday millions of reviews are generated all over the Internet about different products, services and places. This has made the Internet the most important source of getting ideas and opinions about a product or a service. However, as the number of reviews available for a product grows, it is becoming more difficult for a potential consumer to make a good decision on whether to buy the product. Different opinions about the same product on one hand and ambiguous reviews on the other hand makes customers more confused to get the right decision. Here the need for analyzing these contents seems crucial for all e-commerce businesses. Sentiment analysis and classification is a computational study which attempts to address this problem by extracting subjective information from the given texts in natural language, such as opinions and sentiments. Different approaches have been used to tackle this problem from natural language processing, text analysis, computational linguistics, and biometrics. In recent years, Machine learning methods have become popular in semantic and review analysis for their simplicity and accuracy. Amazon is one of the e-commerce giants that people are using every day for online purchases where they can read thousands of reviews dropped by other customers about their desired products. These reviews provide valuable opinions about a product such as its property, quality and recommendations which helps the purchasers to understand almost every detail of a product.

This is not only beneficial for consumers but also helps sellers who are manufacturing their own products to understand the consumers and their needs better. This project is considering the sentiment classification problem for online reviews using supervised approaches to determine the overall semantic of customer reviews by classifying them into positive and negative sentiment.

In today's world there is a lot of impact being made by technology in daily life. There has been tremendous growth in adoption of new technologies in research and development. Technology has been developed to such an extent where it has become a part of our lives. The advancements in Web were made to such a large extent, as a matter of which there is an enormous increase in the volume of sentimental content accessible in the Web. Such a variety of information is found day in day out in the social web / websites / public network in the semblance of movie reviews or product ratings, customer statements, testimonials, critiques in discussion forums etc. By using this kind of information collected from the web in a proper way using appropriate technology, one can bring a huge change in the market by understanding the market trends and

companies or producers can customize their product as preferred by customers. This type of analysis is called Sentiment Analysis.

In today's world, it has become customary to collect opinions and reviews from people through various surveys, polls, social media platforms and analyse them in order to understand the preferences of customers. So, in order to understand the sentiments of customers and their view on the services offered by producers, there comes the need for an accurate and canonical mechanism for speculating and anticipating sentiments which possess the ability to fabricate a positive or negative impact in the market and thus making this kind of analysis important for the pair of producers and consumers. In this paper, the main focus is to anatomize the reviews conveyed by viewers on various movies and to use this analysis to understand the customers' sentiments and market behaviour for better customer experience. This research intends to analyse the reviews of customers on various movies by implementing three algorithms namely K Nearest Neighbours, Logistic Regression and Naive Bayes and provides conclusive remarks.

## 2. LITERATURE REVIEW

### 2.1 To collect the data and train the classifier

Saif, Hassan; He, Yulan and Alani, Harith a method was developed that can automatically collect the corpus and train the sentiment classifier based on multinomial naïve Bayes uses 2 features i.e., n-gram and POS-tagsand the classifier determines the classes of each tweet. This is to prove that a developed technique is more efficient than previously proposed methods.

### 2.2 Overview of Machine Learning techniques

Mejova, Yelena discusses about the overview of the sentiment analysis i.e., they give brief description about the word sentiment and novelty methods that are used to perform analysis on emoticoncharacter text by covering all the challenges that are faced during analysis of product or text etc. Leena A Deshpande, M.R. Narasinga Rao developed a method that determines the variance in the data and retrain the model confer to the drift that is identified. They have used 2 techniques weight-based features and nGram that identifies the unnamed labels in the text which improves its accuracy. T. Sajana, M.R.Narasingarao they have performed survey on detection and prediction of malaria disease using various machine learning techniques, Image Processing techniques. They have observed that machine learning techniques are mostly applicable for critical diagnosis of malaria

### 2.3 Applying Machine Learning techniques

Sayali P. Nazare, Prasad S. Nar, Akshay S. Phate, Prof.Dr. D. R. Ingle have created a dataset by twitter API and collected all tweets regarding the topic blue whale game. Their main aim is to perform analysis on sentimental tweets. They have used Naïve Bayes,Support vector machines, Maximum entropy and Ensemble classifier. SVM and Naive Bayes classifiers are implemented using MATLAB built-in functions. Maximum Entropy classifier is implemented using MaxEntsoftware. Based on comparative results Naïve Bayes has better precision and slightly lower recall and accuracy i.e., 89% and other classifiers are having similar accuracy levels i.e., 90%. The result shows the pie chart which is representing the positive,negative and neutral hashtags with percentages.

## 3. Proposed Solution

### 3.1 Dataset

The dataset can be obtained from the IMDB Dataset of 50K Movie Reviews. The dataset contains 50 thousand movie reviews that have been pre-labeled with negative and positive sentiment class labels.

### 3.2 Data pre-processing

Data preprocessing is the main step to prepare the data as of model requirements. The preliminary step for any dataset is to be pre-processed while applying to any algorithm. This is done by removing unwanted symbols, words or tags. These words or symbols do not affect the result but consumes more time to be executed i.e may slow down the processing of an algorithm. Our dataset involves the following steps

**A) Cleaning the text**

The text usually contains the html tags, which are considered as the unnecessary content , which do not have meaningful weight for analysing sentiment. So it is important to make sure that this unnecessary content should be removed. In this phase the removal of noisy text should be performed. The noisy data includes html tags, strips and brackets which doesn't manipulate the result. In this phase, the special characters are also removed.

**B) Stemming**

Stemming is the process of removing regular or frequent morphological endings from the english words. This process stems the words to root words. For example, cool,cooler,coolest are stemmed to its root word cool.

**C) Remove stop words**

This method removes the most frequently occurring or repeating words, these words

are considered as stop words which doesn't have more use but consumes time, this reduces the size of the dataset which is also known as stopping. As, an, the, is, etc are considered as stop words.

### 3.3 Training and Testing the data

The whole dataset is divided into training dataset and testing dataset, in which we train the algorithm with the training dataset and test the algorithm with the testing dataset. Test dataset is used later to determine the efficiency of each algorithm.

## 4. Applying ML Algorithms

Each researcher has worked on implementing each of the algorithms. So, the used classifiers or algorithms are Logistic Regression, KNN classifier, Multinomial Naive Baysian, SVM classifier.

### A) Logistic Regression

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X.

Contrary to popular belief, logistic regression IS a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as "1". Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

$g(z) = \frac{1}{1+e^{-z}}$

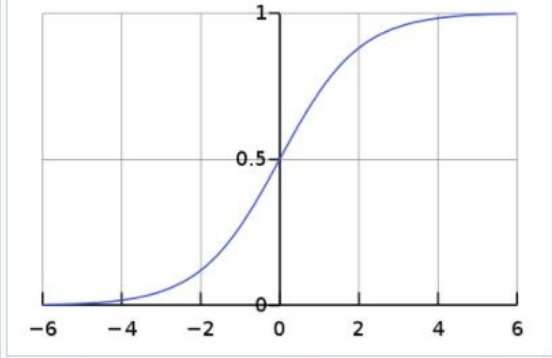


Figure 1. The standard logistic function $\sigma(t)$; note that $\sigma(t) \in (0,1)$ for all $t$.

### Algorithm:

1. The logistic regression algorithm is imported from the scikit-learn package.
2. Split data into training and test data.
3. Generate a logistic regression model.
4. Train or fit the data into the model.
5. Predict the review

### B) SVM Classifier

Support vector machines are a set of supervised learning methods used for classification, regression, and outliers detection.A simple linear SVM classifier works by making a straight line between two classes. That means all of the data points on one side of the line will represent a category and the data points on the other side of the line will be put into a different category.

This means there can be an infinite number of lines to choose from.

### C) Multinomial Naive Bayesian

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

Where $P(c|x)$ = Posterior Probability
$P(c)$ = Class Prior Priority
$P(x|c)$ = Likelihood
$P(x)$ = Predictor Prior Probability
$P(c|x) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$

### Algorithm:

1. Data Preprocessing step
2. Fitting Naive Bayes to the Training set
3. Predicting the test result
4. Test accuracy of the result(Creation of Confusion matrix)
5. Visualizing the test set result.

### D) K Nearest Neighbours Classifier

K Nearest Neighbours classifier is perhaps the simplest and most widely used machine learning algorithm. It can be applied to both classification problems and regression problems. For smaller datasets, it outperforms most of the other classifiers.
KNN can be implemented by finding a group of k objects which are nearest to the test object, and by assigning a label based on predominance of a class in the neighbourhood of the test object.

### Algorithm:

1. The k-nearest neighbor algorithm is imported from the scikit-learn package.
2. Create feature and target variables.
3. Split data into training and test data.
4. Generate a k-NN model using neighbors value.
5. Train or fit the data into the model.
6. Predict the review.

## 5. Results and observations

Logistic regression, SVM Classifier, Naive Bayes, and k-Nearest Neighbours are four types of machine learning classifiers used in this research. The operations of splitting the data set into two parts were tested for a data set which is obtained from preprocessing the dataset. The below table gives the accuracy score for each classifier that is performed on the movie dataset.

| Classifier | Accuracy | Execution Time (sec) |
|---|---|---|
| Logistic Regression | **89%** | **64 sec** |
| Support Vector Machine | **89%** | **3 sec** |
| Multinomial Naive Bayesian | **87%** | **1 sec** |
| K Nearest Neighbours | **77%** | **76 sec** |

## 6. Conclusion

The research paper has an effort to reveal that the classification techniques serve as powerful tools for sentimental analysing the data in Movie Review Data set.By applying different classification algorithms like Logistic regression, Support Vector Machine Classifier, Naive Bayes, and k-Nearest Neighbours it is proven that Logistic regression and Support Vector Machine gave better results in terms of accuracy measurement. But when compared with the factor "time taken for each algorithm to execute", Support Vector Machine took least time. Finally this research work helps in identifying better classifiers among Logistic regression, Support Vector Machine Classifier, Naive Bayes, and k-Nearest Neighbours for sentimental analysis of movie reviews.

## 7. References

*1. Leena A. Deshpande, M.R. Narasingarao "Addressing social popularity in twitter data using drift detection technique" Journal of Engineering Science and Technology Vol. 14, No. 2 (2019) 922 – 934.*

*2. Mejova, Yelena. (2019) "Sentiment Analysis: An Overview".*

*3. Sayali P. Nazare, Prasad S. Nar, Akshay S. Phate, Prof.Dr. D. R. Ingle"Sentiment Analysis in Twitter"International Research Journal of Engineering and Technology (IRJET)Volume: 05, Jan-2018.*

*4. EfthymiosKoulompis, TheresaWilson, Johanna Moore (2017),"Twitter Sentiment Analysis: The Good the Bad and the OMG!," in The Fifth International AAAIConference on Weblogs and Social Media. Pages 538-541.*

*5. Tiruveedhula, Sajana &ramanarasingarao, Manda. (2017). "Machine learning techniques for malaria disease diagnosis - A review" Journal of Advanced Research in Dynamical and Control Systems. 9. 349- 369.*

*6. HumaParveen and ShikhaPandey"Sentiment analysis on Twitter Dataset using Naive Bayes algorithm" 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) page 416-419 @article{Parveen2016SentimentAO}.*

*7. Anuja Prakash Jain and Padma Dandannavar "Application of machine learning techniques to sentiment analysis" 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology(iCATccT) pages 628-632 article{Jain2016ApplicationOM}.*

*8. Dey, Lopamudra&Chakraborty, Sanjay & Biswas, Anuraag& Bose, Beepa& Tiwari, Sweta. (2016). "Sentiment Analysis of Review Datasets Using Naïve Bayes and K-NN Classifier". International Journal of Information Engineering and Electronic Business. 8. 54-62. 10.5815/ijieeb.2016.04.07.*

*9. R. Dey and S. Chakraborty, "Convex-hull &DBSCAN clustering to predict future weather", 6 th International IEEE Conference and Workshop on Computing and Communication, Canada, 2016, pp.1-8.*

*10.Le B., Nguyen H. (2015) "Twitter Sentiment Analysis Using Machine Learning Techniques". In: Le Thi H., Nguyen N., Do T. (eds) Advanced Computational Methods for Knowledge Engineering. Advances in Intelligent Systems and Computing, vol 358. Springer, Cham.*

*11. Gaurav D Rajurkar, Rajeshwari M Goudar, "A speedy data uploading approach for Twitter Trend And Sentiment Analysis using HADOOP", HADOOP, 2015 International Conference on Computing Communication Control and Automation. Pages 580-584*