International Journal of Engineering Sciences & Emerging Technologies, Oct. 2023.
ISSN: 22316604 Volume 11, Issue 2, pp: 405-410 ©IJESET

SENTIMENT ANALYSIS ON IMDB DATASET USING ML

Jayati Bhardwaj[1],Meenakshi Yadav[2], Divyanshu Gupta[3], Gaurvi Bhardwaj[4]
[1,2]Asst. Prof., CSE Department
MIT Moradabad, UP, India
[3,4] Student, CSE Department
MIT Moradabad, UP, India
[1]jayatibhardwaj2@gmail.com, [2]meenakshiyadav2309@gmail.com
[3]divyanshuguptarmp790@gmail.com,[4] gaurvibhardwaj14@gmail.com

ABSTRACT

Analyzing the sentiments out of the text is one of the current emerging area of research which analyzes a large amount of data to provide insights on specific problem statement. Extracting emotion out of the text is the main work in this framework. Many of the NLP & ML techniques are able to perform this work with promising accuracy results. This paper presents three classifiers approaches namely Logistic regression, SVM and Naïve Bayes that are used to classify the emotion of IMDb movie reviews. Classification accuracy is used as the performance measure where the test accuracy of the best classification model turns out to be 89.2%. Along with this, a comparative analysis is carried out among the implemented approaches based on the other performance parameter matrices.

KEYWORDS*—NLP (Natural language Processing), ML (Machine Learning), SVM (Support Vector Machines)*

1. INTRODUCTION

The sentiment analysis is the method of processing and computing etymological data in order to extract opinions or emotions out of the text. It basically merges both the NLP and ML domains under classification problem [1]. Sentiment analysis attracts wide range of human computer interaction applications as business analytics and opinion mining. There are two general approaches for the sentiment analysis- first one, Sentiment analysis based on Lexicon & second one ML based sentiment analysis. In Lexicon approach the text is splitted into tokens, occurrences of each word are counted and then each word is subjectified as per the existing lexicons. While the later one approach uses more sophisticated and complex system. This approach applies training on different classifiers with a data set which is known as training set. Further by using the testing data set, the performance of the classifier is tested i.e how well the classification is being done by the model [2][3]. Beside these some other approaches like Vector learning [4], Ontology based approach [5] and voting ensemble [6]. Although there are many challenges related to the field and are currently under the research domain. Key problems like ambiguity in the meaning of the keyword in reference to the context of the mentioned text and inaccuracy in extracting the exact emotion of the sentence that contains no keyword related to emotion are the major concerns in the work.

In this paper, different classification approaches are applied to analyze the IMDb dataset of the movie reviews. The movie review dataset contains subdivision of data sets training and testing. The former part is used for the different classifiers. Afterwards, the later set is used to calculate the precision of classification of each classifying approach. The respective Accuracy and confusion matrix results are being drawn and compared to outline the best approach for carrying out analysis.

2. METHODOLOGY

2.1Objectives

The objective of the presented work is to build a model of high accuracy that train and identify the polarity of a given text to give positive or negative sentiments. Furthermore, it can easily automate the idea of determining how well did a movie performed on the box office by analyzing the public sentiments behind the movie's reviews which are gathered from a number of platforms. The work in this paper classifies the polarity of the review into two major classes namely positive and negative. The proposed system of ML based work for analyzing text sentiments is presented in the fig.1.

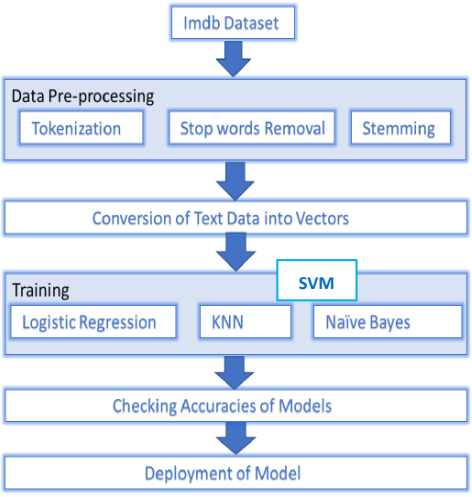

Fig. 1 Flowchart of the working methodology

2.b Dataset

This paper work is being implemented on the kaggle dataset containing 50k movie reviews from IMDb. In which the dataset is divided into both negative and positive reviews set of 25k each. The reviews present in the data set are divided into negative and positive in accordance to the rating system of IMDb. The sentiment distribution of the dataset from Kaggle is represented in the figure 2 below.

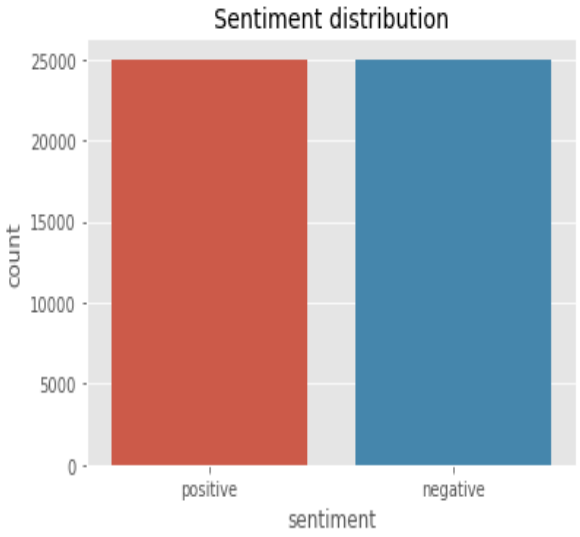

Fig.2 Sentiments distribution of the dataset.

IMDb allows viewers to rate on a scale from 1 to 10, and according to dataset creator any review ≤ 4 stars are labeled negative and ≥ 7 stars is marked as positive. Reviews with ratings out of the above ranges are not included. The glimpses of the positive and negative reviews are shown in table 1 below:

Table 1: Movie reviews and their sentiments

Review	Sentiment
Nice movie. Brilliant plot for a horror movie.	Positive
Worst movie ever….not gonna waste my money on movies like these	Negative
It was a fantastic thriller…a must watch	Positive
What a waste of time! Bad movie of all the times.	Negative

2.c Data Preprocessing

Data cleaning is a crucial step for meaningful analytics results. Data which is not preprocessed already and is also not organized have a possibility of resulting into misleading results. Cleaning of data prevents the errors or confusion that useless information may create during the classification. At first, special symbols are removed then letters are converted to lower case. Thirdly, the hybrid links are removed from text followed by removal of stop words. At last, stemming is applied to convert the words in their original form after removing suffixes and prefixes.

3. CLASSIFICATION MODELS

The analysis has been carried out on three classifiers namely, Naïve Bayes, Logistic Regression and SVM.

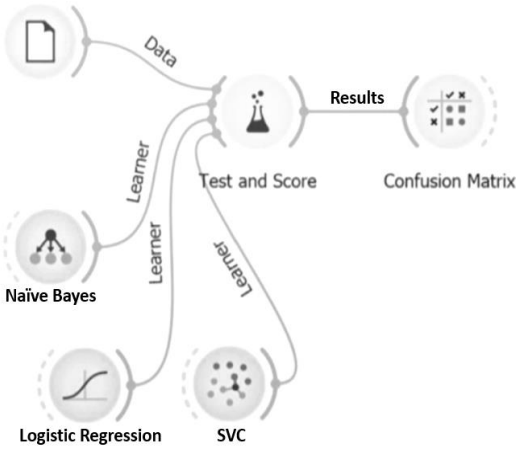

Fig. 3 The classification models

3.a Naïve Bayes

It's a supervised learning technique which is based on Bayes theorem principle. It is used in text classification in high dimensional space. Simplicity and efficiency of the algorithm make it a potential classification model to build ML model and perform predictions. It is a probabilistic classifier that uses probability of an object for classification [7]. The Bayes theorem is defined as follows:

$$P(A|B) = P(A)\, P(B|A)/P(B) \quad (1)$$

where, P(A|B) is Posterori Probability. Probability of A when event B happens.

P(B|A) is the Likelihood Probability. Probability of B when event A happens.

P(A) is Prior Probability.

P(B) is Marginal Probability.

The implemented work uses multinomial Naïve Bayes approach for predicting the labels of the text or email. It calculates the best likelihood of the sample and provides the label with greatest chance. It has its wide and popular application in the field of NLP [9].

3.b Logistic Regression

Logistic Regression is one of the classic supervised learning techniques. This classification method works same as linear regression. This method uses standard logistic function. The logistic function is basically a Sigmoid function. It takes real value (any between zero and one). It is defined as:

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}} \tag{2}$$

By using set of independent variables, it predicts the label of categorical dependent variable. Using discrete and continuous datasets, it can classify the new data. This model is capable to provide the probabilities of classes and perform classification [8].

3.c SVM

The support vector machine can be applied on both the regression and classification problems. It is the supervised learning approach which aims to find the optimal hyperplane in high dimensional space which can separate the data points in different spaces. The hyperplane tries to maximize the distance between the support vectors (closest points) of the classes. The hyperplane dimensions are dependent on the number of features in the data set [10].

3.d Performance Measures

The classification efficiency of any model can be generalized on the basis of the various performance measures/metrics. The major performance metrics on which the proposed models are being compared are mentioned below and their resulting quantified scores are summarized in figures 4, 5 and 6 for each corresponding model:

- ***Classifier Accuracy:*** A classifier accuracy defines how a model is classifying the data correctly or accurately i.e. in the most likeable polarity of the class. It is the percentage of the labels which are classified correctly as per the applied model. It is given by the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \times 100\% \tag{3}$$

In the given Equation (3) FP, FN, TP and TN respectively represents the false positives, false negatives true positives and true negatives, in the predicted resultant labels. The confusion matrix is depicted as follows in table 2:

Table 2: Confusion Matrix

		Predicted condition	
		Positive	**Negative**
Actual	**Positive**	TP	FP
	Negative	TN	FN

- ***Sensitivity & Specificity:*** Sensitivity defines the true positive rate i.e. probability of getting test result as positive. While, Specificity defines the negative true negative rate i.e. probability of getting test result as negative. Both the parameters are inversely proportional to each other[12].
- ***Precision & Recall:*** Both are evaluation parameters for binary and multiclass classification problems.[11] The respective formulas of both the metrics are:

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive + False Negative}} \tag{4) \& (5)}$$

- ***F1-score:*** It combines recall & precision and is defined as follows:

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall} \tag{6}$$

4. RESULTS & CONCLUSION

The sentiment analysis is being carried out on the IMDB movie reviews data set using the three classifier models namely SVM, Logistic Regression and Naïve Bayes. The resultant confusion matrices along with other comparative parameters of the models are shown below in the figures 4, 5 and 6 respectively. The resulting accuracies of the classifiers are 89.2%, 89% & 86.44% respectively. Hence on testing, the performance of the SVM model on the movie review data set turns out as the best when compared with the rest of the two models.

	precision	recall	f1-score	support
1	0.89	0.90	0.90	7513
2	0.90	0.88	0.89	7361
accuracy			0.89	14874
macro avg	0.89	0.89	0.89	14874
weighted avg	0.89	0.89	0.89	14874

Sensitivity : 0.9035005989617996
Specificity : 0.8845265588914549

Fig. 4 SVM confusion matrix and resulting parameters

	precision	recall	f1-score	support
1	0.87	0.86	0.86	7513
2	0.86	0.87	0.86	7361
accuracy			0.86	14874
macro avg	0.86	0.86	0.86	14874
weighted avg	0.86	0.86	0.86	14874

Sensitivity : 0.8570477838413417
Specificity : 0.8760578760578761

Fig. 5 Multinomial Naïve Bayes confusion matrix and resulting parameters

	precision	recall	f1-score	support
1	0.88	0.90	0.89	7513
2	0.90	0.88	0.89	7361
accuracy			0.89	14874
macro avg	0.89	0.89	0.89	14874
weighted avg	0.89	0.89	0.89	14874

Sensitivity : 0.9031012910954346
Specificity : 0.8766471946746366

Fig. 6 Logistic Regression confusion matrix and resulting parameters

REFERENCES

[1] Sailunaz, K., Dhaliwal, M., Rokne, J., & Alhajj, R. (2018). Emotion detection from text and speech: a survey. Social Network Analysis and Mining, 8(1), 28.

[2] Bandhakavi, A., Wiratunga, N., Padmanabhan, D., & Massie, S. (2017). Lexicon based feature extraction for emotion text classification. Pattern recognition letters, 93, 133-142.

[3] Saeed Mian Qaisar, Sentiment Analysis of IMDb Movie Reviews Using Long Short Term Memory, IEEE 2020.

[4] Joshi, P. (2017). Artificial intelligence with python. Packt

[5] Jiang, S., & Chen, Y. (2017, September). Hand Gesture Recognition by Using 3DCNN and LSTM with Adam Optimizer. In Pacific Rim Conference on Multimedia (pp. 743-753). Springer

[6] Onan, A., Korukoğlu, S., & Bulut, H. (2016). A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Systems with Applications, 62, 1-16.]

[7] https://plato.stanford.edu/entries/bayes-theorem/

[8] https://towardsdatascience.com/understanding-logistic-regression-9b02c2aec102.

[9] https://www.upgrad.com/blog/multinomial-naive-bayes-explained/

[10] https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[11] https://en.wikipedia.org/wiki/Precision_and_recall

[12] https://en.wikipedia.org/wiki/Sensitivity_and_specificity