



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: IV Month of publication: April 2025

DOI: <https://doi.org/10.22214/ijraset.2025.68593>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

Classifying User Reviews of Movie applications using Improved Logistic Regression

Yogesh Ramaswamy

Senior DevOps Engineer, Danbury, CT, USA

Abstract: In recent years review classification, analysis and prediction are one of the most common applications of sentiment analysis. It involves detection of sentiments on the reviews made by the users on social networking applications through opinion mining. In general, reviews can have positive, negative or neutral polarity indicators. For classification, the polarity indicators take the form of certain words and emotions that readily show the user's sentiments. Existing works fall short of producing accurate classification results because of two-class problem that affects the performance of evaluation parameters like precision, recall, accuracy and F-measure. Hence there is a need of an efficient classification technique which addresses two-class problem. This work proposes Improved version of Logistic Regression [ILR] that is commonly used for sentiment analysis and classification. The proposed classification technique identifies and replaces the misspelled words in the sentence, support count estimation and classification of reviews along with multiple independent words with similar meaning in parallel. The experimental results show the classification accuracy of the proposed technique to be more accurate compared to the existing logistic regression and naïve bayes classifiers.

Keywords: Sentimental Analysis, Machine Learning, Improved Logistic Regression, POSTagging and Movie Reviews.

I. INTRODUCTION

Data mining is a process of discovering specific patterns in huge data sets. It aims to convert the gathered data from a dataset into a comprehensible form for optimal usage. Web mining is an application of data mining strategies to find interesting patterns in the data which is downloaded from the web. Opinion mining is a sub-discipline of web mining that facilitates searching and discovering user's opinion about a specific topic or a product [17].

Sentiment analysis and opinion mining is the field of computational study of people's opinion expressed in written language or text. Sentiment analysis brings together various research areas such as natural language processing, data mining and text mining. The input of the problem is a collection of written reviews about an object. Sentiment analysis for reviews involves processing of a text document using Natural Language Processing (NLP) techniques that extract only the desirable portion through various machine learning algorithms [1]. Common steps of NLP applied over a document involve tokenization, parts of speech, lemmatization, stop word elimination and vectorization [10, 12 and 13].

Presently a number of machine learning techniques are available for sentiment analysis of reviews [1]. First is lexicon-based approach [15] that includes dictionary, ensemble and corpus-based techniques. Second approach involves machine learning based sentiment analysis with well-known classification algorithms, that is Neural Network (NN), Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM) applied to textual data [16, 9]. Lastly, hybrid approach involves lexicon and machine learning techniques together to provide powerful means of accomplishing sentiment analysis [8, 9 and 11].

In this paper we have examined different papers on movie review analysis, where different machine learning classifiers are used for analysing user reviews over different applications. The main drawback with these classifiers is that they work only for unigram features i.e. they have two-class problem, without considering multiple independent variables with similar meaning and most of the classifiers failed in identifying and replacing misspelled words for classification. As a result of this, the performance parameters such as precision, recall and F-measure and prediction accuracy of these techniques are major issues to be tackled. Our research work aims to address these issues. To address two-class problem in the existing LR classifier, that is the classifier fails when it compares and classifies the reviews with multiple independent variables or this classifier fails when classification is done based on the words which have similar meaning and the existing classifier fails in replacing misspelled words in the sentence. To address this we propose ILR classification which divides the input dataset and classifies the reviews by correlating the variable based on the number of occurrences of a POS tagging, bag-of-words and stop words. The proposed ILR classification technique has different stages like pre-processing, POSTagging, Feature Extraction and classification of reviews by considering multiple independent words with similar meaning.

A case study on web based movie ticket booking is considered in our research work as a real life illustration that incorporates sentiment analysis to look for movie review polarity before the user books a movie. Users can look through their movies of interest, analyse the reviews posted by other users on websites or social media by checking out the ratings, cast, genre, and compare the price of watching the same movie in theatres as well as on online platforms [12].

The main contribution of the proposed work is:

- Identify/Identifies and Replace/replaces the misspelled words by using POS tagging method,
- Support count estimation using feature extraction technique and
- ILR classification of input reviews.

The rest of this article is organized as follows. In section 2 we discuss literature review. Section 3 covers proposed methodology, results and discussion is dealt in section 4 and section 5 consists of conclusion and future work.

II. PRELIMINARIES

The two classification techniques are mainly considered as preliminaries for carrying out the research work are Naive Bayes classification and logistic regression techniques. These techniques work as follows:

- 1) *NB classification algorithm* is based on Bayes probability rule and is used to compute the probability of an event's occurrence under given conditions [2, 10]. The advantages of NB technique are that it is relatively simple and efficient in classification accuracy. Equation 1 represents the Bayes rule producing output $P(C_k/T)$, which represents the probability of textual document T belongs to the class C_k , where $T = \{t_1, t_2, t_3, \dots, t_n\}$ is the feature vector of the text document and $C = \{c_1, c_2, \dots, c_k, \dots, c_n\}$ are the output classes for each k items.

$$P(C_k/T) = [P(T/C_k) * P(C_k)] / [P(T)] \dots \dots \dots (1)$$

The NB classification produces the maximum posterior probability represented as in the equation 2. The document $t_i \in T$ belonging to class C_k , where argmax_y denotes the value of the class mathematically represented by equation 2,

$$y = (\text{argmax}_y P(C_k) \pi^n P_i(t_i|C_k)) \dots \dots \dots (2)$$

- 2) *LR is a linear probability based classifier* that has an additional sigmoid function that represents the input data with a threshold parameter for decision variable [9]. The threshold is applied initially to the regression output in order to restrict the output to the value range [0, 1]. This constitutes the sigmoid function (σ), represented by equation 3,

$$\sigma(z) = \frac{1}{1 + e^{-z}} \dots \dots \dots (3)$$

Where e is base of natural log and e^{-z} is input to the function of sigmoid. It is a regression model that is mainly used for classifying a sample input to its class. The main drawback of the LR classifier is its failure while comparing and classifying the reviews with two independent variables can be referred as two class problem..

III. LITERATURE REVIEW

This section presents various research works related to the classification of reviews in different web based applications. It also provides a comprehensive analysis on various classification techniques and their limitations.

K. L. S Kumar et. al [3] presented the sentiment analysis of end user reviews from Amazon application and classified the output polarity in terms of positive as +1, negative as -1 and 0 for neutral review. They used NB, LR, and SentiWord Net algorithms for evaluating the classification accuracy against different set of movie reviews. The classifiers are trained using sample review data containing each individual polarity class. The dataset is in the form of TSV (Tab Separated Values) files. The NB classification was reported to be better than the other multiple classifiers, where 65% of the classification accuracy is achieved.

allen Rain et. al [8] presented a comprehensive review classification on Amazon's e-commerce site involving a number of different products ranging from books, tablet computers, CDs, and soon.

The website provides their users a scale of 1-5 to rate the product and also post a textual review about it. The approach used for classification makes use of bag-of-words features in order to distinctly represent each review of individual product. The author has extensively worked on finding out the intricate details in review that can serve as features to distinguish the polarity. The adjectives and collocations are also be considered to judge the review as negative or positive.

Sari Widya Sihwi et.al [4] proposed an approach for analysing the sentiments in movie reviews found on Twitter. The work highlights the common drawback of existing classification algorithms for sentiment analysis i.e. as the feature vector size increases; the accuracy of review categorization reduces. The authors have considered the NB algorithm along with information gain as feature selection technique to optimize the accuracy by choosing the important distinct features for review polarity judgment. The data collected using text crawler API is pre-processed to include only the words that exhibit the sentiments expressed by the user. The evaluation of the classifier made it clear that by adjusting the threshold value, the classifier performance at polarity prediction can be optimized.

Marium Nafees et.al [5] has carried out sentiment analysis on the product reviews expressed on Twitter and their polarity prediction using different algorithms. The data collected from Twitter consisting of five products are pre-processed using WEKA Tool. The classification of reviews in the form of tweets was performed using NB, LR, and SVM algorithms through comparisons. The SVM classifier outperforms the other two.

N. Banik et.al [6] proposed a methodology for movie review classification using sentiment analysis over text-based reviews of Bangla movies. The classification is based on NB classifier as well as linear SVM with unigram features used for testing and training. The reviews are pre-processed with the elimination of noise, hash tags, punctuation etc. The processing steps include tokenization, stemming and vectorization. A numerical feature vector for every token after vectorization is obtained. The work evaluates the performance of classification precision of both the classifier and reports that the SVM produces more accurate results than the NB classifier.

Peiman Barnaghi et.al [7] have focused on the dataset consisting of tweets on major hash tags related to FIFA World Cup 2014. The review polarity classification was implemented by LR and NB algorithms. It selected features involving unigram, n-grams and external lexical units. Term Frequency-Inverse Document Frequency (TF-IDF) is used as a part of data pre-processing. The effect on polarity of tweets of the tournament results are evaluated with regard to the user sentiment subject to incidents which happened during the sports.

Chantal Fry et.al [9] proposed clustering approach for Samsung galaxy smart phone product reviews obtained from Amazon e-commerce sites. The methodology involved data collection from Amazon via downloading the product reviews by means of a script. The pre-processing was done on the review set with elimination of hash tags, URLs, stop words and stemming. The clustering was employed using K-means and Peak-searching clustering techniques. The K-means algorithm performance was better than Peak-Searching clustering.

Table I represents comparative study of existing works considering their methodology, advantages, drawbacks and the classification accuracy.

Table I: Comprehensive analysis of existing review based classification techniques

Sl.No	Authors	Paper Title	Methodology	Advantages	Drawbacks & Future Work	Accuracy of Existing works
1	Farkhund Iqbal et.al [1]	Opinion Mining and Sentiment Analysis on Online Customer	Naïve Bayes, Logistic Regression, SentiWordNet	Naïve Bayes classifier proved most efficient	Dataset restricted to product reviews from only one	65%

		Review	classification algorithm with lexical features	Classifier among all three with good precision value on tested on multiple devices.	website Only Textual reviews with no mention of emoticons	
2	Sari Widya Sahwiet.al[4]	Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier	Naïve Bayes with information gain feature selection algorithm	High runtime efficiency with more efficiency.	The neutral review classification accuracy is still improvable	90% training accuracy
3	Marium Nafee set.al[5]	Sentiment Analysis of Polarity in Product Reviews In Social Media	Naïve Bayes SVM, Logistic Regression with text and emoticon review features	Easy classification and visualization using WEKA tool	Large number of features Accuracy of classification improvable	76%
4	N.Baniket.al[6]	Evaluation of Naïve Bayes and Support Vector Machines on Bangla Textual Movie Reviews	Naïve Bayes and Linear SVM with unigram features	Work on unigram features explored Bangla movie reviews Good precision	Only unigram features for small dataset Scope for more semantic details	74%
5	Pieman Barnaghi et.al[7]	Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment	Bayesian Logistic Regression, Naïve Bayes with 3 features-unigrams, n-grams and external lexicons	This kind of sentiment analysis helps sustain Twitter data forecasting patterns based on opinionated texts.	Tested using unigrams and bigram only	72%

6	CallenRain et.al[8]	SentimentAnalysis inAmazonReviewsUsingProbabilisticMachineLearning	Naïve Bayes,decision listclassifierwithasetoffeatures–bag of words,adjectives,collocations,etc.combined	Arichandgood numberofsemantic features	Limits on number of features andrules applied	68%
7	ChantalFryet.al[9]	Can we GroupSimilar AmazonReviews:ACaseStudy withDifferentClusteringAlgorithms	K-means andPeak-SearchingClustering withTF-IDF featurevector	Evaluationusinghumanassessmentand puritymetric forclusteringbothimplemented	Noautomationof topic labelingthroughleveragingexistingsemantic analysis.	66%

In this paper we have examined different papers on movie review analysis, where different machine learning classifiers are used for analysing user reviews over different applications. The main drawback with these classifiers is that they work only for unigram features i.e. they have two-class problem, without considering multiple independent variables with similar meaning and most of the classifiers failed in identifying and replacing misspelled words for classification. As a result of this, the performance parameters such as precision, recall and F-measure and prediction accuracy of these techniques are major issues to be tackled. Our research work aims to address these issues.

IV. PROPOSED METHODOLOGY

This section discusses the proposed technique of Improved logistic regression that identifies and replaces the misspelled word by using POS tagging method, support count estimation and classification of input reviews.

1) ILR Workflow model

The system architecture diagram depicted in figure 1 describes the workflow model of how the ILR technique works on movie dataset considered from the standard movie based application and then applied with data pre-processing on the data set considered, feature selection of the attributes from the review and then classifying them based on the proposed ILR algorithm.

The first step in analysing the movie reviews is to construct the dataset for the model. The dataset considered is from standard website “<http://www.ai.stanford.edu/~amaas/data/sentiment>” [22]. The dataset contains 50,000 reviews from IMDB database for 1850 different English movies and divided into 25,000 training set and 25,000 test set. Because some of the movies receive substantially more reviews than others, the dataset is limited for including at most 30 reviews from any movie in the collection. The attributes considered for the creation of the dataset are various features of the reviews like rating, number of reviews per movie and then stored in a text form as training set and test dataset and then applied the proposed technique for classification of positive and negative based reviews. Later this dataset can be used for classification and prediction of movie reviews.

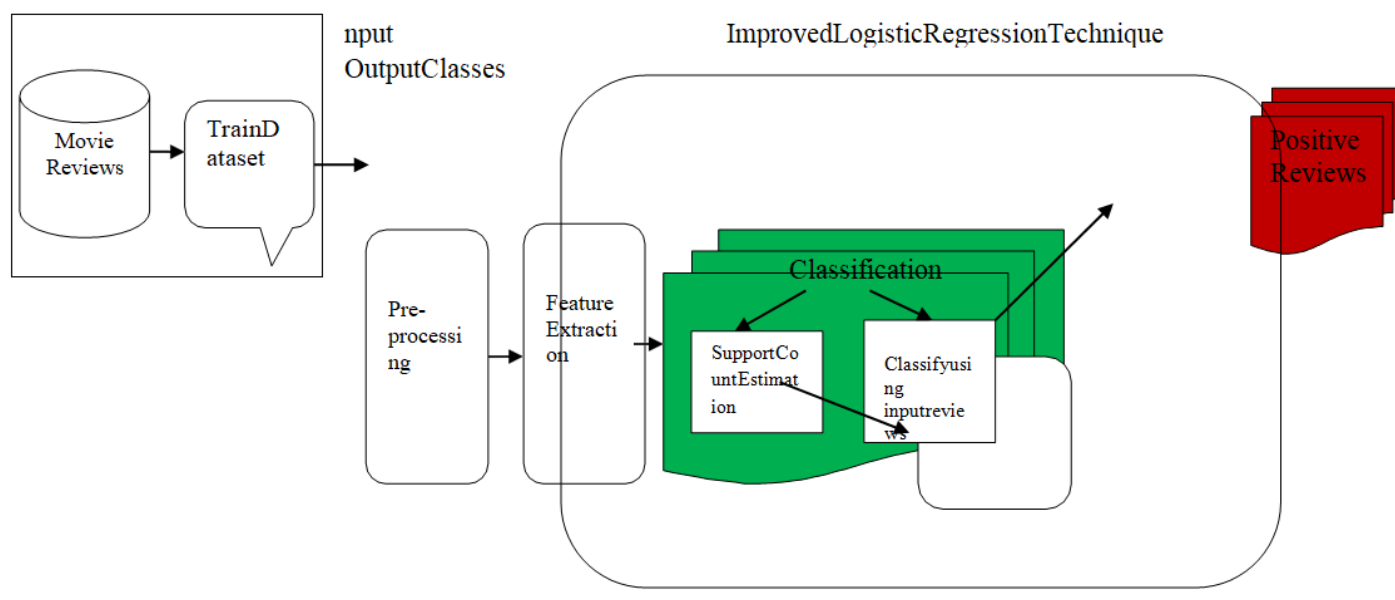


Figure 1: ILR workflow model

2) ILR Stages

This proposed technique carried out in different stages like data preprocessing, feature extraction and classification which are explained as follows.

a) Data Pre-processing

The most important and computational part of the analysis is pre-processing of the input data, which is done as follows:

- *Tokenization of words:* This is mainly used to identify all the noun words in given input reviews. These words are then referred as tokens or the units for the given input.
- *Removal of stop words:* This is the important process of preprocessing which is mainly used to eliminate frequently occurring words such as nouns, prepositions, articles and adverbs. These words depend on the language used for reviews.
- *Stemming of the tokens:* This is used for the standardization of the tokens into the text, in which different variants of tokens are reduced as common term (called stem). For grammatical reasons, documents or texts use different forms of a word, such as 'stems', 'stemmer', 'stemming', 'stemmed' where the root word is 'stem'.
- *POS Tagging:* This is the final step of preprocessing the input, which identifies the misspelled words in the sentence to provide a proper representation of given input dataset. This can be implemented in following ways.
- Words like nouns and pronouns usually do not contain any sentiment. It is able to filter out such words with the help of a POS tagger;
- A POS tagger can also be used to distinguish words that can be used in different parts of speech. For instance, as a verb, "enhanced" may conduct different amount of sentiment as being of an adjective.
- POS Tagging has been integrated with dictionary to identify and replace the misspelled words in the sentence that helps in achieving good classification accuracy.

b) Feature Extraction

Feature Extraction is the process of extracting relevant features. In the existing research on sentiment analysis considered as all speech words are features. The proposed model retrieves three different parts of words as features. The verbs, adverbs and adjectives play an important role in opinions. The WordNet dictionary is used to perform tagging and extracts all the Verbs (V), Adverbs (A), Adjectives (AJ) and their combinations Adverbs + Adjectives (AAJ), Adverbs + Verbs (AV), Adverbs + Adjectives + Verbs (AAJV) and Adjectives + Verbs (AJV) as sentiment features of movie application then these features are used for classifying the user reviews.

c) Classification

Once the features are extracted, the classification of the movie reviews is done using ILR algorithm. The classification technique is implemented by combining both joint distribution and the input to output mapping techniques. Which means the selected feature for classifying the review will be compared with similar words as well as the word with similar meaning. This is done by using the integration of POS tagging which will be classified as a similar group of review. This will be carried out using different steps which is described as follows:

3) Support count for splitting the input dataset

Support count is the value for splitting the input dataset which will be determined based on the size and number of reviews used in the training dataset. Before selecting features like target variable for the classification, we need to set the support count for splitting the input dataset. In this work, the support count is set based on the number of reviews considered for analysis and splitting the input dataset, we can process the data faster or we can do parallel processing.

$$vect = \text{CountVectorizer}(\text{min_df} = \text{count_value}) \dots \dots \dots (4)$$

The equation 4 specifies *vect* variable which takes count of vectorizer that can be referred as a simple way to tokenize a collection of text documents and build the vocabulary of known words. *min_df* defines the support count value for the input dataset which is considered for classification.

4) Classifying based on input reviews

This module describes the unlabeled input dataset that is taken for analysis and will be classified based on the type of reviews. Here POS module is integrated for classifying the reviews based on multiple independent variables with similar meaning which can be classified as a similar group described in equation 5. Here *ngram_range* describes the lower and upper boundary of the range of 2-values for different *n*-gram to be extracted. In the proposed technique we have considered (2,2) as upper and lower bound as a cutoff, because the proposed technique works for bigram features.

$$ngram_range = (a, b) \dots \dots \dots (5)$$

The ILR is also based on a bilinear equation module with multiple independent input parameters as in linear regression to predict the probability of the input belonging to a specific class. A possible output that represents a class. Using bilinear function, the output range can vary from less than 1 to values over 0. The Improved logistic function can be expressed as in equation 6,

$$\sigma(z) = \frac{1}{1 + e^{-z}} \dots (P(X|Y_b) * P(Y_b)) / [P(X)] \dots \dots \dots (6)$$

Equation 6 represents the rule producing output $P(x / y)$, the probability of textual document X belonging to the class Y , where $X = \{x_1, x_2, x_3, \dots, x_n\}$ is the feature vector of the text document and $Y = \{y_1, y_2, \dots, y_k, \dots, y_n\}$ is the output class for each b items. It is combined with existing LR classifier that has an additional sigmoid function (e^z) representing the input data with a threshold parameter for decision variable.

The working of ILR based classification model is described below considering an example of user review for a particular movie. User review is "the movie was good, but the cinematography was too worst music was horrible, comedy was better and music was too good, overall the movie is once watchable"

This review is classified using ILR through following steps:

- Step 1: Apply pre-processing steps discussed in 3.2.1 section that results in removal of frequently occurring words like 'the', 'was', 'is' etc, the misspelled word *movi* is replaced by the correct word *movie* after applying POS tagging technique 18 words out of 27 words will be retrieved. Output after applying pre-processing: *movie good but cinematography too worst music horrible, comedy better and music too good, overall movie once watchable*
- Step 2: Apply feature extraction process that groups the combinations Adverbs + Adjectives (AAJ), Adverbs + Verbs (AV), Adverbs + Adjectives + Verbs (AAJV) and Adjectives + Verbs (AJV) as sentiment features of movie based applications.

- Step 3: Apply the support count for the input review. By referring equation (4), we have considered support count value as 5 for parallel processing of reviews. Then 5 words out of 18 words are separated into four different groups for parallel processing.
- Step 4: Next, multiple independent words with same meaning are processed at a time. Considering the value of a and b as 2 in equation (5), the review “good”, “too good” and better are treated similar words during classification for the input review considered, hence total words during classification will become 15 out of 18. Output after applying pre-processing: movie good but cinematography too worst music horrible, comedy better and music too good, overall movie once watchable
- Step 5: Equation (6) is considered to classify the negative and positive set of reviews based on the prediction attributes of the dataset. If we apply this to the input review, the probability of positive occurrence of positive words is $1/2$ and the probability of negative occurrence of words is $3/15$. Hence the given review is classified as positive because of more positive words in the review. By this we can achieve around 85% classification accuracy.
In the proposed work we have considered 25000 movie reviews, where we have achieved 88% classification accuracy, through the proposed technique we can able achieve good prediction accuracy when we train the dataset with more number of input reviews.
- Step 6: Plot the graph against the classification accuracy, time taken for classification, precision, recall and F-measure of proposed ILR and compare with existing LR and NB classifiers.

V. EXPERIMENTAL RESULTS

The implementation of proposed work is carried out using anaconda 4.3.8, python 3.6.3 and the open source libraries suitable for analyzing the movie reviews. Matplotlib toolkit is used for drawing the results. The below Table 2 provides the parameters considered for the implementation of the proposed work.

Table 2: Implementation parameters

Dataset	Movie Dataset
Source:	http://ai.stanford.edu/~amaas/data/sentiment
Total Number of Reviews	50000
Number of reviews considered for training:	25000
Number of reviews considered for testing	25000
Number of maximum reviews considered for a single movie	30
Total number of movies considered	850
Technology used	Python 3.6.3

The performance of proposed ILR is compared with existing logistic regression and naïve bayes classifiers for different set of reviews against various performance parameters like classification accuracy, time taken for classification, precision, recall and F-measure.

A. Classification Accuracy:

The Figure 2 describes the accuracy of classification for movie based reviews, where x-axis represents different set of test reviews considered and y-axis represents the classification accuracy. Though the proposed ILR an average of 88% classification accuracy has been achieved, which is 15% more when compared with existing LR and NB classifiers.

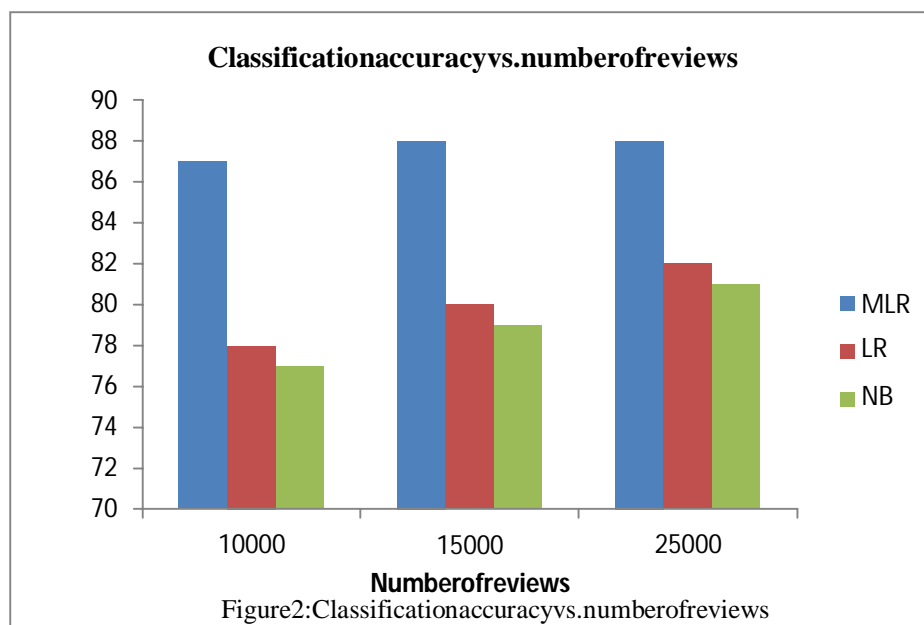


Figure 2: Classification accuracy vs. number of reviews

B. Time-taken for Classification:

The Figure 3 describes the time taken to classify the various instance of test reviews, where x-axis represents the time taken to classify various instance of reviews using proposed ILR technique, existing LR and Naïve Bayes classifiers against the various instance of reviews and proves the proposed ILR is taking less time for classification because of parallel processing when compared to existing techniques even after varying the size of the dataset with different number of reviews.

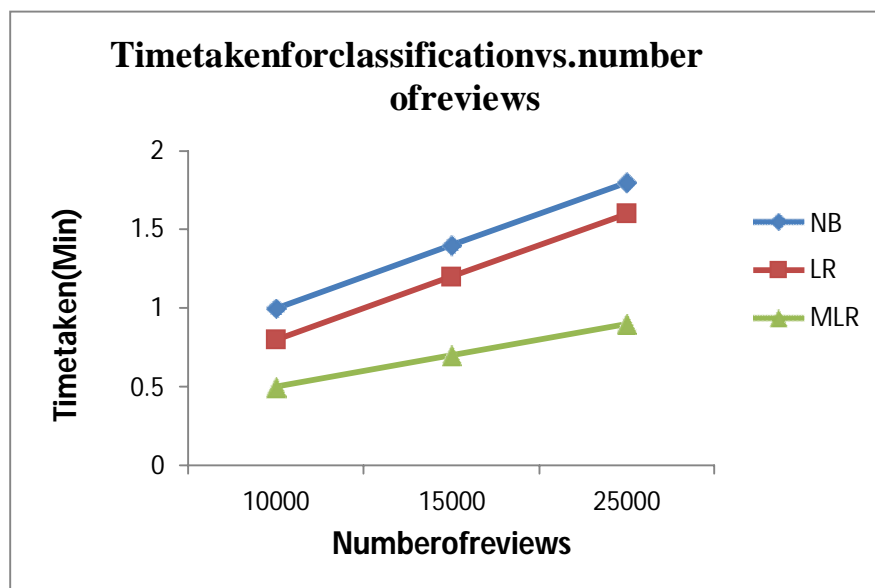


Figure 3: Time taken for the classification of input reviews vs. number of reviews

C. Precision

It is defined as the ratio of correctly classified over number of all classifications which can be expressed as:

Precision = $\frac{\text{correctly classified}}{\text{correctly classified} + \text{Errorly classified}}$

The below Figure 4 describes the accuracy of precision value in percentage against proposed ILR, existing LR and NB classifiers and proves the proposed ILR is having more precision value because of less number of errorly classified words when compared with existing technique.

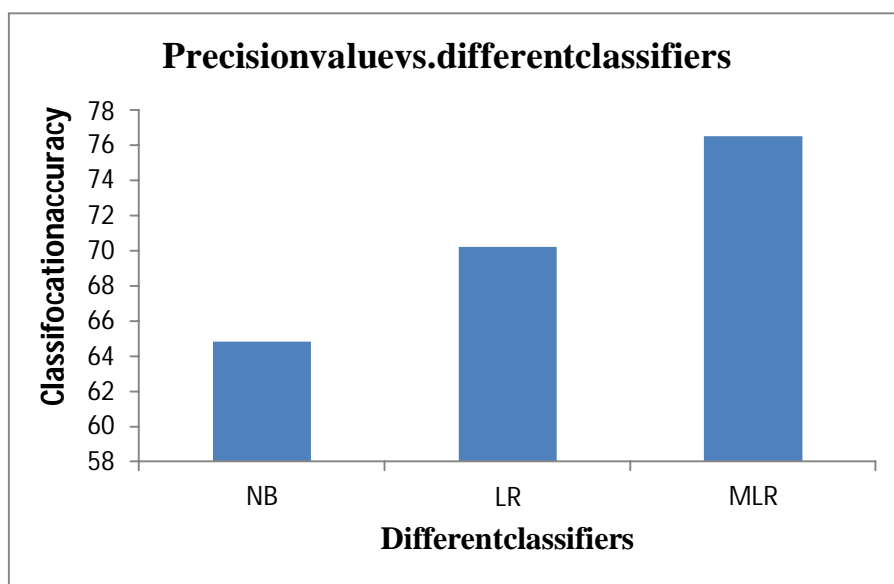


Figure4:Precisionvaluevs. DifferentClassifiers

D. Recall

It is considered to determine the number of true positive function which can be expressed as:

$\text{Recall} = \frac{\text{correctly classified}}{\text{correctly classified} + \text{Missed classified}}$

The below Figure 5 describes the accuracy of recall value in percentage against proposed ILR, existing LR and NB classifiers and proves the proposed ILR is having more recall value because of less number of missclassified words when compared with existing techniques.

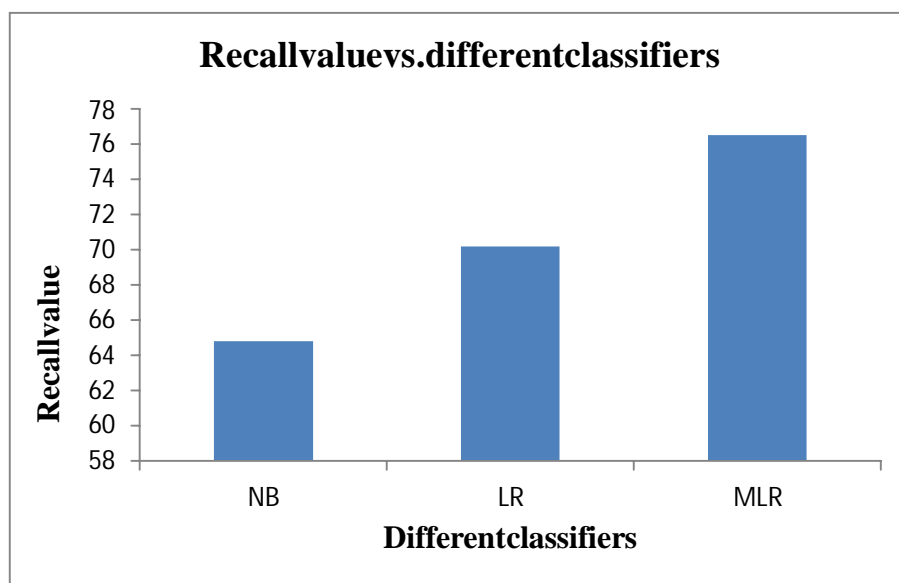


Figure5:Recallvaluevs.DifferentClassifiers

E. F-Measure

It is a combined measure for precision and recall values which can be expressed as:

$\text{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$

The below Figure 6 describes the accuracy of F-measure value in percentage against proposed ILR, existing LR and NB classifiers and proves the proposed ILR is having more F-measure value because of more precision and recall values when compared with existing techniques.

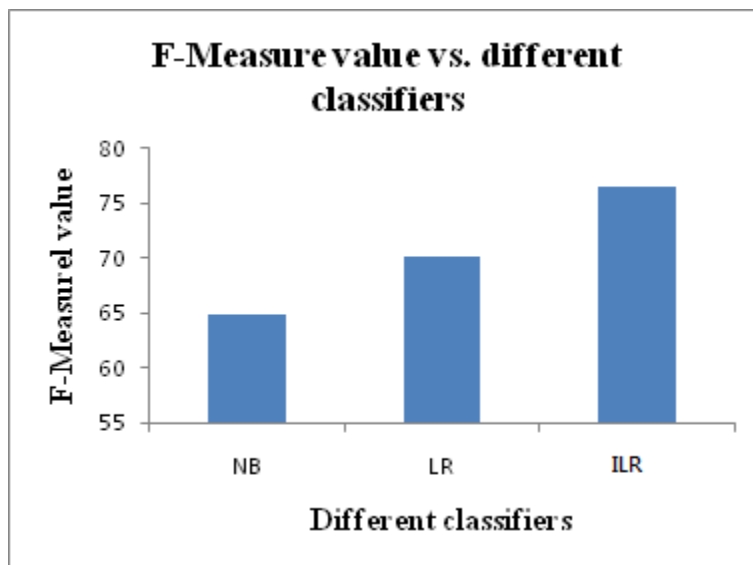


Figure6:F-measurevaluevs.Different Classifiers

VI. CONCLUSION AND FUTURE WORK

The analysis and classification of various movie based reviews is taken from different moviebased applications. Different classifiers are used to classify the reviews on the movies like Naive Bayes, Logistic Regression, Support Vector Machine etc.,. The existing classifiers fail in achieving the desired accuracy, because the classifiers do not work properly with multiple independent variables i.e. word with similar meaning is treated as separate for the classification that affects the performance parameters. While classification, the proposed work addressed the two-class problem which is the main drawback in the existing LR classifier. With the proposed classifier achieved an average classification accuracy of 88% by varying the size of the reviews. The proposed classifier accuracy has been evaluated with different evaluation parameters and achieved better performance. In future, this work can be extended on mining the reviews from multiple applications such as Bookmyshow, Paytm etc. Further improved machine learning algorithms can be incorporated to improve the efficiency, which will help in deciding the best classification classifier in sentiment analysis.

BIBLIOGRAPHY

- [1] Farkhund Iqbal, Jahanzeb Maqbool, Benjamin C. M. Fung, Rabia Batool, Asad Masood Khattak, Saiqa Aleem, Patrick C. K. Hunga, "A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction", IEEE, vol. 7, pp. 14637-14652, 2019.
- [2] Tu Nguyen Thi Ngoc, Ha Nguyen Thi Thu, Viet Anh Nguyen, "Mining aspects of customer's review on the social network", Journal of Big Data, vol. 6, Springer, Number 1, pp. 6-22, Article number: 22, 2019.
- [3] K. L. S. Kumar, J. Desai and J. Majumdar, "Opinion mining and sentiment analysis on online customer review," IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-4, 2016.
- [4] Sari Widya Sihwi, Insan Prasetya Jati, Rini Anggrainingsih, "Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier", IEEE International Conference on Application for Technology of Information and Communication (iSemantic), pp. 190-195, 2018.
- [5] Mariam Nafees, Hafsa Dar, Ikram Ullah Lali, Salman Tiwana, "Sentiment Analysis of Polarity in Product Reviews in Social Media", 14th International Conference on Emerging Technologies (ICET), pp. 1-6, 2018.
- [6] N. Banik and M. Hasan Hafizur Rahman, "Evaluation of Naïve Bayes and Support Vector Machines on Bangla Textual Movie Reviews," International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, pp. 1-6, 2018.
- [7] Peiman Barnaghi, John G. Breslin, Parsa Ghaffari, "Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment", Oxford, Second International Conference on Big Data Computing Service and Applications, pp. 52-57, 2016.
- [8] Wang, Yequan, Aixin Sun, Jialong Han, Ying Liu, and Xiaoyan Zhu. "Sentiment analysis by capsules." In Proceedings of the 2018 world wide web conference, pp. 1165-1174, 2018.
- [9] Chantal Fry, Sukanya Manna, "Can we Group Similar Amazon Reviews: A Case Study with Different Clustering Algorithms", Tenth International Conference on Semantic Computing, pp. 374-377, 2016.
- [10] Asha S Manek, P Deepa Shenoy, M Chandra Mohan, Venugopal K R, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier", World Wide Web, vol. 20, Springer, Number 2, pp. 135-154, 2017.
- [11] Haiyun Peng, Erik Cambria, Amir Hussain, "A Review of Sentiment Analysis Research in Chinese Language", Cognitive Computation, vol. 9, Springer, Number 4, pp. 423-435, 2017.

- [12] J. Zheng and L. Zheng, "A Dictionary-Based Convolution Recurrent Neural Network Model for Sentiment Analysis", 2019 International Conference on Communications, Information System and Computer Engineering (CISCE), Haikou, China, pp. 606-611, 2019
- [13] N. Mtetwa, A.O. Awukam and M. Yousefi, "Feature Extraction and Classification of Movie Reviews", 5th International Conference on Soft Computing & Machine Intelligence (ISCMI), Nairobi, Kenya, pp. 67-71, 2018
- [14] S. Rajalakshmi, S. Asha, N. Pazhaniraja, "A Comprehensive Survey on Sentiment Analysis", 4th International Conference on Signal Processing, Communications and Networking (ICSCN -2017), pp. 1-5, 2017.
- [15] Harpreet Kaur, Veenu Mangat, Nidhi, "A survey of sentiment analysis techniques", International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), , pp. 921-925, 2017.
- [16] Vikas K Vijayan, K. R. Bindu, Latha Parameswaran, "A comprehensive study of text classification algorithms", IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), , pp. 1109-1113, 2017.
- [17] X. Lei, X. Qian and G. Zhao, "Rating Prediction Based on Social Sentiment From Textual Reviews", in IEEE Transactions on Multimedia, vol. 18, Number 9, pp. 1910-1921, Sept. 2016.
- [18] Parkhe V. & Biswas B. "Sentiment analysis of movie reviews: finding most important movie aspects using driving factors", Soft Computing, vol. 20, Springer, pp. 3373-3379, 2016.
- [19] Ketan Sarvakar, Urvashi K Kuchara, "Sentiment Analysis of movie reviews: A new feature-based sentiment classification", International Journal of Scientific Research in Computer Science and Engineering, vol. 6, Issue. 3, pp. 8-12, 2018.
- [20] Doaa Mohey El-Din Mohamed Hussein, "A survey on sentiment analysis challenges", Journal of King Saud University-Engineering Sciences, vol. 30, Elsevier, pp. 330-338, 2018
- [21] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, vol. 5 Elsevier, Issue 4, pp. 1093-1113, 2018 <http://ai.stanford.edu/~amaas/data/sentiment-Dataset> considered for classification.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)