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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.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 commonapplications of sentiment analysis. It involves detection of sentiments on the reviews made bythe users on social networking applications through opinion mining.In general,reviews canhave positive, negative or neutral polarity indicators. For classification, the polarity indicatorstake the form of certain words and emotions that readily show the user's sentiments. Existingworks fall short of producing accurate classification results because of two-class problem thataffects the performance of evaluation parameters like precision, recall, accuracy and F-measure.Hencethereisaneedofanefficientclassificationtechniquewhichaddressestwo-classproblem. Thiswork proposes ImprovedversionofLogisticRegression[ILR]thatiscommonly used for sentiment analysis and classification. The proposed classification techniqueidentifies and replaces the misspelled words in the sentence,supportcountestimation andclassificationofreviewsalongwithmultipleindependentwordswithsimilarmeaninginparallel. The experimental results show the classification accuracy of the proposed technique tobemoreaccuratecomparedtothe existinglogistic regressionandnaïvebayesclassifiers.

Keywords: SentimentalAnalysis,MachineLearning,ImprovedLogisticRegression,POSTaggingandMovieReviews.

I. INTRODUCTION

Data mining is a process of discovering specific patterns in huge data sets. It aims to convert thegathered data from a dataset into a comprehensible form for optimal usage. Web mining is anapplication of data mining strategies to find interesting patterns in the data which is downloadedfrom the web. Opinion mining is a sub-discipline of web mining that facilitates searching anddiscoveringuser'sopinionaboutaspecifictopicora product[17].

Sentiment analysis and opinion mining is the field of computational study of people's opinionexpressed in written language or text. Sentiment analysis brings together various research areassuch as natural language processing, data mining and text mining. The input of the problem is acollectionofwrittenreviewsaboutanobject.Sentimentanalysisforreviewsinvolvesprocessingof atextdocumentusingNaturalLanguageProcessing(NLP)techniquesthatextract only the desirable portion through various machine learning algorithms [1]. Common steps ofNLP applied over a document involve tokenization, parts of speech, lemmatization, stop wordeliminationandvectorization[10,12and13].

Presently a number of machine learning techniques are available for sentiment analysisofreviews [1]. First is lexicon-based approach [15] that includes dictionary, ensemble and corpusbased techniques. Second approach involves machine learning based sentiment analysis withwell-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 powerfulmeansof accomplishingsentimentanalysis[8,9and11].

In this paper we have examined different papers on movie review analysis, where differentmachine learning classifiers are used for analysing user reviews over different applications. Themain drawback with these classifiers is that they work only for unigram features i.e. they havetwo-class problem,without considering multiple independent variables with similar meaningandmostoftheclassifiersfailedinidentifyingandreplacingmisspelledwordsforclassification. As a result of this, the performance parameters such as precision, recall and F-measureandprediction accuracyofthese techniquesare majorissuestobetackled. Ourresearchworkaimstoaddresstheseissues.To address two-class problem in the existing LR classifier, that is the classifier fails when itcompares and classify the reviews with multiple independent variables or this classifier failswhen classification is done based on the words which have similar meaning and the existingclassifier fails in replacing misspelled words in the sentence. To address this we propose ILRclassificationwhichdividesthe inputdatasetandclassifiesthe reviewsby correlating thevariable based on the number of occurrences of a POS tagging, bag-of-words and stop words.TheproposedILRclassificationtechniquehasdifferentstageslikepre-processing,POSTagging,Feature Extraction and classification of reviews by considering multiple independentwordswithsimilarmeaning.

A case study on web based movie ticket booking is considered in our research work as a real lifeillustration that incorporates sentiment analysis to look for movie review polarity before the userbooks a movie. Users can look through their movies of interest, analyse the reviews posted byother users on websites or social media by checking out the ratings, cast, genre, and compare thepriceofwatchingthesame movie intheatreaswellasonlineplatforms[12].

The maincontributionof theproposedworkis:

- Identify/IdentifiesandReplace/replacesthemisspelledwordsbyusingPOStaggingmethod,
- Supportcountestimationusingfeatureextractiontechniqueand
- ILRclassificationofinputreviews.

The rest of this article is organized as follows. In section 2 we discuss literature review. Section3 covers proposed methodology, results and discussion is dealt in section 4 and section 5consistsof conclusionandfuture work.

II. PRELIMINARIES

The two classification techniques are mainly considered as preliminaries for carrying out theresearchworkareNaivebayesclassificationandlogisticregressiontechniques.Thesetechniquesworkasfollows:

1) *NB classification algorithm* is based on bayes probability rule and is used to compute theprobability of an event's occurrence under given conditions [2, 10]. The advantages of NBstechnique are that it is relatively simple and efficient in classification accuracy. Equation 1represents the Bayes rule producing output *P(Ck|T),* which represents the probability of textualdocument *T* belongs to the class *Ck,* where *T={t1, t2, t3,…tn}* is the feature vector of the textdocumentand*C={c1,c2,…,ck,…cn}*aretheoutputclassesforeach*k* items.

$$P(C_k|T)=[P(\,T|C_k)*P(C_k)\,]/[P(T)]\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (1)$$

The NB classification produces the maximum posteriorprobability represented as*y* in theequation 2. The document *ti€T* belonging to class *Ck*, where *argmax* denotes the value of theclassismathematicallyrepresentedbyequation2,

$$y=(argmax_y P(C_k)\pi^n P_{i=1}(t_i|C_k))\ldots\ldots..(2)$$

2) LRisalinearprobability based classifierthathas an additional sigmoid function thatrepresents the input data with a threshold parameter for decision variable [9]. The threshold isapplied initially to the regression output in order to restrict the output to the value range [0, 1].Thisconstitutesthesigmoidfunction*(σ)*,representedbyequation3,

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

Where *e* isbase of natural log and e^{-z}isinputto the function of sigmoid.Itis a regressionmodelthatismainlyusedforclassifyingasampleinputtoitsclass.Themaindrawbackofthe LR classifier is its failure while comparing and classifying the reviews with two independentvariablescanbereferredastwoclassproblem..

III. LITERATURE REVIEW

This section presents various research works related to the classification of reviews in differentweb based applications. It also provides a comprehensive analysis on various classificationtechniquesandtheirlimitations.

K. L. S Kumar et. al [3] presented the sentiment analysis of end user reviews from Amazonapplication and classified the output polarity in terms of positive as +1, negative as -1 and 0 forneutralreview. TheyusedNB,LR ,andSentiWord Netalgorithmsforevaluatingtheclassification accuracy against different set of movie reviews. The classifiers are trained usingsample 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 othermultiple classifiers,where65%of theclassificationaccuracyisachieved.

allen Rain et.al [8] presented a comprehensive review classification on Amazon's e-commercesite 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 textualreview about it. The approach used forclassificationmakes use of bag-of-words features inorderto distinctly represent each review of individual product. The authorhas extensivelyworked on finding out the intricate details in review that can serve as features to distinguish thepolarity. The adjectives and collocations are also be considered to judge the review as negativeorpositive.

Sari Widya Sihwi et.al [4] proposed an approach for analysing the sentiments in movie reviewsfoundonTwitter. Theworkhighlightsthecommondrawbackofexistingclassificationalgorithms for sentiment analysis i.e. as the feature vector size increases; the accuracy of reviewcategorization reduces. The authors have considered the NB algorithm along with informationgain as feature selection technique to optimize the accuracy by choosing the important distinctfeaturesforreviewpolarityjudgment. ThedatacollectedusingtextcrawlerAPIispre-processed to include only the words that exhibit the sentiments expressed by the user. Theevaluation of the classifier made it clear that by adjusting the threshold value, the classifierperformanceatpolarity predictioncanbeoptimized.

MariumNafeeset.al[5]hascarriedoutsentimentanalysisontheproductreviewsexpressedon Twitter and their polarity prediction using different algorithms. The data collected from Twitterconsistingoffiveproductsarepre-processedusingWEKATool.Theclassificationof reviewsinthe form of tweets was performed using NB, LR, and SVM algorithms through comparisons.TheSVMclassifieroutperformsthe othertwo.

N. Banik et.al [6] proposed a methodology for movie review classification using sentimentanalysis over text-based reviews of Bangla movies. The classification is based on NB classifieras 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 includetokenization, stemming and vectorization. A numerical feature vector for every token aftervectorization is obtained. The work evaluates the performance of classification precision of boththeclassifierandreportsthattheSVM producesmoreaccurateresults thantheNBclassifier.

PeimanBarnaghi et.al [7] have focused on the dataset consisting of tweets on major hash tagsrelated to FIFA World Cup 2014.The review polarity classification was implemented by LR andNB algorithms. It selected features involving unigram, n-grams and external lexical units. TermFrequency–Inverse Document Frequency (TF-IDF) is used as a part of data pre-processing. Theeffect on polarity of tweets of the tournament results are evaluated with regard to the usersentimentssubjecttoincidentswhichhappenedduringthesports.

Chantal Fry et.al [9] proposed clustering approach for Samsung galaxy smart phone productreviews obtained from Amazon e-commerce sites. The methodology involved data collectionfrom Amazon via downloading the product reviews by means of a script. The pre-processingwas done on the review set with elimination of hash tags, URLs, stop words and stemming. Theclustering wasemployed using K-meansand Peak-searching clustering techniques. The K-meansalgorithmperformance wasbetterthanPeak-Searchingclustering.

Table1representscomparativestudyofexistingworksconsideringtheirmethodology,advantages,drawbacksandtheclassificationaccuracy .

Table1:Comprehensiveanalysisof existingreviewbasedclassificationtechniques

Sl.No.	Authors	PaperTitle	Methodology	Advantages	Drawbacks &FutureWork	AccuracyofExistingworks
1	FarkhundIqbalet.al[1]	OpinionMiningandSentimentAnalysis on Online Customer	Naïve Bayes,LogisticRegression, SentiWordNet	NaïveBayesclassifierprovedmost efficient	Datasetrestricted toproductreviews fromonlyone	65%

		Review	classificationalgorithm withlexiconfeatures	Classifieramong all threewithgoodprecisionvalue on tested onmultipledevices.	website OnlyTextualreviews with nomentionofemoticons	
2	SariWidyaSahwiet.al[4]	Twitter SentimentAnalysis ofMovieReviewsUsing InformationGainandNaïveBayesClassifier	Naïve Bayeswith informationgain featureselectionalgorithm	High runtime efficiencywith moreefficiency.	The neutralreviewclassificationaccuracy stillimprovable	90% trainingaccuracy
3	MariumNafeeset.al[5]	SentimentAnalysis of Polarity inProductReviewsInSocialMedia	NaïveBayesSVM,LogisticRegressionwithtext and emoticonreviewfeatures	Easyclassificationandvisualizationusing WEKAtool	Large numberoffeatures Accuracy ofclassificationimprovable	76%
4	N.Baniket.al[6]	EvaluationofNaiveBayesandSupportVector MachinesonBanglaTextualMovieReviews	Naïve BayesandLinear SVM with unigramfeatures	Work onun exploredBanglamoviereviews Goodprecision	Onlyunigramfeaturesforsmalldataset Scope for moresemanticdetails	74%
5	PiemanBarnaghiet.al[7]	OpinionMiningandSentimentPolarity on Twitter andCorrelationBetween Eventsand Sentiment	BayesianLogisticRegression,Naïve Bayeswith3features-unigrams, n-grams andexternallexicons	Thiskindofsentimentanalysishelpsustouse Twitter data forextractingpatternsbased onopinionatedtexts.	Tested using unigrams andbigramsonly	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 differentmachine learning classifiers are used for analysing user reviews over different applications. Themain drawback with these classifiers is that they work only for unigram features i.e. they havetwo-class problem, without consideringmultiple independent variables with similar meaningandmostoftheclassifiersfailedinidentifyingandreplacingmisspelledwordsforclassification. As a result of this, the performance parameters such as precision, recall and F-measureandpredictionaccuracyofthesetechniquesaremajorissuestobetackled.Ourresearchworkaimstoaddresstheseissues.

IV. PROPOSED METHODOLOGY

Thissection discussestheproposed techniqueofImprovedlogisticregressionthatidentifies and replaces themisspelled word by using POS taggingmethod, supportcountestimationandclassificationofinputreviews.

1) ILRWorkflowmodel

The system architecture diagram depicted in figure 1 describes the workflow model of how theILR technique works on movie dataset considered from the standard movie based applicationand then applied with data pre-processing on the data set considered, feature selection of theattributesfromthereviewandthenclassifyingthembasedontheproposedILRalgorithm.

Thefirststepinanalysingthemoviereviewsistoconstructthedatasetforthemodel.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 Englishmovies and divided into 25,000 training set and 25,000 test set. Because some of the moviesreceive substantially more reviews than others, the dataset is limited for including at most 30reviews from any movie in the collection. The attributes considered for the creation of thedataset are various features of the reviews like rating, number of reviews per movie and thenstored in a text form as training set and test dataset and then applied the proposed technique forclassificationofpositiveandnegativebasedreviews.Laterthisdatasetcanbeusedforclassificationandpredictionofmoviesreviews.

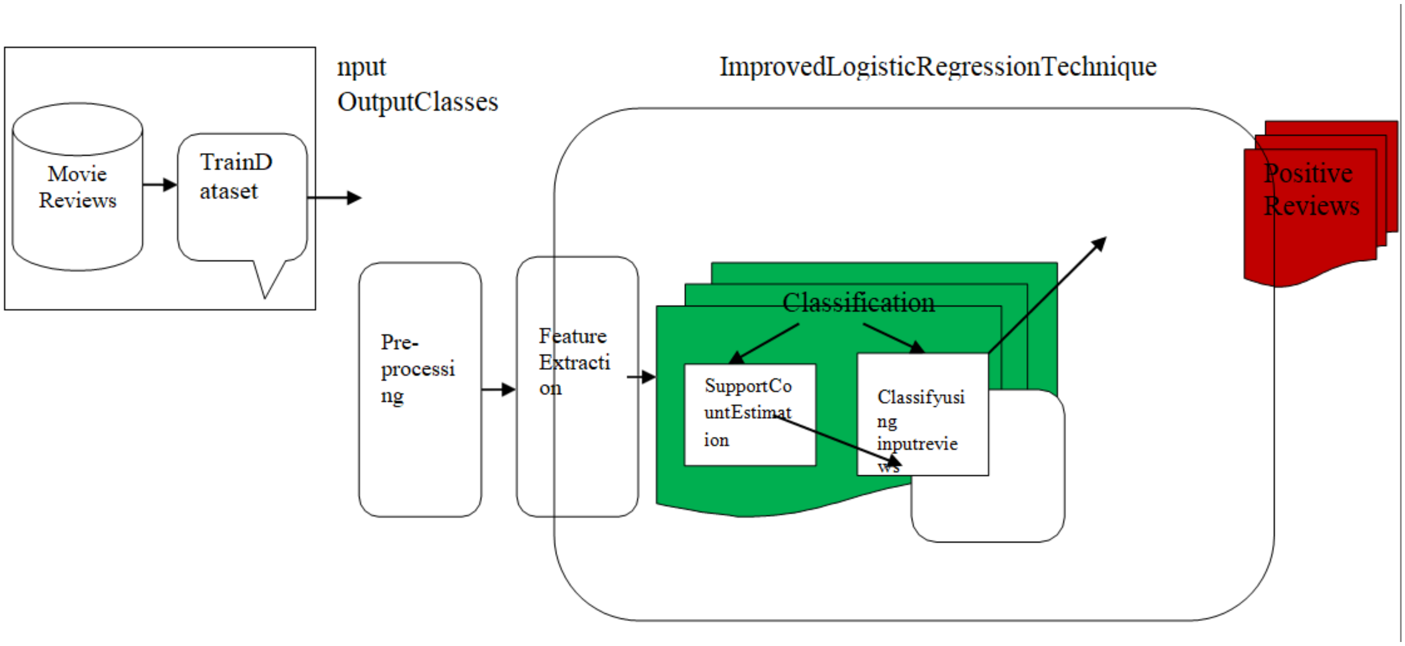

Figure1:ILRworkflowmodel

2) *ILRStages*

Thisproposedtechniquecarriedoutindifferentstageslikedatapreprocessing,featureextractionandclassificationwhichareexplainedasfollows.

a) *DataPre-processing*

Themostimportantandcomputationalpartoftheanalysisispre-processingoftheinputdata,whichisdoneasfollows:

- *Tokenization of words:* This is mainly used to identify the all the noun words in giveninputreviews.Thesewordsarethenreferredas tokenortheunitsforthegiveninput.
- *Removal of stop words:* This is the important process of preprocessing which is mainlyused to eliminate frequently occurring words such as nouns, prepositions, articles andadverbs.These wordsdependonthelanguageusedforreviews.
- *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). Forgrammatical reasons,documents ortexts uses different forms of aword,suchas'stems', 'stemmer','stemming','stemmed'wheretherootwordis'stem'.
- *POS Tagging:* This is the final step of preprocessing the input, which identifies themisspelled words in the sentence to provide a proper representation of given inputdataset.Thiscanbeimplementedinfollowingways.
- Words like nouns and pronouns usually do not contain any sentiment. It is able tofilteroutsuchwordswiththe helpof aPOStagger;
- A POS tagger can also be used to distinguish words that can be used in differentparts of speech. For instance, as a verb, "enhanced" may conduct different amountofsentimentasbeingofanadjective.
- POSTagginghasbeenintegratedwithdictionarytoidentifyandreplacethemisspelledwordsinthesentencethathelpsinachievinggoodclassificationaccuracy.

b) *FeatureExtraction*

Feature Extraction is the process of extracting relevant features. In the existing research onsentiment analysis considered as all speech words are features. The proposed model retrievesthree different parts of words as features. The verbs, adverbs and adjectives play an importantrole in opinions. The WorldNet dictionary is used to perform tagging and extracts all the Verbs(V),Adverbs (A),Adjectives(AJ)andtheircombinationsAdverbs +Adjectives (AAJ),Adverbs +Verbs(AV), Adverbs+ Adjectives+Verbs(AAJV)and Adjectives+Verbs(AJV)assentiment features of movie application then these features are used for classifying the userreviews.

c) Classification

Once the features are extracted, the classification of the movie reviews isdone using ILRalgorithm. The classification technique is implemented by combining both joint distribution andthe input to output mapping techniques. Which means the selected feature for classifying thereview will be compared with similar words as well as the word with similar meaning. This isdonebyusingtheintegrationPOSTaggingwhichwill beclassifiedasasimilargroupofreview.Thiswillbecarriedoutusingdifferentstepswhichisdescribedasfollows:

3) Support count forsplitting theinputdataset

Support count is the value for splitting the input dataset which will be determined based on thesize and number of reviews used in the training dataset. Before selecting features like targetvariable for the classification, we need to set the support count for splitting the input dataset. Inthis work, the support count is set based on the number of reviews considered for analysis andsplittingtheinputdataset,wecanprocessthedatafasterorwecandoparallelprocessing.

$$vect=CountVectorizer(min\_df=count\_value)......................(4)$$

The equation 4 specifies *vect*variable which takes count of vectorizer that can be referred as asimple way to tokenize acollection of textdocuments and build the vocabulary of knownwords. *min_df*defines the support count value for the input dataset which is considered forclassification.

4) Classifyingbasedoninput reviews

Thismoduledescribestheunlabeledinputdatasetthatistakenforanalysesandwillbeclassified based on the type of reviews. Here POS module is integrated for classifying thereviews based onmultiple independentvariables with similarmeaning which can be classifiedas similar group described in equation 5. Here *ngram_range*describes the lower and upperboundaryoftherangeof2-valuesfordifferentn-gramstobeextracted.Intheproposedtechniquewe have considered (2,2) as upperand lowerbound asacutoff,because theproposedtechnique worksforbigramfeatures.

$$ngram\_range=(a,b)....................................(5)$$

The ILR is also based on a bilinear equation module with multiple independent input parametersas in linear regression to predict the probability of the input belonging to a specific class. Apossible output that represents a class. Using bilinear function, the output range can vary fromlessthan1tovaluesover0.TheImprovedlogisticfunctioncanbeexpressedasin equation6,

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

Equation 6 represents the rule producing output $P(x \mid y)$, the probability of textual document *X*belonging to the class*Y,* where $X = \{x1,x2,x3,...xn\}$ is the feature vectorof the text documentand $Y = \{y1, y2,...,yk,...yn\}$ is the output class for each *b* items. It is combined withexisting LR classifierthat has an additional sigmoid function (e^{z})representing the input datawithathresholdparameterfordecisionvariable.

The working of ILR based classification model is describes below considering an example ofuser review for a particular movie. User review is “the movie was good, but the cinematographywas too worst music was horrible, comedy was better and music was too good, overall the moviisonce watchable”

This reviewis classifiedusingILRthroughfollowingsteps:

- Step1:Applypre-processingstepsdiscussedin3.2.1sectionthatresultsinremovaloffrequently occurring words like ‘the’, was, ‘is’ etc, the misspelled word movi is replaced by thecorrect word movie after applying POS tagging technique 18 words out of 27 words will beretrieved. Outputafterapplyingpre-processing:movie good butcinematography too worstmusichorrible,comedybetterandmusictoogood,overallmovieoncewatchable
- Step 2: Apply feature extraction process that groups the combinations Adverbs + Adjectives(AAJ), Adverbs + Verbs (AV), Adverbs + Adjectives + Verbs (AAJV) and Adjectives + Verbs(AJV)assentimentfeaturesofmovie basedapplications.

- Step 3: Apply the support count forthe input review. By referring equation (4), we haveconsidered support count value as 5 for parallel processing of reviews. Then 5 words out of 18wordsareseparatedintofourdifferentgroupsforparallelprocessing.
- Step4:Next,multipleindependentwordswithsamemeaningareprocessedatatime.Considering the value of a and b as 2 in equation (5),the review "good" , "too good" and betterare treated similar words during classification for the input review considered, hence total wordsduringclassificationwillbecome15outof18.Outputafterapplyingpre-processing:movie good butcinematography too worstmusichorrible,comedybetterandmusictoogood,overallmovieoncewatchable
- Step 5: Equation (6) is considered to classify the negative and positive set of reviews based onthe prediction attributes of the dataset. If we apply this to the input review, the probability ofpositiveoccurrenceofpositivewordsis12/15andtheprobabilityofnegativeoccurrenceof words is 3/115. Hence the given review is classified as positive because of more positive wordsinthereview.Bythiswecanachievearound85%classificationaccuracy.
  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 predictionaccuracywhenwetrainthedatasetwithmorenumberofinputreviews.
- Step6:Plotthegraphagainsttheclassificationaccuracy,timetakenforclassification,precision,recallandFmeasureofproposedILRandcomparewithexistingLRandNBclassifiers.

V. EXPERIMENTAL RESULTS

The implementation of proposed work is carried out using anaconda 4.3.8, python 3.6.3 and theopen source libraries suitable for analyzing the movie reviews. Matplotlib toolkit is used fordrawingtheresults.Thebelow Table2 providestheparametersconsideredfortheimplementationof theproposedwork.

Table2:Implementationparameters

Dataset	MovieDataset
Source:	http://ai.stanford.edu/~amaas/data/sentiment
TotalNumberofReviews	50000
Number of reviewsconsideredfortraining:	25000
Number of reviews consideredfortesting	25000
Numberofmaximumreviews consideredforasinglemovie	30
Total number of movies considered	850
Technologyused	Python3.6.3

The performance of proposed ILR is compared with existing logistic regression and naïve bayesclassifiers for different set of reviews against various performance parameters like classificationaccuracy,timetakenforclassification,precision,recallandF-measure.

A. *Classification Accuracy:*

The Figure 2 describes the accuracy of classification for movie based reviews, where x-axisrepresentsdifferentsetoftestreviewsconsideredandy-axisrepresentstheclassificationaccuracy.TroughtheproposedILRanaverageof88%classificationaccuracyhasbeenachieved,whichis15% morewhencomparedwithexistingLRandNBclassifiers.

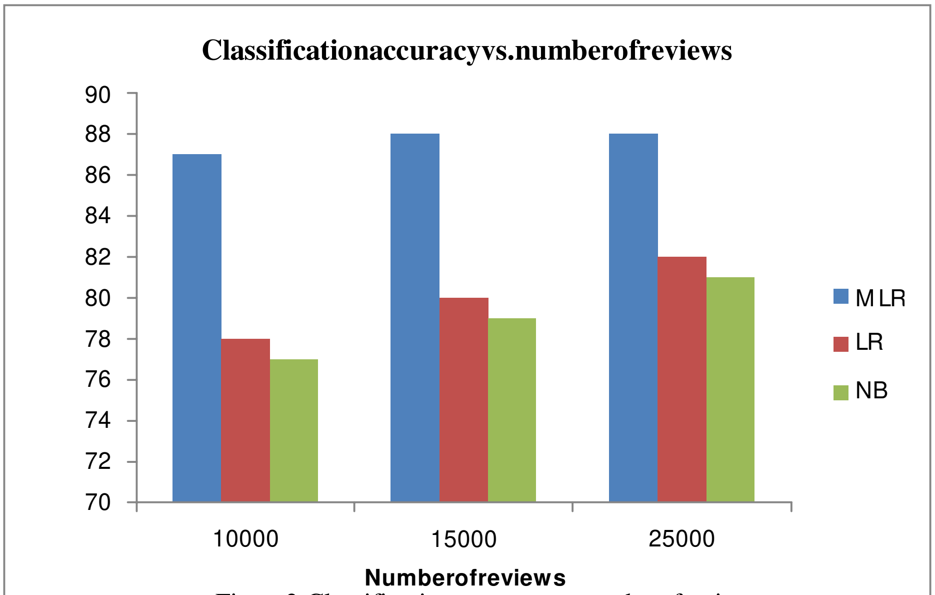

Figure2:Classificationaccuracyvs.numberofreviews

B. Time-takenforClassification:

The Figure 3 describes the time taken to classify the various instance of test reviews, where x-axisrepresentsthe time taken to classifyvariousinstance of reviewsusing proposedILRtechnique, existing LR and Naïve Bayes classifiers against the various instance of reviews andproves the proposed ILR is taking less time for classification because of parallel processingwhen compared to exiting techniques even after varying the size of the dataset with differentnumberof reviews.

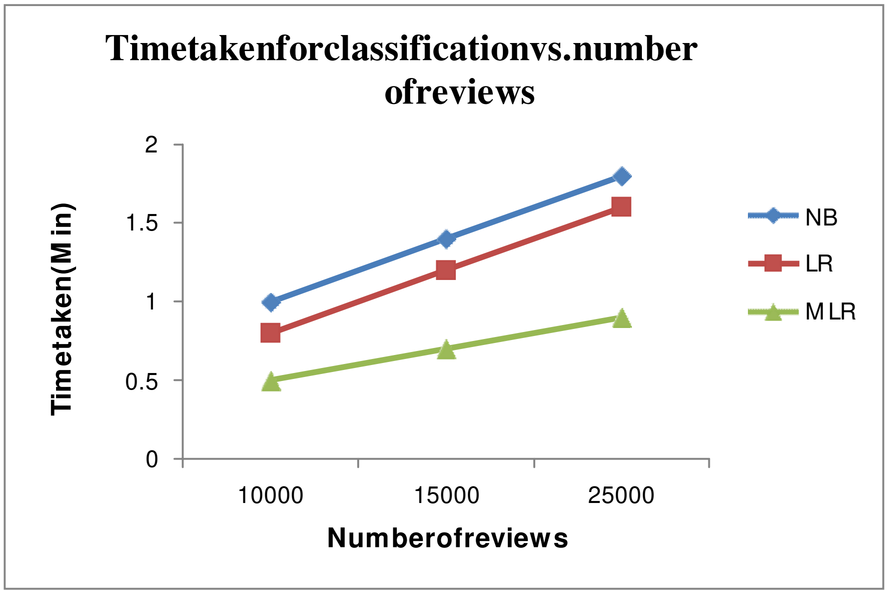

Figure3:Timetakenfortheclassificationofinput reviewsvs.numberofreviews

C. Precision

Itisdefinedastheratioofcorrectlyclassifiedovernumberofallclassificationswhichcanbe expressedas:

Precision= correctlyclassified/(correctlyclassified+Errorlyclassified)

The below Figure 4 describes the accuracy of precision value in percentage against proposedILR, existing LR and NB classifiers and proves the proposed ILR ishaving more precisionvaluebecauseoflessnumbero fErrorlyclassifiedwords whencomparedwithexitingtechnique.

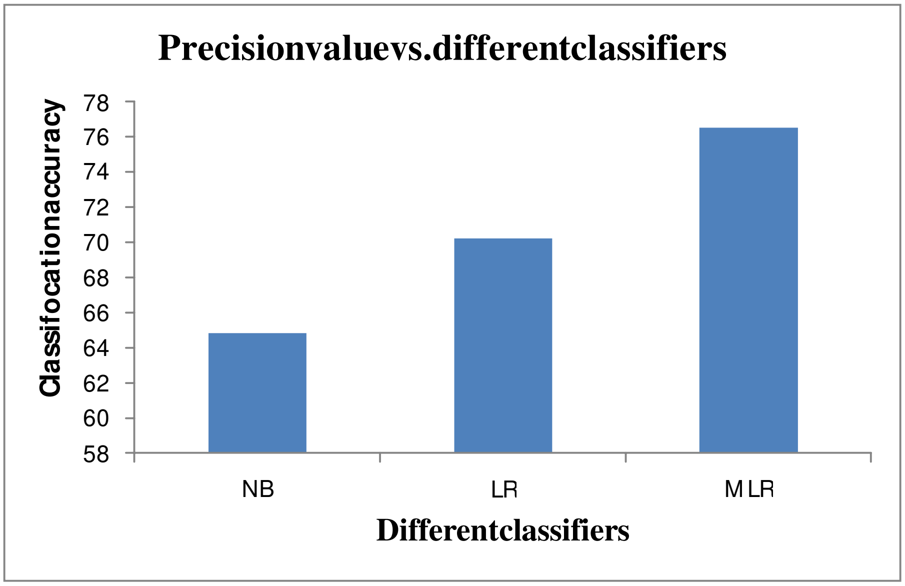

Figure4:Precisionvaluevs. DifferentClassifiers

D. Recall

Itisconsideredtodeterminethenumberoftruepositivefunctionwhichcanbeexpressedas:

Recall= correctlyclassified/(correctlyclassified+ Missedclassified)

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 becauseoflessnumberofmissclassified wordswhencomparedwith exitingtechniques.

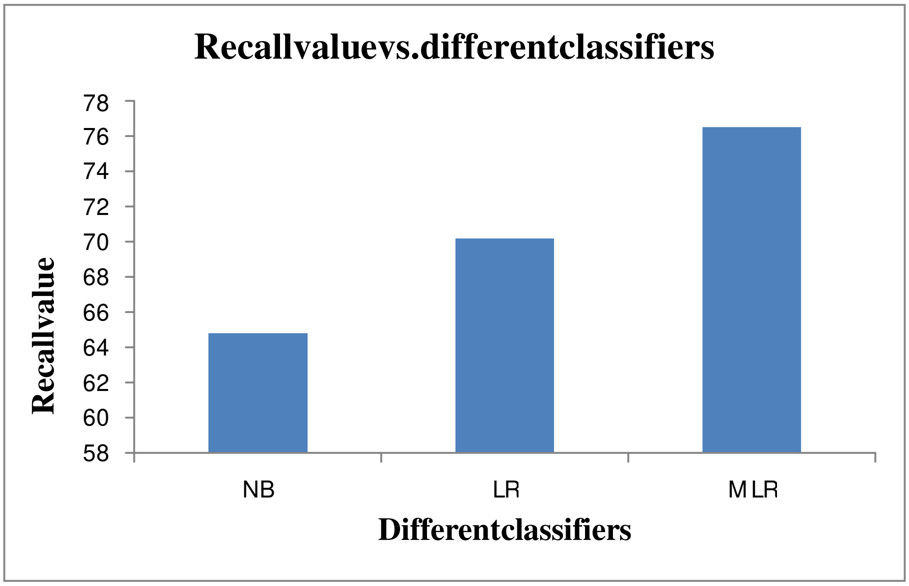

Figure5:Recallvaluevs.DifferentClassifiers

E. F-Measure

Itisacombinedmeasureforprecisionandrecallvalueswhichcanbeexpressedas:

F-Measure=2*Precision*Recall/(Precision+ Recall).

The below Figure 6 describes the accuracy of F-measure value in percentage against proposedILR, existing LR and NB classifiers and proves the proposed ILR is having more F-measurevalue because of more precisionand recallvalues when Compared with exiting techniques.

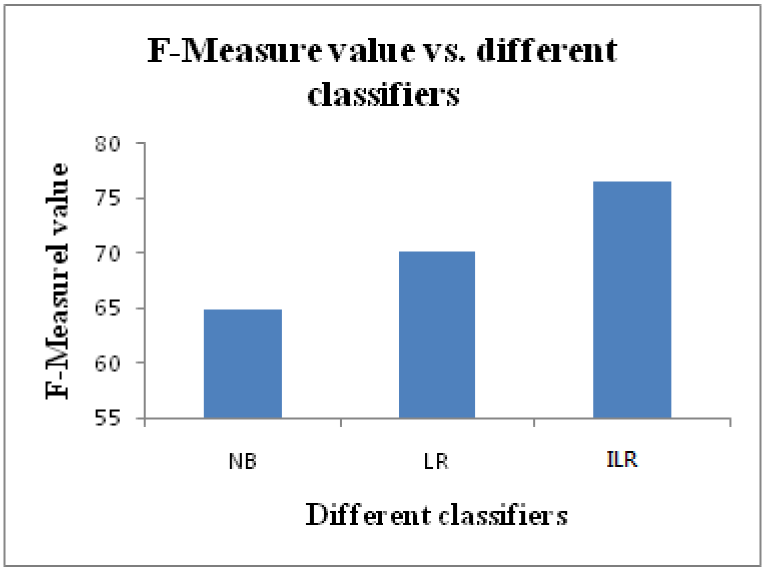

Figure6:F-measurevaluevs.Different Classifiers

VI. CONCLUSION AND FUTURE WORK

The analysis and classification of various movie based reviews is taken from different moviebased applications.Differentclassifiersare used toclassify the reviewson the movies likeNaive bayes, Logistic Regression, Support Vector Machine etc., The existing classifiers fails inachieving the desired accuracy, because the classifiers does not work properly with multipleindependent variables i.e. word with similar meaning is treated as separate for the classificationthat affects the performance parameters. While classification, the proposed work addressed thetwo-class problem which is the main drawback in the existing LR classifier.With the proposedclassifier 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 andachieved better performance.In future, this work can be extended on mining the reviews frommultiple applications such as Bookmyshow, Paytm etc. Further improved machine learningalgorithms can be incorporated to improve the efficiency, which will help in deciding the bestclassificationclassifierinsentimentalanalysis.

BIBILOGRAPHY

[1] Farkhund Iqbal, JahanzebMaqbool,Benjamin C. M. Fung,RabiaBatool,Asad Masood Khattak,SaiqaAleem, Patrick C. K. Hunga, “A Hybrid Framework for Sentiment Analysis Using GeneticAlgorithmBasedFeatureReduction”,IEEE,vol.7,pp.14637-14652,2019.

[2] Tu Nguyen Thi Ngoc, Ha Nguyen Thi Thu, Viet Anh Nguyen, “Mining aspects of customer’s reviewonthesocialnetwork”,JournalofBigData, vol. 6, Springer, Number 1, pp 6-22.Articlenumber: 22,2019

[3] K. L. S. Kumar, J. Desai and J. Majumdar, "Opinion mining and sentiment analysis ononlinecustomer review," IEEE International Conference on Computational Intelligence and ComputingResearch(ICCIC),pp. 1-4, 2016

[4] Sari Widya Sihwi, InsanPrasetyaJati, RiniAnggrainingsih, “Twitter Sentiment Analysis of MovieReviews Using Information Gain and Naïve Bayes Classifier”, IEEE International Conference onApplication forTechnologyofInformationandCommunication(iSemantic),pp.190-195,2018

[5] MariumNafees,HafsaDar,IkramUllahLali,Salman Tiwana,“Sentiment Analysisof Polarity inProductReviewsInSocialMedia”, 14thInternationalConferenceonEmergingTechnologies(ICET), pp. 1-6, 2018

[6] N. Banik and M. Hasan Hafizur Rahman, "Evaluation of Naïve Bayes and Support Vector Machineson Bangla Textual Movie Reviews," International Conference on Bangla Speech and LanguageProcessing(ICBSLP),Sylhet, pp. 1-6,2018

[7] PeimanBarnaghi, John G. Breslin, ParsaGhaffari, “Opinion Mining and Sentiment Polarity on Twitterand Correlation Between Events and Sentiment”, Oxford, Second International Conference on BigDataComputingServiceandApplications, pp. 52-57,2016.

[8] Wang,Yequan,AixinSun,JialongHan,YingLiu,andXiaoyanZhu."Sentimentanalysisbycapsules."InProceedings ofthe2018worldwidewebconference,pp.1165-1174.2018

[9] Chantal Fry, Sukanya Manna, “Can we Group Similar Amazon Reviews: A Case Study with DifferentClusteringAlgorithms”, TenthInternationalConferenceonSemantic Computing,pp.374-377,2016.

[10] Asha S Manek, P Deepa Shenoy, M Chandra Mohan, Venugopal K R, “Aspect term extraction forsentiment analysis in large movie reviews using Gini Index feature selection method and SVMclassifier”,WorldWideWeb,vol.20,Springer, Number2, pp.135-154, 2017

[11] Haiyun Peng, Erik Cambria, Amir Hussain, “A Review of Sentiment Analysis Research in ChineseLanguage”,CognitiveComputation,vol.9, Springer,Number4,pp.423-435, 2017

[12] J. Zheng and L. Zheng, "A Dictionary-Based Convolution Recurrent Neural Network Model forSentiment Analysis", 2019 International Conference on Communications, Information System andComputerEngineering(CISCE),Haikou, China,pp. 606-611,2019

[13] N. Mtetwa,A.O. Awukam andM.Yousefi,"Feature ExtractionandClassificationof MovieReviews,"5thInternational ConferenceonSoftComputing &MachineIntelligence(ISCMI), Nairobi, Kenya, pp.67-71,2018

[14] S. Rajalakshmi, S.Asha, N.Pazhaniraja, “A Comprehensive Survey on Sentiment Analysis”, 4thInternational Conference on Signal Processing, Communications and Networking (ICSCN -2017),pp.1-5,2017.

[15] Harpreet Kaur, VeenuMangat, Nidhi, “A survey of sentiment analysis techniques”, InternationalConference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), , pp. 921-925,2017.

[16] Vikas K Vijayan,K. R. Bindu,LathaParameswaran, “A comprehensive study of text classificationalgorithms” ,IEEEInternationalConferenceonAdvancesinComputing,CommunicationsandInformatics (ICACCI), , pp. 1109-1113,2017.

[17] X. Lei, X. Qian and G. Zhao, "Rating Prediction Based on Social Sentiment From Textual Reviews,"inIEEETransactions onMultimedia,vol.18,Number9, pp.1910-1921, Sept.2016.

[18] Parkhe V. & Biswas B. “Sentiment analysis of movie reviews: finding most important movie aspectsusingdrivingfacto rs”,SoftComputing,vol.20,Springer,pp.3373-3379, 2016.

[19] KetanSarvakar, Urvashi K Kuchara, “Sentiment Analysis of movie reviews: A new feature-basedsentiment classification”, International Journal of Scientific Researchin ComputerScience andEngineering,vol.6, Issue.3,pp.8-12, 2018.

[20] DoaaMohey El-Din Mohamed Hussein, “A survey on sentiment analysis challenges”, Journal ofKingSaudUniversity–EngineeringSciences,vol.30,Elsevier, pp330–338, 2018

[21] WalaaMedhat, Ahmed Hassan, HodaKorashy,“Sentiment analysis algorithms and applications: Asurvey”,AinShamsEngineeringJournal,vol.5 Elsevier,Issue4,pp1093-1113,201 8http://ai.stanford.edu/~amaas/data/sentiment-Datasetconsideredfor classification.