Dr. U. D. Prasan, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 13, Issue 4, April 2023, pp. 60-64

RESEARCH ARTICLE OPEN ACCESS

SentimentAnalysisofAmazonProductReviewsUsingMachineLearningAlgorithms

Dr.U.D.Prasan[1],M.Ruchita[2],T.Dinesh[3],S.Manohar[4],S.Prasanth[5],Saurab[6]
[1]Professor&HODofComputerScienceEngineering,AdityaInstituteofTechnologyandManagement(AITAM),Tekkali,AndhraPradesh,India
[2,3,4,5,6] Students, IV B. Tech CSE, Aditya Institute of Technology and Management (AITAM), Tekkali, AndhraPradesh,India

ABSTRACT
Sentimentanalysisor opinionminingisoneofthemajor tasksofNLP(NaturalLanguageProcessing).Sentimentanalysis has gained much attention in recent years. Nowadays customers wanted to purchase anything just at oneclick of a mouse button. Online shopping is becoming even more popular due to its high level of convenience.Online sellers and merchants ask their purchasers to share their opinions about the products they have bought. Asa result, millions of reviews are generated daily, which makes it difficult for a customer to make a good decisionor whether to buy the product or not. Analyzing large number of reviews is also hard and time consuming forproduct manufacturers. The goal of our project is to understand and analyze the Amazon user review data set toidentifywhethermostofthereviewsarepositiveornegativewiththehelpoflogisticregressionandsupportvectormachineclassificationalgorithms.

Keywords **–** Classification, Natural Language Processing, Logistic Regression, Sentiment analysis, SupportVectorMachine

---

Date of Submission: 01-04-2023 Date of acceptance: 11-04-2023

I. INTRODUCTION

Sentiment is an emotion or attitude prompted bythe feelings of the customer. Sentiment analysis isoftenreferredtoasopinionmining,becausetheopinion given by the customer will be mined to revealthe polarity of the opinion. It comes under machinelearning.Sincetheonlinedataistremendouslygrowingday-by-day,itisconsideredtobeveryimportant. As customers express their reviews andthoughtsaboutthebrandmoreopenlythaneverbefore,sentimentanalysishasbecomeapowerfultooltomonitorandunderstandonlineconversations.Sentimentanalysisisaprocesswhere thedatasetconsists of emotions, attitudes or assessment whichtakesintoaccountthewayahumanthink.Inasentence, trying to understand the positive and thenegativeaspectisaverydifficulttask.Peopleareabletoposttheirowncontentthroughvarioussocialmedia, such as forums, micro-blogs, or online socialnetworkingsites.Onlinedatahaveseveralflaws.Thefirstflawisthatsincepeoplecanfreelyposttheirowncontent,thequalityoftheiropinionscannotbeguaranteed. The second flaw is that ground truth ofsuchonlinedataisnotalwaysavailable.Agroundtruth is more like A tag of a certain opinion aboutindicatingwhether the opinion is positive or negative. The purposeof our project is to analyze and understand the overallsentimentexpressedbycustomersaboutaparticularproductontheAmazonplatform.Sentimentanalysisisatechniquethatinvolvesusingnaturallanguageprocessing (NLP) and machine learning algorithms toautomaticallyidentifyandextractthesubjectiveinformationfromtextualdatasuchasproductreviews.

II. LITERATURESURVEY

Therearemanyworksdoneintheareaofsentimentanalysis.

R.SinghandA.Singh[1]paperdiscussestheuseof a decision tree algorithm for sentiment analysis ofAmazon product reviews. The authors first discusstheimportanceofsentimentanalysisintoday'sworldand how it can be used to make informed decisions.They thenexplainthedatasetusedinthestudy,which consisted of 5000 Amazon product reviews.The dataset was preprocessed to remove stop wordsandperformstemming.Theresultsofthestudyshowed that the decision tree algorithm achieved anaccuracy of 77.6%, with precision,recall,andF1scoreof0.78,0.77,and0.77respe

ctively. Theauthors finally conclude that decisiontreealgorithmisausefultoolforsentiment analysis of Amazon product reviewsandcanbeusedtomakeinformeddecisionsabou tproducts.

S. S. Shrivastava and S. K. Shukla [2] paperfocuses on the use of the random forest algorithmfor sentiment analysis of Amazon product reviews.Theauthorsbeginbyintroducingtheimportanc eofsentiment analysis in today's world and how it canbe used to make informed decisions. They thendiscussthedatasetusedinthestudy,whichconsisted of 1000 Amazon product reviews. Thedataset was preprocessed to remove stop words,performstemming,andconvertthetextintoanum erical representationusing the bag-of-wordsmodel. The results of the study showed that therandom forest algorithm achieved an accuracy of84.8%, with precision, recall, and F1 score of 0.87,0.85,and0.84respectively.

S.S.ReddyandV.Krishnaiah[3]paperdiscusses the use of the K-nearest neighbor (KNN)algorithmforsentimentanalysisofAmazonprod uctreviews. The results of the study showed that theKNN algorithm achievedanaccuracy of82.3%,with precision, recall, and F1 score of 0.83, 0.82,and 0.82 respectively. The authors finally concludethat the KNN algorithm isa useful tool for sentimentanalysis of Amazon product reviews andcanbeusedtomakeinformeddecisionsaboutproduc ts.

A. Mukherjee and S. K. Das [4] paper focusesontheuseoftheNaiveBayesalgorithmforsentim entanalysisofAmazonproductreviews.Theauthorsbeg inbydiscussingtheimportanceofsentiment analysis and its applications in variousfields. They then describe the dataset used in thestudy, which consisted of 1000 Amazon productreviews. The dataset was preprocessed to removestop words and perform stemming. The results ofthe study showed that the Naive Bayes algorithmachievedanaccuracyof81.7%,withprecision ,recall,andF1scoreof0.82,0.82,and0.81respectively. The authors conclude that the NaiveBayesalgorithmisausefultoolforsentimentanaly sis of Amazonproduct reviews and can beusedtomakeinformeddecisionsaboutproducts.

M.H.Rahman,M.M.Rahman,andM.H.Kabir [5]paperpresentsacomparativestudyofthreepopularm achinelearningalgorithms,namelyK-NearestNeighbor(KNN),DecisionTree,andRandom Forest, for sentiment analysis of Amazonproduct reviews. The authors start by introducing theimportanceofsentimentanalysisanditsapplicationsi nvariousfields.Theythendescribethedatasetusedinthes tudy,whichconsistedof2000Amazonproductreviews. Thedatasetwaspreprocessedtoremovestopwordsandt o performstemming.The resultsofthestudyshowedthattheRandom Forest performed the best with an accuracyof87.35%,followedbyDecisionTreewithana ccuracy of 82.5%, and KNN with an accuracy of80.5%. The authors conclude that Random Forest isthe most suitable algorithm for sentiment analysis ofAmazonproductreviews.

S. S. Patil, K. R. K. Singh, S. S. Deokate [6]paper proposes a machine learning-based approachfor sentiment analysis of Amazon product reviews.The authors start by introducing the importance ofsentiment analysis and its applications in variousfields. They then describe the dataset used in thestudy, which consisted of 3000 Amazon productreviews. The dataset was preprocessed to removestop words, punctuation, and perform stemming.NaiveBayesandLogisticRegressionachiev edaccuracies of 78.4% and 80.2%, respectively. Theauthorsconclude that SVM is themostsuitablealgorithmforsentimentanalysisofAm azonproductreviews.

B.C.PatilandV.R.Reddy[7]paperproposesaK-NearestNeighbor(K-NN)classification algorithm for sentiment analysis ofAmazon reviews. The authors start by introducingtheimportanceofsentimentanalysisanditsa pplications in various fields. They then describethe dataset used in the study, which consisted of1500Amazonreviews.Thedatasetwaspreprocessed to remove stop words, punctuation,and performstemming.The results of the studyshowedthattheK-NNalgorithmachievedthehighestaccuracyof76.8%wh enK=7.Theprecision,recall,andF1scoreforthepositive ,negative, and neutral sentiment categories were alsoreportedfordifferentvaluesofK.

A.BhattacharyaandS.Basak[8]paperproposesacom parisonofNaiveBayes algorithmandDecision Tree algorithm for sentiment analysis ofAmazonproductreviews.The authorsbeginbydiscussingtheimportanceofsentimenta nalysisandthechallengesassociatedwiththe sentiment analysis.Theythendescribethedatasetusedinthestudy, whichconsistedof500Amazonproductreviews.Thedat aset was preprocessed by removing stop words,punctuation, white spaces and performing stemming. The accuracy of Naive Bayes algorithm is greater than the accuracy of the Decision Tree

algorithm. The resultsofthestudyshowedthattheNaiveBayesalgorithmachieved an accuracy of 74.4%, while the DecisionTreealgorithmachievedanaccuracyof72.8%. Theprecision,recall, andF1scoreforthepositive, negative,andneutralsentimentcategorieswerealso reportedforboth Naive Bayes algorithm and Decision Tree algorithm.

N.Naeem,N.Naureen,andM.Shahbaz[9]paperproposedasentimentanalysisapproach using the random forest algorithm toclassify Amazon product reviews as positive,negative,orneutral.The authorscollectedadatasetofAmazonproductreviews,preprocessed the data, extracted features, andtrainedtherandomforestmodel.Theexperimental results showed that the proposedapproach achieved good accuracy in sentimentanalysis of Amazon product reviews. The paperconcluded that the random forest algorithm canbe usedeffectively forsentimentanalysisofAmazon productreviews.

III. IMPLEMENTATION

Amazon is one of the largest E-commerce sitesas for that there are innumerous amount of reviewsthatcanbe seen. We useddata namedAmazonproduct data which was provided by Kaggle. Ourdataset comes from Consumer Reviews of AmazonProducts. This datasethas 30,847 rows and 12attributes. Each example includes the type, name ofthe product as well as the text review and the ratingof the product etc. For preparing the desired data asimple code was written in python to remove theuselessfeatures.WeusedvariousNaturalLanguageProcessing(NLP)techniquestopreparethedata.

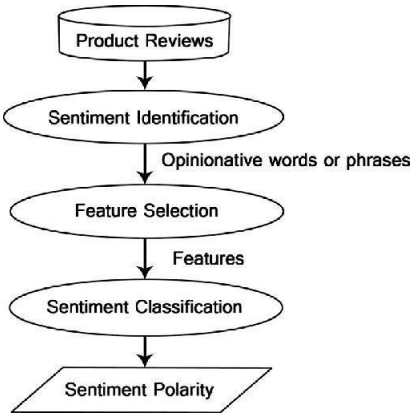

Fig:workflow

Tokenization:Itistheprocessofseparatingasequence of strings into individuals such as words,keywords, phrases, symbols and otherelementsknown as tokens. Tokens can be individual words,phrases or even whole sentences. In the process oftokenization,somecharacterslikepunctuationmarks are discarded. The tokens work as theinputfordifferentprocesslikeparsingandtextmining.

Stop word removal: Stop words are those objects in asentencewhicharenotnecessaryinanysectorintextmining.Sowegenerally ignorethesewordstoenhance the accuracy of the analysis. In different formatthere are different stop words depending on the countrylanguage etc. POS tagging: The process of assigning oneof the parts of speech to the given word is calledParts ofSpeechtagging.ItisgenerallyreferredtoasPOStagging. Parts of speech generally contain nouns, verbs,adverbs,adjectives,pronouns,conjunctionandtheirsubcategories. Parts of speech tagger or POS tagger is aprogramthatdoesthisjob.TF-IDF:TF-IDFisaninformation retrieval technique which weighs a term'sfrequency(TF)andalso inverse documentfrequency(IDF). Each word or term has its own TF and IDF score.The TF and IDF product scores of a term is referred totheTF*IDFweightofthatterm.Simplywecanstatethat the higher the TF*IDF score (weight) the rarer the termand vice versa. TF of a word is the frequency of a word.IDFofawordisthemeasureofhowsignificantthatermisthroughoutthecorpus.Convertingallthecapitalletterstoalowercase.Stemmingandreducinginflectionalforms to a stemma form. Lemmatizing to group togetherthe different inflected forms of a word so they can beanalyzedasasingleitem.Theinputdatacanbetransformedintoareducedsetoffeatures(featurevectors).This process is called feature extraction.Forfeaturelabeling,weusetwofilescontainingpositiveandnegativewordscollectedfromthedictionary.Theresultant set of pre- processing is compared with thesefiles.Thepositivewordsare labeledas'1'whereasnegativearelabeledunder'0'.

IV. ALGORITHMS

Logistic Regression and Support Vector Machineare the best supervised machine learning algorithmstoclassifythedata.

4.1 LogisticRegression

Logistic Regression is one of the most popular machine learning algorithms. Which comes under the supervisedLearningtechnique.Itisusedforpredictiontheca tegoricaldependentvariableusingagivensetofindep endent variables. Logistic Regression predicts theoutputofacategoricaldependentvariable.Therefore,

theoutcome must be a categorical or discrete value. It can beeither yes orno,0 or 1, true or false etc. but instead ofgivingtheexactvalueas0and1itgivestheprobabilistic valueswhichliebetween0and1.LogisticRegression ismuch similar to the Linear Regression except that howtheyareused.LinearRegressionexcept thathowtheyareused.Linear Regressionis used to solve Regressionproblems.WhereasLogisticregressionisus edforsolvingtheclassificationproblems.In Logistic Regression, instead of fitting a regression line, we fit an "S"shapedlogisticfunctionwhichpredictstwomaximu m values (0 or 1). Logistic Regressionisasignificant machine learning algorithm because it has the ability to provide the probabilities and classify newdatausingcontinuousanddiscrete datasets.

4.2 Support VectorMachine

Support Vector Machine is a universal learner.Support Vector Machine has defined both inputand output format. The output is either positive ornegativeandinputisvectorspace.Thetextdocument is not suitable for learning. Those textsare transformed into a structured format. The textis transformed into a format which matches intotheinputofmachinelearningalgorithm.Thescore of the texts is calculated and then the scoreisgivenasinputtoSupportVectorMachine. SupportVectorMachine(SVM)hasbeenproved oneofthemostpowerfullearningalgorithmsfortextCate gorizationbut textcategorization sometimemayproduceoccurs.Todecidewhichoneisbet terbetween texts a comparison of text classifier isrequired. SVMs can efficiently perform nonlinearclassification. SVM simultaneously minimize theempirical classification error and maximize thegeometric margin. So SVM is called MaximumMarginClassifiers.

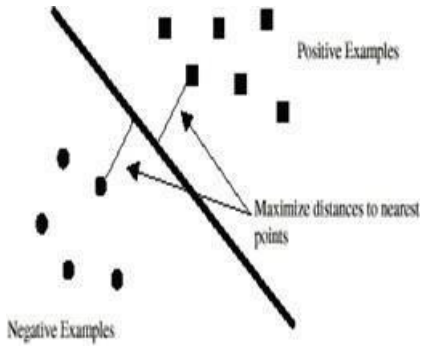

Fig:supportvectorclassifier

SVMchoosestheextremepoints/vectorsthath elpin creating the hyperplane. These extreme casesare called as support vectors, and hence algorithmistermedasSupportVectorMachine.Thegoal ofthe SVM algorithm is to create the best line ordecisionboundarythatcansegregaten-dimensionalspace intoclassessothatwecaneasilyputthenewdatapointinth ecorrectcategory in the future. The performance measureisusedinthiscase.

V. CONCLUSION

Sentiment analysis deals with the classification oftexts based on the sentiments they contain. Thisfocusesonatypicalsentimentanalysismodel consistingofthreecoresteps,namelydatapreparation,re viewanalysisandsentimentclassification and describes representative techniquesinvolvedinthosesteps.Sentimentanalysisis anemergingresearchareaintextminingandcomputatio nallinguistics,andhasattractedconsiderable research attention in the past few years.Futureresearchshallidentifyneutralreviewsalso andexploresophisticatedmethodsforopinionandprodu ctfeatureextraction,aswellasnewclassificationmodels that can address the ordered labels property inrating inference. Applications that utilize results fromsentiment analysis is also expected to emerge in thenearfuture.

Algorithm	Accuracies
LogisticRegression	91%
RandomForest	87%
KNN	84%
Decisiontree	85%
SupportVectorMachine	91%

Fig: tableofaccuracyscoresofvariousmodels

Thus, we get maximum accuracy in Logistic RegressionandLinearSVC.So,wechooseLogisticRegr ession andLinearSVCasourfinalalgorithmsforsentimentanal ysis ofamazonproductreviews.

REFERENCES

[1]. R. Singh and A. Singh, "Sentiment analysis ofAmazonproductreviewsusingdecisiontreeal gorithm,"20193rdInternationalConferenceon ComputingMethodologiesandCommunication (ICCMC),Erode,India,2019,pp.1064-1067,doi:10.1109/ICCMC.2019.8725549.

[2]. S. S. Shrivastava and S. K. Shukla, "SentimentanalysisofAmazonproductreviews usingrandomforestalgorithm,"20175thInternat

ionalConferenceonAdvanced Computing & Communication Technologies(ACCT),Rohtak,India,2017,pp.95-98,doi:10.1109/ACCT.2017.16.

[3]. S.S.ReddyandV.Krishnaiah,"SentimentanalysisofAmazonproductreviewsusingKNNalgorithm,"2016IEEEInternationalConferenceonRecentTrendsinElectronics,Information&CommunicationTechnology(RTEICT),Bangalore,India,2016,pp.804807.Poi.10.1109/RTEICT.2016.7807884.

[4]. I.MukherjeeandS.KDas.,"SentimentanalysisofAmazonproductreviewsusingNaiveBayesalgorithm," 2016 IEEE 7th Annual InformationTechnology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, 2016, pp. 1-6,doi:10.1109/IEMCON.2016.7746337.

[5]. M. H. Rahman, M. M. Rahman, and M. H.Kabir, "A comparative study of KNN, Decision Treeand Random Forest algorithms for sentiment analysisof Amazon product reviews," 2020 IEEE InternationalConference on Computer, Communication, Chemical,MaterialsandElectronicEngineering (IC4ME2),Dhaka,Bangladesh,2020,pp.1-4,doi:10.1109/IC4ME2.2020.9334648.

[6]. S.S.Patil,K.R.K.Singh,S.S.Deokate,S. S. Ingle, "Sentiment Analysis of Amazon ProductReviews using Machine Learning," 2021 InternationalConferenceonComputingandInformationTechnologies (ICCIT), Wollongong, Australia, 2021,pp.1-6,doi:10.1109/ICCIT52534.2021.9422872.

[7]. B.C.PatilandV.R.Reddy,"SentimentAnalysisofAmazonReviewsusingK-NNClassification,"2021InternationalConferenceonEmergingTrendsinInformationTechnologyandEngineering(icETITE),Pune,India,2021, pp.1-5,doi:10.1109/icETITE51432.2021.9451919.

[8]. A.BhattacharyaandS.Basak,"SentimentAnalysisofAmazonProductReviewsusingNaiveBayesandDecisionTree,"2021InternationalConferenceonInventiveResearchinComputingApplications (ICIRCA), Coimbatore, India, 2021, pp.239-243,doi:10.1109/ICIRCA52104.2021.9603899.

[9]. N.Naeem,N.Naureen,andM.Shahbaz,"SentimentAnalysisofAmazonProductReviewsUsingRandomForest,"2020InternationalConferenceonAdvancesinComputing, CommunicationControl and Networking(ICACCCN),Bangalore,India,2020,pp170174,doi:10.1109/ICACCCN49741.2020.9289365.