Sentimental Analysis-Based Recommended System for Products Using Machine Learning

B. G. Mamatha Bai, S. R. Likhith, and Salma Itagi

Abstract Sentiment Analysis is widely used in the process of mining the data, to predict emotion of a sentence through Natural Language Processing (NLP). The main aim is to find the accurate polarity of a sentence. Therefore, to find the polarity or sentiment of a user or customer for a product there is a need for automated data analysis techniques. In this paper, a detailed analysis of classification techniques is used in Sentimental Analysis of Amazon Product Reviews with recommendation for a best buy product on web. Multinomial Naïve Bayes, Random Forest, Logistic Regression, Decision Tree, and SVM classifiers are tested and compared. Random Forest gives the best accuracy of 94.94%. Web scraping extracts five Amazon products on Amazon.com and recommends the best buy product on the basis of polarity score of each product and here Samsung Galaxy M01 is recommended as the best buy product.

Keywords Sentimental analysis · Polarity · Machine learning · Classification · Web scraping

B. G. Mamatha Bai
Department of Computer Science Engineering, Nitte Meenakshi Institute of Technology, Bengaluru, Karnataka, India

S. R. Likhith (✉)
Department of School of Computer Science and Engineering and Information Science, Bio-Intelligence Lab Presidency University, Bengaluru, Karnataka, India
e-mail: anandlikith111@gmail.com

S. Itagi
Department of Computer Science Engineering, Sai Vidya Institute of Technology, Bengaluru, Karnataka, India
e-mail: salma.itagi@saividya.ac.in

N. R. Shetty et al. (eds.), *Advances in Computing and Information*, Lecture Notes in Electrical Engineering 1104, https://doi.org/10.1007/978-981-99-7622-5_14

1 Introduction

Online shopping is in the rise in today's day to day life. Electronic business districts like Amazon are made to satisfy the rising need. Subsequently, something specific can be bought on a few districts and the expenses could change. As clients need the best quality essentially cost at this point, all around, can't directly check it, reviews from various clients show up, obviously, to be the strongest technique for picking the decision about whether to buy the thing. Thusly, Feeling Evaluation has shown essential to comprehend a thing's remaining among the buyers beginning with one side of the planet then onto the following. The mark of this paper is to explore expecting contemplative examination is possible for the request for thing reviews from Amazon.com. Multinomial Naïve Bayes, Linear Support Vector Machines, Decision Trees, Random Forest, and Logistic Regression classifiers are implemented and are compared; the performance of these classifiers are computed on the classification (positives, negatives, and neutral) for the product reviews from Amazon.com.

The rest of the document is formatted in the same way. In Sect. 2, the most current efforts for Sentimental Analysis are discussed. The proposed system is outlined in Sect. 3. The process for designing the system is discussed in Sect. 4. The results of the constructed system are examined in Sect. 5. In Sect. 6, the conclusion is formed about the future scope.

2 Related Work

Feeling Evaluation is one of the most quickly making study areas in PC programming, making it hard to keep up of the movements generally. A client input examinations on thing, uses assessment mining, message mining, and feelings. It gives reads up for 4 groupings which are PCs, Streak drives, Mobiles, and Gadgets [1]. Legitimate Bayes' classifier method [2] is being utilized with a definitive goal of assortment. The dataset is from Amazon contains around 600 records. The general accuracy is 0.89. Proposing a language word reference. Pondering assessment [3]-based approach with n grams. Man-made consciousness assessments utilized Inconsistent Backwoods region student with word vector, Choice tree student with report vector, and Sporadic Woods with n-gram are among the choices. The Amazon Product Review dataset was used, and the Lexicon dictionary-based technique was shown to be the most accurate. Tamara Kati recommends that Amazon review sentiment analysis and document representation approaches be compared. [4]. Combining traditional models like bag-of-words, bag-of-n grams, and their TF-IDF variations with linear classifiers like Logistic Regression and SVM. The recurrent neural network Long Short-Term Memory (LSTM) is used. An examination of two AI calculations for assessing the opinion of Amazon item surveys has been proposed [5]. The Naive

Bayes classifier and the Support Vector Machine were used to categories the participants' sentiments into binary groups (SVM). The purpose of this study is to compare SVM with Naive Bayes' classifiers based on statistical measurement to find a better machine learning approach. The accuracy of SVM is higher. Ensemble learning is used for classification of positive, negative, or neutral behavior. Voting is an ensemble machine learning [6] approach that combines five classifiers: Naive Bayes, Support Vector Machines (SVMs), Random Forest, Bagging, and Boosting are some of the techniques used. In the event of employing a unigram and stopping word removal, the Random Forest technique has the maximum accuracy, but the voting algorithm has the best performance in the other circumstances.

Nowadays surprising electronic business battles grant clients to sign in to their area through client's virtual redirection account like twitter, Facebook. This is for overall confirmation of electronic business on obliging affiliations. This lead to disorder of current figured plans to guarantee things to clients who sign in using virtual redirection account on electronic business site. This issue is reliably known as cross site cold where starting situation is either thing or client is new and thought structure have no previous information about them and as such fails to thought about things. Our strategy settles this issue using microblogging relationship of things accumulated from virtual redirection stage, Twitter. Both sound or video overviews and text frames in English and Hindi tongues are considered in commonly thought. Visit with message planning is done on sound or video studies to get nostalgic party. Cross Lingual Liking/Point Model is used to do clever party of Hindi vernaculars studies. Sorting out of thing's parts and client's parts is completely finished the errands of action factorization structure. Our starter results on dataset framed from microblogging stage Twitter moreover, online business page Amazon show that our model on a particularly chief level work on the precision of thought structure in cool starting situation [7].

Quick assessment targets see the examinations of different clients. This paper presents my examination work on the usage of quick appraisal on book outlines. I have applied both free (Semantic Course—Pointwise Shared Data—Data Recovery) and made due (Sponsorship Vector Machine and Fundamental Bayes) robotized thinking framework on two obviously open book outline datasets from Good Reads and Amazon. The near appraisal of the strategies on the datasets shows that free strategy performs better on Good Reads dataset with a precision of 73.23% while directed approach gives other than made results on Amazon dataset with Guiltless Bayes giving the best exactness which goes from 73.72 to 74.73% by sensibility of 5-folds and 10-wrinkles clearly [8].

This paper gives a piece that can be utilized by different on the web shopping stages to annihilate the surveys given by the purchasers, counting nostalgic assessment to remain mindful of astonishing help among their clients. Nostalgic Evaluation is possibly of the most moving appraisal locale in the space of Standard Language Making due. Portrayed as the procedure helps in the evaluation of individuals' feelings, speculations from messages. In this paper, different PC-based information demand assessments have been utilized for tracking down the farthest spot of the examinations. Specifically, relative assessment of evaluations, for instance,

Stochastic Inclination Jump, Picked Fall away from the conviction, Multinomial Pure Bayes, and Sponsorship Vector Machine has been done. Execution appraisal of these assessments has been done on the clarification of the accuracy achieved. The procured results show that the Stochastic Enjoying Fall with Store of Words model beats various assessments and shows the most head accuracy of 88.76% [9].

Appraisal mining is perhaps of the key assignment of typical language making due, which is for the most part called assessment examination, used to see about what individuals have an impression about their affiliations and things in electronic entertainment stages. To further make showing structures utilizing thing studies, strong regions for a approach ought to utilized for expect past what many would think about conceivable. In this evaluation article, a man-made thinking framework called Help Vector with machining (SVM) is utilized to plan a model and this model has been finished on an Internet business application. The information utilized in this study are electronic thing audits which are assembled from Amazon.com. The appraisals of assessment are performed for two degrees obviously of progress: outline level and sentence level. The wild eyed spot of blend of this paper is to introduce a dependable sharp assessment on the thing outlines of electronic business application so the client experience can be upheld [10].

Looking at the strong circumstance it's phenomenally easy to appear around the end that everything turns data. Data is being made at a prominent rate. Not unequivocally how much data being made is heavenly yet other than the rate and assembling are besides accelerating. This correspondingly scrutinizes up a tremendous improvement for advancement of affiliation, clients, and online overviews. More than trillions of outlines of various things are seen as on the web. Earlier thought structures regularly set up adequacy issues. In this paper, we present the thought plan to the clients considering their benefit. Nostalgic appraisal is used to propose considering the top k conflicted with assessments considering top positive plans of a geographically quick. Here we use the Hadoop Improvement to make our development more fit and flexible [11].

Online business is one of the quickly making business areas of current world. At this point a-days individuals buy a great deal of things from web shopping battles. Plans of online things are most frequently study driven. Hence, seeing misleading investigations is getting more significance step by step. Feeling evaluation has amazing significance in counterfeit audit seeing affirmation framework. This paper presents a tendency assessment model that can limit perfect and dubious nostalgic assessments handily. It shows an examination of evaluation dissipating for phony and fair outlines. Besides, the proposed feeling model is utilized to find the effect of probabilistic evaluation score in counterfeit electronic outline confirmation utilizing an inn survey dataset [12].

This paper needs to seclude the diagrams of various motion pictures besides, pick the parts that impacted the assessments of the movies. It also needs to shape an improvement to deal with the client experience. The layouts ought to be a gigantic part in the accomplishment or chaos of a film; a fair propensity assessment model ought to

work with film studies. This paper has destroyed film focuses on using various frameworks like Clashing Forest area district, chose apostatize, Straight SVM, Stochastic grade, additionally, Multinomial NB [13].

Web business is a motorized stage where individuals can trade things on the web. Web business awards the clients to buy things whenever furthermore, from any spot. It moreover gives the purchasers the ability to survey vehemently or unfavorably on anything over the stage. Since individuals see the electronic examinations as a gigantic data on thing, it helps the client in going with choice. Client outlines are their appraisal on something express or alliance. In this paper, the perspective-based evaluation assessment is performed. The systems are pre-made due likewise, sometime later farthest point is settled utilizing feeling assessment. Assessment procedure is finished utilizing fair Bayes (NB) and backing vector machine (SVM) assessment depicts the audits into positive or negative ones. Then, at that point, the arranging is finished by the getting sorted out assessment considering their significance [14].

Thing surveys are basic data enteral focuses for clients as they seek after their buying choices. Regardless, two or three scheming firms use counterfeit correspondents to make lopsided positive examinations to raise their thing and to hurt the thing notorieties of their foes. From the perspective of web-based thing survey stage suppliers, it is fundamental for keep the stage fair and impartial by seeing phony outlines and keeping counterfeit specialists from spreading lopsided surveys. In the consistent review, we attempt to utilize transient and appraisal assessments as prompts to detach counterfeit studies from certifiable thing audits. Guaranteed case information of phony concentrates on in Taiwan was utilized for this normal and feeling assessment. Taking into account the appraisal results, we believe that to be phony correspondents ordinarily made and offered all due appreciation to counterfeit inspectors during common work hours. On the other hand, standard clients in a way made and offered all due appreciation to a little level of traditional thing surveys during work hours. They conveyed and offered all due appreciation to customary thing surveys the most during off-work hours and terminations of the week. In addition, the persistent concentrate moreover uncovered that a large portion of phony specialists addressed others' reactions to their own phony audits no later than in one day or less. The research results revealed that normal and feeling assessments could possibly go about as signs to perceive counterfeit audits and counterfeit specialists. [15].

3 Proposed System

Sentimental analysis of Amazon product reviews using 5 machine learning algorithms, with the use of web scraping for analysis and the choice of best buy product (mobile_phones). The objectives of the paper are; firstly to find all the sentiments of Amazon product reviews, secondly to analyze which machine learning classifier gives the best accuracy for classification and lastly to recommend a best buy product. The proposed system flow chart is shown in Fig. 1.

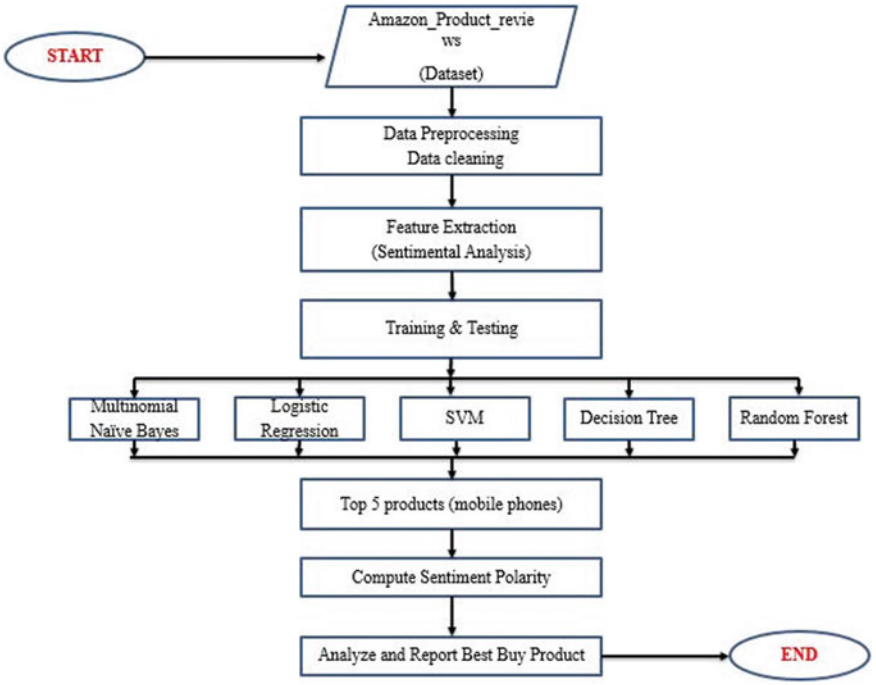


Fig. 1 Flowchart of the proposed system

4 Methodology

Sentimental analysis of Amazon product reviews using 5 machine learning algorithms, with the use of web scraping for analysis and the recommendation of best buy product (mobile_phones). The objectives are firstly to compute the sentiments of all the Amazon_Product Reviews (Positive, Negative & Neutral), secondly to analyze which machine learning classifier gives the best accuracy and finally decide on the basis of sentiment polarity and review ratings to analyze which product is the best to buy.

4.1 Dataset

This dataset has 413,841 data points in total. Each example includes the type, name of the product as well as the text review and the rating of the product. Then, we found that there are some data points which have no ratings when we went through the data. After eliminating those examples, we have 344,509 data points in total. Figure 2, shows that we have 5 classes rating 1 to 5 as well as the distribution among them.

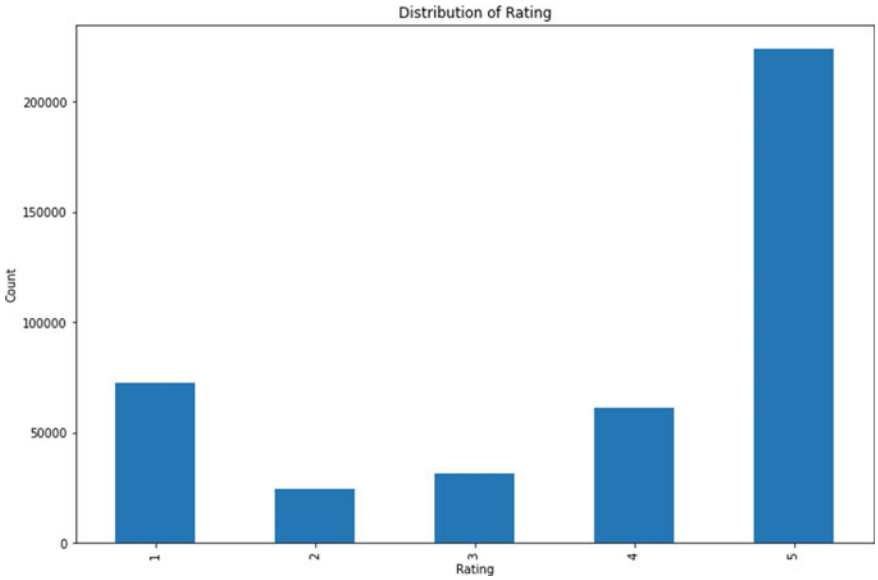


Fig. 2 Distribution of rating

4.2 Workflow

Sentimental Analysis: This is done using bag-of-words (BoW).

Step 1: Raw reviews must be cleaned before they can be used.
Step 2: CountVectorizer TF-IDF vectorizer is used to create BoW.

CountVectorizer which carries out both tokenization and event including in a solitary class. The result is a meager lattice portrayal of a report. **TF-IDF** is a statistical measure that evaluates how relevant a word is to a document in a collection of documents.

Step 3: Convert the message in the survey into mathematical portrayals (highlight vectors).
Step 4: Feature vectors should be fitted to the supervised learning algorithm.

Hence before using machine learning algorithms for classification split the dataset into train and test set.

4.3 Supervised Machine Learning Classification Algorithms

4.3.1 Multinomial Naïve Bayes

We use five classification algorithms here, i.e., Multinomial Naïve Bayes, SVM, Logistic Regression, Random Forest, Decision Tree. Naïve Bayes'—the classifier algorithm is a group of probabilistic algorithms based on Bayes' theorem and the 'naive' assumption of conditional independence between each pair of features. For NLP tasks involving text classification, multinomial Naive Bayes is explicitly utilized.

4.3.2 Support Vector Machine

Support Vector Machine (SVM) is a classification method for both linear and nonlinear data. If the data is arranged in a linear fashion, the SVM looks for the linear optimal separating hyperplane (the linear kernel), which is a decision boundary that divides data into different classes.

4.3.3 Decision Tree

Inner nodes of the Decision Tree classifier were labeled the edges that exited the node were labeled as trial on the data set weight, and the edges that exited the node were marked with features. The leaves of the tree are labeled using categorization. Starting at the bottom of the tree and working your way down through its branches until you reach the top a leaf node is reached, this categorizes the entire document. A Decision Tree classifier is used as an expected model in Decision Tree learning.

4.3.4 Logistic Regression

A logistic regression yields a logistic curve with values only ranging from 0 to 1. The curve is formed differently in logistic regression than in linear regression. Rather than employing the probability, the natural logarithm of the target variable's odds is used. Maximum likelihood estimation (MLE) is used in logistic regression to provide model coefficients that link predictors to the target.

4.3.5 Random Forest

The Random Forest classifier was picked since it beat a solitary choice tree with regards to precision. It's basically a packing-based gathering approach.

5 Results and Analysis

Figure 3 gives a visualization of the review ratings for top 20 brand names in the dataset which are commonly first shown when purchasing mobile phones on Amazon, then refining further to get top 50 products under each brand as shown in Fig. 4.

The confusion matrix and validation accuracy is calculated for the sentiments and plotted. Multinomial Naïve Bayes classification algorithm gives an accuracy score of 85% as shown in Fig. 5, Logistic Regression gives an accuracy score of 88% as shown in Fig. 6, SVM gives an accuracy score of 89% as depicted in Fig. 7, Random Forest classifier gives an accuracy score of 94% as depicted in Fig. 8, lastly Decision Tree gives an accuracy of 92% as depicted in Fig. 9.

Table 1 gives the comparison of accuracy scores for all the five algorithms and Random Forest gives the highest accuracy of all as validated by the confusion matrix too.

Figure 10 shows the sentiments score of all the 5 mobile products which is got by web scraping the reviews of the five phones with the ASIN number.

Figure 11 gives a word cloud of all the 5 products for comparison, it gives a clear visualization as Samsung Galaxy M01 has more positive words than negative so it helps in analyzing and determining that it is the best of all other phones. Hence can be reported as the best buy product in comparison. The Polarity Score for Best Buy Product is computed as shown in the plot in Fig. 12 having ASIN as B08BV2T63X FOR Panasonic Eluga i7, B086KCCMCP for Samsung Galaxy M01, B086KDZGTZ

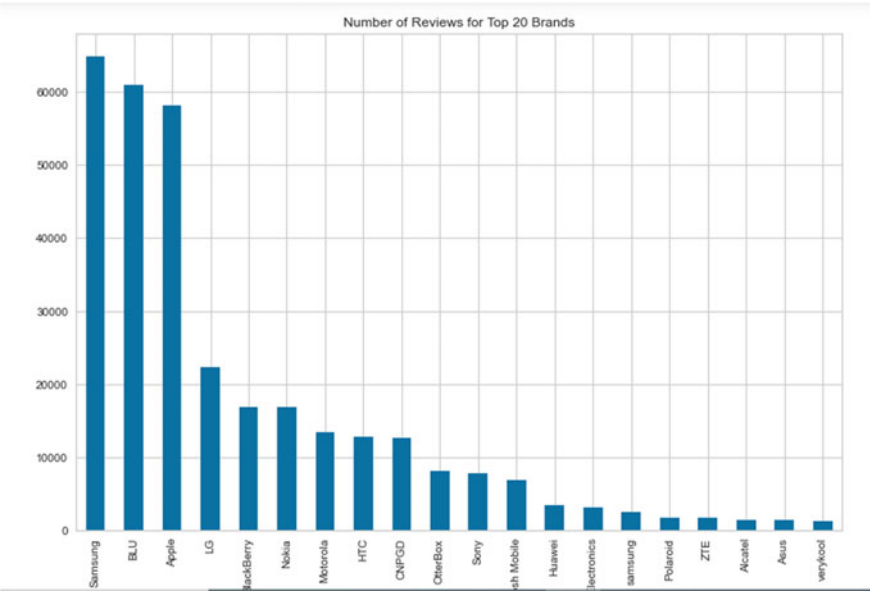


Fig. 3 Plot of Top 20 brand names on Amazon

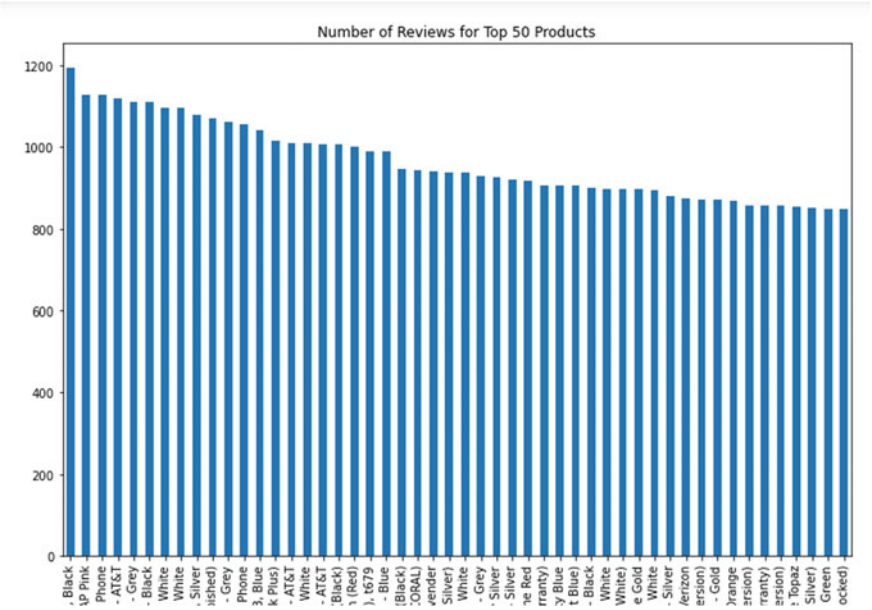


Fig. 4 Plot of Top 50 products on Amazon

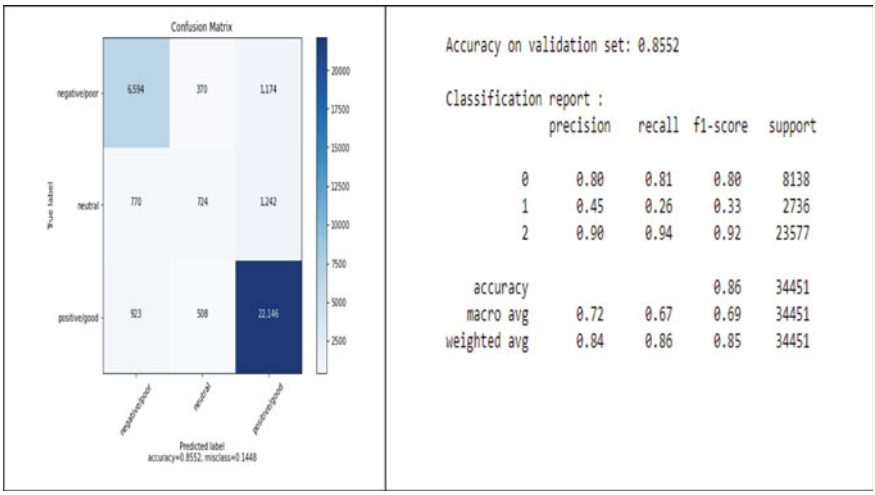
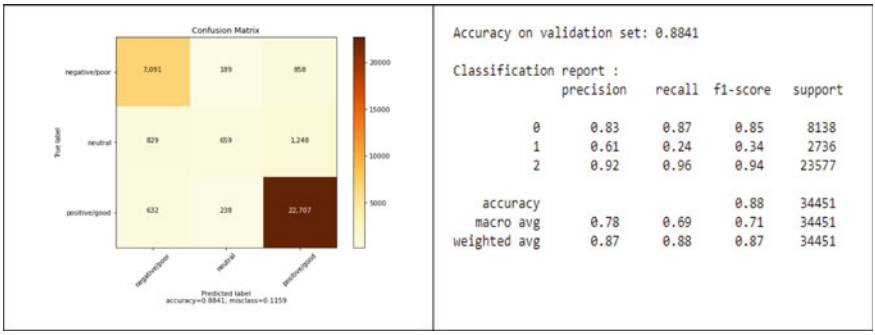


Fig. 5 Confusion matrix and validation accuracy of Multinomial Naïve Bayes classifier

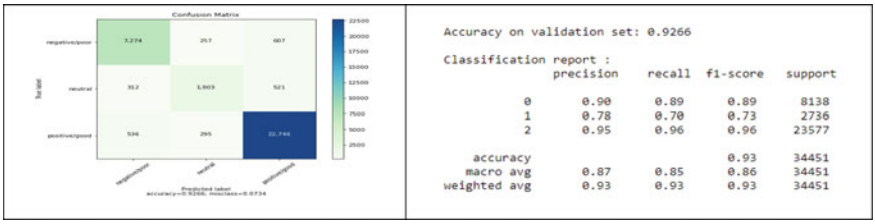


Accuracy on validation set: 0.8841

Classification report :

	precision	recall	f1-score	support
0	0.83	0.87	0.85	8138
1	0.61	0.24	0.34	2736
2	0.92	0.96	0.94	23577
accuracy			0.88	34451
macro avg	0.78	0.69	0.71	34451
weighted avg	0.87	0.88	0.87	34451

Fig. 6 Confusion matrix and validation accuracy of Logistic Regression

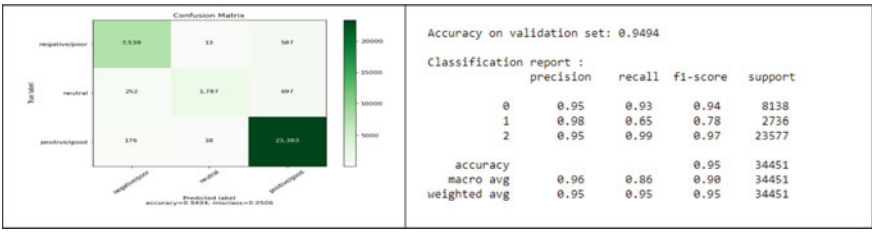


Accuracy on validation set: 0.9266

Classification report :

	precision	recall	f1-score	support
0	0.90	0.89	0.89	8138
1	0.78	0.70	0.73	2736
2	0.95	0.96	0.96	23577
accuracy			0.93	34451
macro avg	0.87	0.85	0.86	34451
weighted avg	0.93	0.93	0.93	34451

Fig. 7 Confusion matrix and validation accuracy of Decision Tree



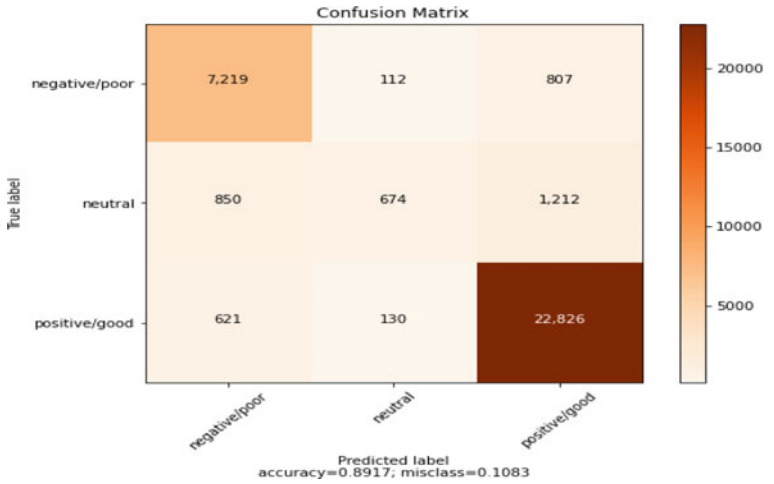
Accuracy on validation set: 0.9494

Classification report :

	precision	recall	f1-score	support
0	0.95	0.93	0.94	8138
1	0.98	0.65	0.78	2736
2	0.95	0.99	0.97	23577
accuracy			0.95	34451
macro avg	0.96	0.86	0.90	34451
weighted avg	0.95	0.95	0.95	34451

Fig. 8 Confusion matrix and validation accuracy of Random Forest

for Vivo Y50, B08GT28WQQ for Nokia, B07PP2K69Z for Oppo A5S. We can analyze the review rating and average sentiment polarity score being highest for **Samsung Galaxy M01,** hence ASIN number **B086KCCMCP** is the best buy product with regard to highest positive polarity score.



Accuracy: 0.8916722301239441
F1 Score: 0.8771983350517051

Fig. 9 Confusion matrix and validation accuracy of SVM

Table 1 Comparison of accuracy

Classification Algorithm	Accuracy
Multinomial Naïve Bayes	0.8552
Logistic Regression	0.8841
Support Vector Machine	0.8917
Random Forest	**0.9494**
Decision Tree	0.9266

```
Processing Sentiments for: Panasonic Eluga i7 (2GB RAM, 16GB Storage, Finger Print Sensor, 4000mAh Battery) (Black)
Average Positive Polarity:0.2
Average Negative Polarity:0.09
Average Neutral Polarity:0.71
Processing Sentiments for: Samsung Galaxy M01 (Black, 3GB RAM, 32GB Storage) with No Cost EMI/Additional Exchange Offers
Average Positive Polarity:0.19
Average Negative Polarity:0.07
Average Neutral Polarity:0.74
Processing Sentiments for: Vivo Y50 (Pearl White, 8GB RAM, 128GB Storage) with No Cost EMI/Additional Exchange Offers
Average Positive Polarity:0.19
Average Negative Polarity:0.17
Average Neutral Polarity:0.64
Processing Sentiments for: Nokia 5.3 Android One Smartphone with Quad Camera, 6 GB RAM and 64 GB Storage - Charcoal
Average Positive Polarity:0.2
Average Negative Polarity:0.06
Average Neutral Polarity:0.74
Processing Sentiments for: OPPO A5S (Black, 3GB RAM, 32GB Storage) With No Cost EMI/Additional Exchange Offers
Average Positive Polarity:0.21
Average Negative Polarity:0.1
Average Neutral Polarity:0.69
```

Fig. 10 Sentiment score of 5 Amazon products

Fig. 11 Word cloud of 5 Amazon products

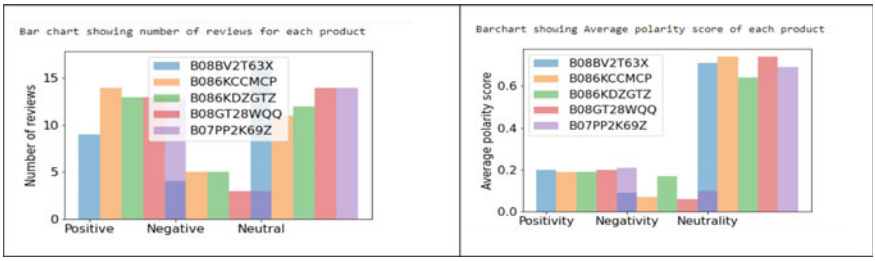


Fig. 12 Sentiment polarity with score

6 Conclusion and Future Scope

Classification of reviews, as well as sentimental analysis, improved the system's accuracy, resulting in more accurate reviews for the user. The practice of classifying texts based on the emotions they contain is known as sentiment analysis. The fundamental methodologies used in an ordinary inclination assessment model are depicted in this survey, which contains three chief advances: data preparation, study assessment, and assessment grouping. Random Forest classifier gives the best accuracy of all the 5 algorithms. From the Word cloud and analysis of sentiment positive polarity percent of Samsung Galaxy M01 is the best to buy product compared to other mobile phones on Amazon. Before long, applications that use assessment examination results are moreover expected to make. Hence for future scope it can be suggested to use this recommendation system to be used as a front end in an html application or to develop an android app. Further we can test classification using Deep Learning techniques to improve accuracy.

References

1. Pankaj R, Pandey P et al (2019) Sentiment analysis on customer feedback data: Amazon product reviews, Manav. In: 2019 International conference on machine learning, big data, cloud and parallel computing (Com-IT-Con), India, 14–16 Feb 2019. Rachna International Institute of Research and Studies, Faridabad, Haryana
2. Surya Prabha PM, Subbulakshmi B (2019) Sentimental analysis using Naive Bayes classifier. In: 2019 International conference on vision towards emerging trends in communication and networking (ViTECoN)
3. Ejaz A, Turabee Z et al (2020) Opinion mining approaches on amazon product reviews: a comparative stu. In: 2020 International conference on computer network, electronic and automation (ICCNEA)
4. Katić T, Milićević N (2018) Comparing sentiment analysis and document representation methods of amazon reviews. In: SISY 2018, IEEE 16th International symposium on intelligent systems and informatics, September 13–15, 2018. Subotica, Serbia
5. Dey S, Wasif S et al (2020) A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews. In: 2020 International conference on contemporary computing and applications (IC3A). Dr. A.P.J. Abdul Kalam Technical University, Lucknow
6. Alrehili A, Albalawi K (2019) Sentimental analysis of customer reviews using ensemble method. ISBN: 978-1-5386-8125-1
7. Kumbhar N, Belerao K (2017) Microblogging reviews based cross-lingual sentimental classification for cold-start product recommendation. In: 2017 International conference on computing, communication, control and automation (ICCUBEA), pp 1–4. https://doi.org/10.1109/ICCUBEA.2017.8463697
8. Kaur VD (2018) Sentimental analysis of book reviews using unsupervised semantic orientation and supervised machine learning approaches. In: 2018 Second International conference on green computing and internet of things (ICGCIoT), pp 519–524. https://doi.org/10.1109/ICGCIoT.2018.8753089
9. Gupta A, Rastogi A, Katal A (2021) A comparative study of Amazon product reviews using sentiment analysis. In: 2021 International conference on advances in computing,

communication, and control (ICAC3), pp 1–6. https://doi.org/10.1109/ICAC353642.2021.9697155
10. Jabbar J, Urooj I, JunSheng W, Azeem N (2019) Real-time sentiment analysis on e-commerce application. In: 2019 IEEE 16th International conference on networking, sensing and control (ICNSC), pp 391–396. https://doi.org/10.1109/ICNSC.2019.8743331
11. Pradhan R, Khandelwal V, Chaturvedi A, Sharma DK (2020) Recommendation system using Lexicon based sentimental analysis with collaborative filtering. In: 2020 International conference on power electronics & IoT applications in renewable energy and its control (PARC), pp 129–132. https://doi.org/10.1109/PARC49193.2020.236571
12. Hassan R, Islam MR (2021) Impact of sentiment analysis in fake online review detection. In: 2021 International conference on information and communication technology for sustainable development (ICICT4SD), pp 21–24. https://doi.org/10.1109/ICICT4SD50815.2021.9396899
13. Kumar S, Sharma K, Veragi D, Juyal A (2022) Sentimental analysis of movie reviews using machine learning algorithms. In: 2022 International conference on machine learning, big data, cloud and parallel computing (COM-IT-CON), pp 526–529. https://doi.org/10.1109/COM-IT-CON54601.2022.9850878
14. Yuvashree E, Preethika S, Nirupama A, Cloudin S (2021) Product aspect ranking using sentimental analysis. In: 2021 International conference on system, computation, automation and networking (ICSCAN), pp 1–4. https://doi.org/10.1109/ICSCAN53069.2021.9526429
15. Wang C-C, Day M-Y, Chen C-C, Liou J-W (2017) Temporal and sentimental analysis of a real case of fake reviews in Taiwan. In: 2017 IEEE/ACM International conference on advances in social networks analysis and mining (ASONAM), pp 729–736