Predicting the Usefulness of E-Commerce Products' Reviews Using Machine Learning Techniques

Dimple Chehal, Parul Gupta, Payal Gulati
Department of Computer Engineering, J.C. Bose University of Science and Technology, YMCA, Faridabad, India
E-mail: dimplechehal@gmail.com, parulgupta_gem@yahoo.com, gulatipayal@yahoo.co.in

Keywords: classification, e-commerce, machine learning, recommender system, usefulness, user reviews

Received: May 5, 2022

User-generated reviews are an essential component of e-commerce platforms. The presence of a large number of these reviews creates an information overload problem, making it difficult for other users to establish their purchase decision. A review voting mechanism, in which users can vote for or against a review, addresses this issue (as helpful or not). The helpful votes on a review reflect its usefulness to other users. As voting on usefulness is optional, not all reviews receive this vote. Furthermore, reviews posted recently by users are not associated with any vote (s). The aim of this paper is to predict the usefulness of user reviews through machine learning techniques. Using the Amazon product review dataset of cell phones, classification models are built on eight features and compared on seven performance measures. As per results, all the classification models performed well, except Linear Discriminant Analysis. The classification performance of Logistic Regression, Decision Tree, Random Forest, AdaBoost, and Gradient Boost was unaffected by feature selection or outlier removal. The performance of Linear Discriminant Analysis improved after feature selection but decreased after outlier removal, whereas ET and KNN classifiers improved in both cases.

Povzetek: Uporaba tehnik strojnega učenja za napovedovanje uporabnih ocen izdelkov e-trgovine.

1 Introduction

Online consumer reviews have evolved for e-commerce users and its stakeholders as an electronic word of mouth (eWoM) [32],[30]. Product reviews comprise of detailed experience of the customer(s) with the product(s). They help the consumers in their purchase decision, indicate any improvement required in the products' quality, thereby helping the business organizations in improving the products' sales. Mining customer reviews through sentiment analysis or topic modeling techniques reveal the customer's inclination towards a product. This helps in building the customer profile and understanding his/her preference for unseen products. Many platforms such as Amazon, Yelp, TripAdvisor, IMDB and Netflix are hosting a large number of user reviews [35]. However, the ever-growing rise in the number of products, customers and product reviews on the e-commerce platform, has led to the information overload problem and has made it infeasible for the customers to browse all the product reviews. To overcome this problem, voting a review as helpful by other customers had been introduced. While the rating of a product depicts a user's experience with a product, the votes gained by a review indicate its usefulness. The solution can be browsing user reviews according to their helpfulness or usefulness. But, due to factors such as humongous volume of electronic word of mouth, voluntary helpfulness voting mechanism, level of visibility and their recentness, all reviews do not receive this vote [5],[27]. Hence, the objective of this study is to categorize the product review according to its usefulness. This will not only help the customers to identify the products as useful or useless even if the review has not gained any votes but can also be fed as input to the recommender system for generating useful recommendations to the users. Also, through this study, the following questions have been answered-

- Which is the most efficient machine learning algorithm for the forecasting the usefulness of a product review?
- Which features should help in determining the usefulness of product review?

The results to the above questions have been obtained with the help of cell phone and accessories dataset taken from Amazon [3]. Eight different machine learning models, namely, Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), AdaBoost (ADA), Gradient Boost (GB), Extra Trees (ET), K Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA), have been trained and tested on existing and derived features and have been evaluated on seven evaluation metrics such as Area under the Curve (AUC), Accuracy (ACC), F1-score (F1) , Precision (P), Recall (R), Mathew's Correlation Coefficient (MCC) and Kappa score [19]. The best model has been fine tuned for prediction of usefulness of review. This study's contributions are stated as follows:

1. Through this research, features such as overall rating, user review, review summary, review

votes, word count of review, character count of review, review's sentiment score, average word count of review have been used to predict a review's usefulness.

2. Along with the already existing features in the Amazon dataset such as overall rating, user review, review summary and review votes, other features have been derived from user review and used in combination as input to the prediction model.
3. This study enables customers to identify useful reviews and e-commerce managers, merchants, retailers to improve the listing of product reviews based on the review usefulness.

The rest of the study is structured as follows: Section 2 consists of the related work, Section 3 details the research methodology followed. While section 4 discusses the result of different experiments conducted on the dataset. Lastly, section 5 concludes by discussing the limitations and future work.

2 Related work

Online user testimonials have gained much-needed prominence in the literature as they instill trust in other potential consumers in the online community [9], [17]. Product reviews can be viewed as a type of passive recommendation process or visibility of user sentiment for their past purchases [12]. Critical management choice for policy-makers is to manage user review to improve customer review efficacy. The academic evidence on review usefulness is largely driven and aided by review hosting platforms, which offer users' opinions on reviews' helpfulness explicitly. For instance on Amazon, customers not only access the rating and text content of each user review, but they also view the number of votes the review obtains from the fellow users and the number of helpful votes [35], [25]. Consumers benefit from informative reviews while making buying decisions. Some customers believe that favorable and unfavorable reviews are useful because such deeply divided records help to validate or invalidate purchase options. Others, on the contrary, find mixed reviews useful because they illustrate both the positives and negatives of the product under consideration. The perceived importance of a review to the end-user is also conveyed through the review's usefulness [28], [18]. This functionality, in particular, makes use of crowd-sourcing to assess the usefulness of reviews [6]. Every review includes the question, "Was this review helpful to you?" Customers who read the reviews may up vote or down vote the review [9], [12].

The research on review usefulness is roughly classified into two categories, prediction-based techniques to ascertain the review's usefulness and understanding of review usefulness. Machine learning classifiers, regression and deep learning approaches have been used to predict review helpfulness in the past [10], [8], [14], [16]. The review length, review timestamp, reviewer's expertise, and manner of writing reviews all have been used previously to predict helpful reviews [5]. Early indicators used to identify review usefulness through review length and review star rating also [24]. Deviation from the mean review length of a product, review's polarity and rating from the same user or on the same item to estimate review helpfulness helped in filtering high-quality ratings thereby improving the collaborative item recommendation process [27]. The moderate-length reviews outperform brief and lengthy ones as review length has inverted-U-shaped impact on usefulness [15]. Further, the more the review matches the language style of the target user, the more it is said to be readable. As a result, it is classified as a domain-specific indicator [22]. The semantic analysis of reviews comprises a wide range of methodologies that make use of structural characteristics like the count of product features cited in a review and its length [34]. The most useful reviews are said to be medium in length, have a lower score, and are negative or neutral in polarity [13]. Both critical evaluations containing data on service failures and favorable reviews highlighting essential product functionalities, technical elements, and aesthetics are seen as beneficial for usefulness prediction [1]. Besides the semantic aspect, neutral polarity reviews are regarded to be also useful [31]. The inclusion of adjectives, status and action verbs, as well as grammatical structure, are vital predictors of helpfulness, particularly when paired with factors such as review age, rating, readability, and subjectivity [21]. Highly readable reviews have proven to be the most beneficial. Based on previously performed emotion-based analysis, it has been concluded that male readers were more inclined to reviews with positive emotions, whereas female readers were more attracted to reviews with negative emotions. Previous findings also indicated that several features such as the review title's polarity, the review's sentiment and polarity, and the cosine similarity between the product review and the product title are contributing factors to determine the usefulness of user reviews [24]. As per the literature review, previous studies are deficient in terms of the combination of natural language processing tools and machine learning techniques for estimation of review usefulness. This study considering the above employs user voting as the target label to build the helpfulness or usefulness prediction system.

This depiction of helpfulness votes differs across platforms. Some platforms show the most helpful votes for a review, whereas others represent the usefulness as the "X of Y" concept. However, in prior methods, a ratio of 0.6 was considered as a helpfulness threshold in the "X of Y" approach of the consumer voting mechanism. Review usefulness, in particular, is critical in product rankings and recommendations [12]. Prediction of review usefulness enables users to compose meaningful reviews that shall assist retailers in intelligent website management by guiding it s users in purchase decisions [24]. The incorporation of a usefulness estimation model can aid in increasing the effectiveness of a collaborative filtering-based recommender system through optimization of data selection for user ratings estimation. This is a great resource for identifying relevant user

Table 1: Comparison of existing studies on identification of useful reviews

S. No	Paper	Model	Dataset	Input features	Performance Metric	Key points	Classification / Regression problem
1.	[20]	MLP, CNN with Trans E	Amazon dataset: CDs, Electronics, Video Games, Books	Product, Review, Reviewer features	Accuracy, F1-score	Dependence solely on hand crafted features leads to poor accuracy. Along with CNN another technique is required for mapping between the reviewer, product and reviews	Regression
2.	[7]	R:LNR, C:Log Reg, Both: DT, RF, GBT, NN	Yelp Shopping reviews	Product, Review, Reviewer features	RMSE, MSE, RAE, RSE, RRSE, MAE,R2 and CC (R), Accuracy, AUC, Precision, Recall, and F1 score (C)	Authors examine the impact of friends on review usefulness by introducing social network features. For classification, reviews receiving more than 3 votes are marked as helpful, 0 votes as unhelpful and discarded otherwise	Classification, Regression
3.	[11]	MLP, CNN	SiteJabber.com, ConsumerAffairs.com (DomainsDating, Wedding Dresses, Marketplace, Car Insurance, Travel Agencies, Mortgages)	Review features	Accuracy	Adjacent or neighbor reviews impact a user's helpfulness perception of a review. For classification, reviews receiving more than 2 helpful votes labeled as helpful and unhelpful otherwise.	Classification
4.	[27]	Linear Support Vector Regression, RF Regression	Yelp hotel stores reviews, Yelp food stores reviews	Review features	Pearson and Spearman correlation values	Deviations in star ratings, review's length and review's polarity with respect to user and item impact usefulness. Authors do not consider reviewer features. Random Forest was a better helpfulness predictor. Integration of such an estimation model improves the CF system performance.	Regression
5.	[24]	Multivariate adaptive regression, 'C' and 'R' tree, RF, NN, deep NN	Amazon multidomain sentiment analysis dataset	Review, Reviewer, Product features	MSE, RMSE,RRSE	Review type characteristics standout as effective characteristics to determine review's helpfulness as compared to reviewer and product characteristics. Combining all three characteristics yield best performance.	Regression
6.	[2]	DT, RF	Amazon Product dataset (Books, Office Products)	Review, Reviewer features	Accuracy, F-measure	Helpfulness threshold ratio set to value of 0.6. Features such as text, reviewer, and readability perform better than summary features. RF performed better than decision trees	Classification
7.	[26]	MLP, CART	Contributed dataset of	Review, Reviewer,	MSE, RAE, RMSE,	More the comments, polarity and sentiments in a review, more are the	Regression

		, Multivariate adaptive regression, Generalized Linear model, Ensemble model	34 product categories from Amazon.com	Product features	RRSE, MAE	helpful votes. Reviews with at least 10 votes are selected. Best results were obtained using hybrid features with ensemble model performing the best	
8.	[34]	RF	Dataset from JD.com	Review features, informativeness, length	Accuracy, AUC	Classification threshold for search and experience products to be different. Threshold of 4 for search products such as electronics and 2 for experience products such as skin gave the best model performance	Classification

reviews for decision-making [27]. Table 1 lists the key takeaway points from the existing literature.

In Table 1, AUC stands for Area Under the Curve, 'C'-Classification, CNN-Convolutional Neural Network, CC-Correlation Coefficient, DT-Decision Tree, GBT-Gradient Boosted Tree, Log Reg-Logistic Regression R-Regression, RAE-Relative Absolute Error, RF-Random Forest, RMSE-Root Mean Square Error, R2 -R Squared, RSE-Relative Squared Error, RRSE-Root Relative Squared Error, MAE-Mean Absolute Error, MLP-Multi Linear Perceptron, MSE-Mean Squared Error and NN-Neural Network.

3 Research methodology

The review-based recommendation methods in the studied literature utilize review contents and do not incorporate the associated helpfulness or usefulness scores. Incorporating this information of reviews helps in better exploitation of the user reviews. Since, several reviews don't have helpfulness scores, it is essential to predict the usefulness of these reviews [16]. The steps undertaken as part of prediction of usefulness of user reviews are shown in Figure 1 and are as follows:

3.1 Data collection

Data collection and its processing are the initial steps in all the machine learning methods [36]. Amazon cell phone and accessories dataset has been considered for this task [29] [3]. As shown in Table 2, the dataset with `(1048572, 12)` size has the following columns:

{
"reviewerID": "`A62MUEQU8I52E`*",*
"asin": "`B007SJZUSI`*",*
"reviewerName": " H. Moyer *",*
"vote": 3,
"style": { 'Color:': ' Gold' },
"reviewText": "Not a huge capacity power bank but a very good capacity for its very compact size. Exactly what I need, to have with me all of the time, just in case. One micro USB power input port for charging it, and one standard USB port for charging another device, either one using the same most standard cable in the industry. For most of us, power banks are for emergency only, so multiple output ports just add size unnecessarily. One state of charge gauge with 4 LED indicator lights, and one pushbutton. Very simple.",
"overall": 5.0,
"summary": " SIMPLE, COMPACT, AND POWERFUL FOR ITS SIZE ",
"unixReviewTime": `1490659200.0`*,*
"reviewTime": " 03 28, 2017 ",
"image": nan,
"verified": True
}

Table 2: Dataset description

Column name	Column description
reviewerID	Specifies the reviewer's unique identifier e.g. A284QS51P9P9V1
asin	Specifies the product's unique identifier e.g. B00UVSNVHA
reviewerName	Represents the name of the user/reviewer
vote	Represents the count of helpful votes received by a review
style	Represents a dictionary of the product metadata
reviewText	Implies the text contained in the review
overall	Represents the star rating given to a product
summary	Represents the textual summary of a product review
unixReviewTime	Represents the time at which review was generated (unix time)
reviewTime	Represents the time at which review was generated (raw)
image	Represents the product images that users post when they review the product

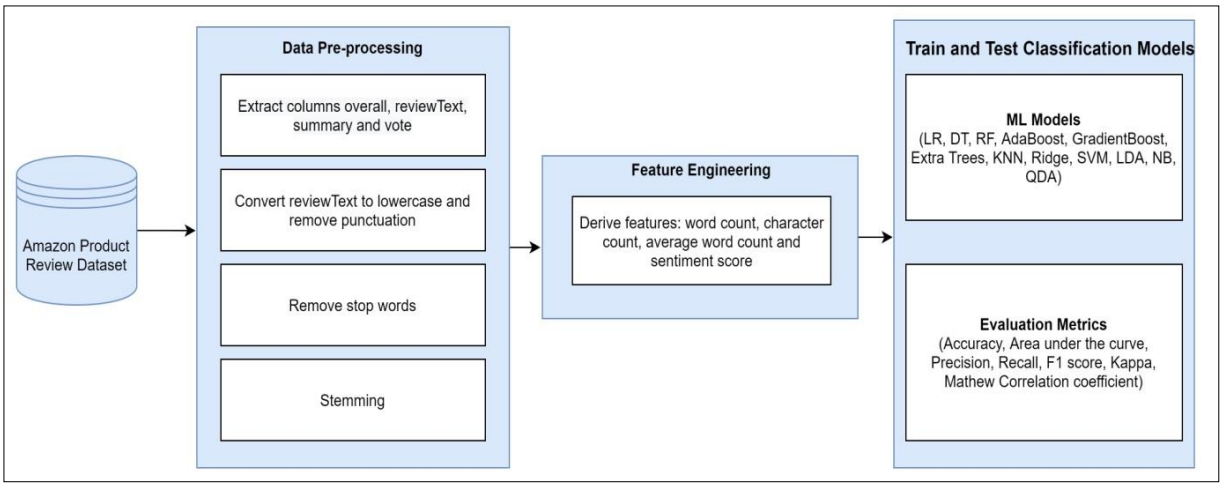

Figure 1: Research method

3.2 Pre-processing

The dataset consists of 12 columns and 1048572 rows. In order to categorize the reviews as useful or useless the following pre-processing steps have been undertaken:

1. Out of the 12 columns available, only columns- *overall*, *reviewText*, *summary* and *vote* have been utilized.
2. Next, *reviewText* column has been converted to lower case and punctuation has been removed.
3. After performing the below mentioned feature engineering steps, stop words using Python's *nltk* library have been removed.
4. Step 3 has been followed by stemming process in which porter stemmer has been used to apply stemming on *reviewText* column.

3.3 Feature engineering

Apart from the columns considered during the pre-processing phase, below mentioned columns have been derived:

1. *Word count*: This column represents number of words in a review
2. *Char count*: This column indicates number of characters in a review
3. *Avg word count*: This column stands for average word length of a review
4. *Sentiment score*: This column represents polarity of a review ranging from minus one (indicating extremely negative) to plus one (indicating extremely positive) which has been determined with the help of Python's *vaderSentiment* library

3.4 Preliminary analysis

1. The top ten most frequently occurring words, as shown in Table 3, after removal of stop words from the dataset are given below:

Table 3 Top ten frequently occurring words

Word	Frequency
Phone	165691
case	117779
one	62104
screen	57831
like	51122
use	43841
great	39611
battery	39595
would	38616
good	37078

As seen in Table 3, as the dataset is related to cell phones, the top ten frequently occurring words are related to this domain. The users have provided reviews mostly related to phone, case, screen and battery. To obtain these words, the frequency of words in the user reviews was obtained and then the top ten words were extracted.

2. The ten least frequently occurring words in the dataset, with only single occurrence are- Performancebattery, gummybearlike, amazonsunvalleytek, knive , terd, hh, nomy, 4siphone, Loosey, caseseems
3. The percentage of overall rating provided by users is provided in Table 4. The review dataset contains the majority of user reviews with the highest rating of the product, that is, 5, followed by user rating 4. The dataset contains more one-

star ratings than compared to three-star and two-star ratings.

Table 4 Distribution of user rating in the dataset

Rating	Count	Percentage (%)
5	49894	55.02
4	15243	16.81
3	8094	8.93
2	5509	6.08
1	11942	13.17

Supervised learning algorithms require input and output examples for training the model. In order to predict the review usefulness, the target column has been contributed which identifies each review as useful or not. To help the classification models learn if a review is useful or useless all the reviews with more than 10 votes have been marked as useful else useless.

3.5 Model selection

Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), AdaBoost (ADA), Gradient Boost (GB), Extra Trees (ET), k Nearest Neighbor (KNN) and Linear Discriminant Analysis (LDA) were used to categorize the usefulness of user reviews [23], [4]. All the models have been implemented in Python using *pycaret* library.

3.6 Data setup

Classification estimators were used in this study to predict the user review's usefulness. The target type is binary, with two possible values as useful or useless. The data has been partitioned into 70:30 partitions to obtain the training and testing sets. To allow row shuffling during the train-test split, the data split shuffle was set to true. The predictive models' performance was evaluated using stratified ten-fold cross-validation.

4 Result and discussion

Usefulness is treated as dependent variable and overall, reviewText, summary, vote, word count, character count, average word length and sentiment score are treated as independent variables. The model's performance can be assessed using a variety of evaluators, some of which are more appropriate than others. The models have been assessed in terms of accuracy (calculated using (1)), area under the curve, precision (calculated using (2)), recall (calculated using (3)), f1-score (calculated using (4)), kappa score (calculated using (5)) and MCC (calculated using (6)) as shown in Table 5. The Table also displays the time taken (TT) in seconds for the models to be trained.

- *Accuracy:* It is the most widely used performance metric and is calculated as the number of correct predictions over all predictions [33].

$$Accuracy = \left(\frac{TP + TN}{TP+TN+FP+FN}\right) * 100 \quad (1)$$

Where, TP stands for true positive, TN stands for true negative, FP stands for false positive and FN stands for false negative.

- *Area under the Curve:* The plot of sensitivity versus (1-specificity) is given by Receiver Operating Characteristic curve. AUC converts the curve to a numeric value. The ranges of the curve and their corresponding interpretations are grouped as excellent for range varying from 1 to 0.90; good from 0.90 to 0.80; fair from 0.80 to 0.70; poor from 0.70 to 0.60 and fail from 0.60 to 0.50.

- *Precision:*

$$Precision = TP/TP + FP \quad (2)$$

- *Sensitivity:* Sensitivity is the ratio of actually true classes that are identified correctly. Another name for sensitivity is true positive rate or recall. To reframe, it measures how often true predictions are correct.

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

- *F1 Score:* It's an accuracy metric that considers the trade-off between precision and recall.

$$F1\ Score = 2 * \frac{(Precision*Recall)}{Precision+Recall} \quad (4)$$

- *Kappa:* The Kappa score handles multi-class as well as imbalanced class problems.

$$Kappa = p_o - p_e/1 - p_e \quad (5)$$

Where, p_o and p_e denote the observed and expected agreement, respectively. In general, it reflects how a classifier performs as compared to another classifier that simply guesses at random based on each class's frequency. Cohen's kappa is never greater than 1. When the value of kappa is zero, the classifier is useless.

- *Matthews Correlation Coefficient (MCC):* The Matthews correlation coefficient assesses the quality of a binary classification problem; it is a balanced measure for unbalanced dataset as well. It outputs a value between minus one and plus one where, plus one indicates complete agreement between predicted and observed value, minus one indicates total disagreement, and zero value indicates random predicted values [33].

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (6)$$

Table 5: Performance of ML models

Model	Accuracy	AUC	Recall	Precision	F1-Score	Kappa	MCC	TT (sec)
LR	0.99	0.99	0.99	0.99	0.99	0.99	0.99	11.5
DT	0.99	0.99	0.99	0.99	0.99	0.99	0.99	**0.19**
RF	0.99	0.99	0.99	0.99	0.99	0.99	0.99	2.47
ADA	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.23
GB	0.99	0.99	0.99	0.99	0.99	0.99	0.99	6.48
ET	0.9657	0.9995	0.9997	0.9612	0.98	0.8596	0.8683	9.14
KNN	0.9263	0.9471	0.9912	0.9263	0.9576	0.6762	0.7004	1.92
LDA	0.5788	0.517	0.6078	0.6427	0.6233	0.3348	0.3439	27.73

As shown in Table 5 and Figure 2, most of the classification models are performing decently when contrasted according to the evaluation parameters. In order to test the model's robustness, ten-fold cross-validation has been employed. Due to lack of sufficient system RAM, the model has been fed a random sample of 5000 rows, leading to the above performance. Also, the methods' black-box state diminishes the results' interpretability. In comparison to others, LDA is unable to provide a reasonable prediction. The models have been trained again after performing feature selection and outlier removal to check the improvement in their performance. The near perfect performance of these models can be attributed to the size of data being fed to these models. Decision Tree model takes the least amount of time i.e. 0.19 seconds for generating the above results.

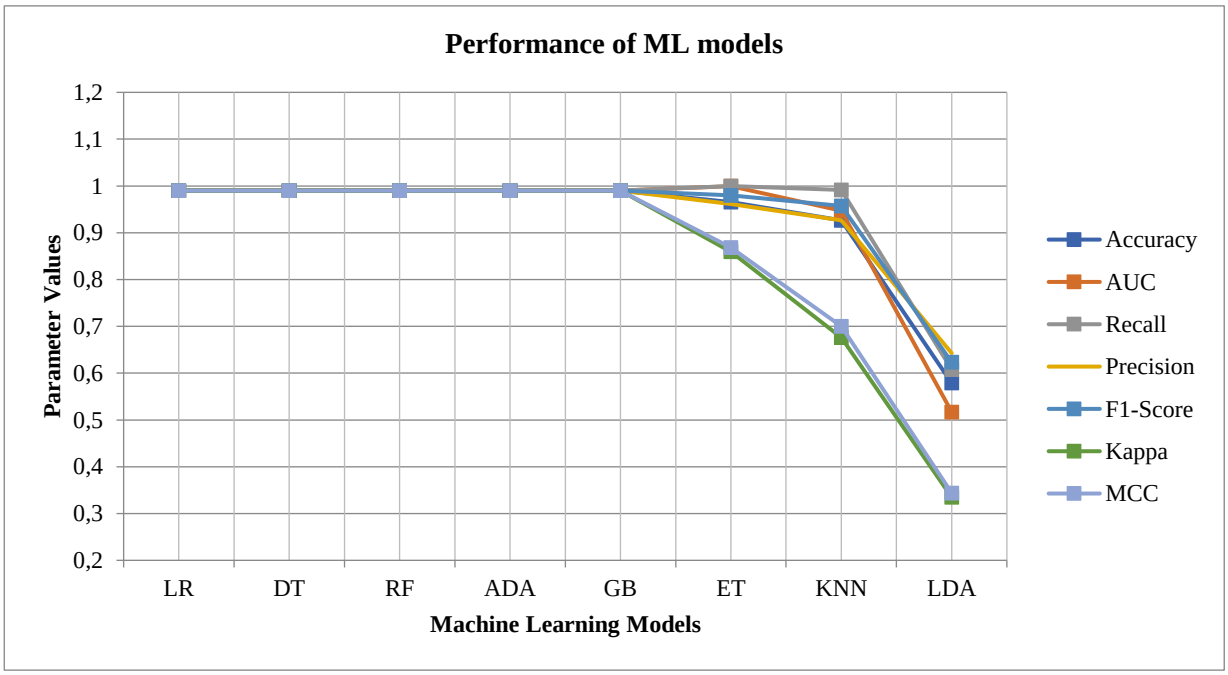

Figure 2: Performance of ML models

In order to improve the performance and reduce the training time of the above models, feature selection has been performed. Upon performing feature selection, the accuracy of LDA model jumps to 0.8411, AUC increases to 0.732, recall, precision, f1-score, kappa and MCC turn out to be 0.892, 0.842, 0.866, 0.638 and 0.658 respectively. The threshold value used for feature selection is set to 0.8 and the classic method of permutation feature importance techniques is used. Even after performing feature selection, the performance of LR, DT, RF, ADA, and GB classifiers remains unaffected as shown in Table 6.

As seen from Table 5 and Table 6, the training time of all the models reduced. Training time of model- LR reduced to 6.32 from 11.5 (without feature selection), DT remained the same as 0.19, ADA classifier remained the same as 0.23, ET reduced to 9.05 from 9.14, KNN reduced to 1.90 from 1.92 and LDA reduced to 26.75 from 27.73 seconds. Only two models RF and GB had their training time increased to 2.62 from 2.47 and 6.51 from 6.48 respectively. This increase

Table 6: Performance of ML models after feature selection

Model	Accuracy	AUC	Recall	Precision	F1-Score	Kappa	MCC	TT (sec)
LR	0.99	0.99	0.99	0.99	0.99	0.99	0.99	6.32
DT	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.19
RF	0.99	0.99	0.99	0.99	0.99	0.99	0.99	2.62
ADA	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.23
GB	0.99	0.99	0.99	0.99	0.99	0.99	0.99	6.51
ET	0.9634	0.99	0.99	0.959	0.979	0.848	0.858	9.05
KNN	0.9729	0.991	0.996	0.973	0.984	0.892	0.895	1.90
LDA	0.8411	0.732	0.892	0.842	0.866	0.638	0.658	26.75

Table 7 represents performance of classifiers after removal of outliers. Outliers from the training data have been reduced using the Singular Value Decomposition and the outlier threshold has been set to 0.05, that is, five percent of the outliers have been removed from the training dataset. Again, the performance of LR, DT, RF, ADA, and GB classifiers remained unaffected. While the accuracy of ET and KNN classifiers increased, that of LDA decreased significantly. This implies that ET, KNN and LDA classifiers are affected due to removal of outliers whereas the rest of the classifiers are not affected with this processing step.

Table 7: Performance of ML models after outlier removal

Model	Accuracy	AUC	Recall	Precision	F1-Score	Kappa	MCC	TT (sec)
LR	0.99	0.99	0.99	0.99	0.99	0.99	0.99	6.25
DT	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.18
RF	0.99	0.99	0.99	0.99	0.99	0.99	0.99	2.37
ADA	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.22
GB	0.99	0.99	0.99	0.99	0.99	0.99	0.99	6.42
ET	0.9759	0.9999	0.99	0.9726	0.9861	0.8969	0.9019	6.43
KNN	0.9801	0.9938	0.9979	0.9793	0.9885	0.9168	0.9192	1.83
LDA	0.7523	0.7173	0.7673	0.8541	0.806	0.4644	0.4877	23.48

As shown in Table 5, Table 6 and Table 7, LR, DT, RF, ADA and GB are performing perfectly for the sample dataset provided to the models with and without feature selection and outlier removal process. LDA model showed performance improvement after feature selection process, but degradation after outlier removal and accuracy of ET and KNN models improved after removal of outliers.

5 Limitations and future work

In this study, review usefulness prediction models were built and compared using collected features from the publicly available Amazon's cell phone and accessories dataset such as overall, reviewtext, summary, and vote, as well as derived features such as word count, character count, average word count, and sentiment score. Seven different performance measures namely, accuracy, area under the curve, precision, recall, f1-score, Kappa score and MCC were used to compare eight machine learning models- Logistic Regression, Decision Tree, Random Forest, AdaBoost, Gradient Boost, Extra Trees, K nearest Neighbor and Linear Discriminant Analysis. All the classification models performed well except LDA. Feature selection and outlier removal techniques had no effect on the classification performance of Logistic Regression, Decision Tree, Random Forest, AdaBoost, and Gradient Boost. The performance of LDA improved after feature selection but decreased after outlier removal, whereas ET and KNN depicted improvement in both cases. The results of this research can assist e-commerce platforms in gaining more clarity of the usefulness of online reviews. They can automatically analyze the usefulness of product reviews by utilizing prediction models as stated above. A system that uses ML models to predict useful reviews will benefit all stakeholders, including end users, product owners, and e-commerce platform regulators. In cases where the review has received no votes from people, such a system would be beneficial. In that instance, stakeholders might utilize the models' predictions to find reviews of interest or usefulness. This would ultimately save a significant amount of time spent reviewing the enormous number of available user reviews. This study was limited due to lack of sufficient system RAM; the models were fed a random sample of 5000 rows. Also, the methods' black-box state diminishes the results' interpretability. The study can be strengthened by improving the prediction models by removing fake reviews, incorporating emoticons for online review helpfulness prediction, employing unsupervised learning techniques instead of supervised learning, and developing deep learning models.

Availability of data

The dataset is available through URL: https://jmcauley.ucsd.edu/data/amazon/

References

[1] Ahmad, S.N. and Laroche, M. 2017. Analyzing electronic word of mouth: A social commerce

construct. *International Journal of Information Management*. 37, 3 (Jun. 2017), 202–213. DOI:https://doi.org/10.1016/J.IJINFOMGT.2016.08.004.

[2] Akbarabadi, M. and Hosseini, M. 2020. Predicting the helpfulness of online customer reviews: The role of title features. *International Journal of Market Research*. 62, 3 (2020), 272–287. DOI:https://doi.org/10.1177/1470785318819979.

[3] Amazon Review Data: 2018. *https://jmcauley.ucsd.edu/data/amazon/*. Accessed: 2021-05-14.

[4] Ampomah, E.K. et al. 2021. Stock market decision support modeling with tree-based AdaBoost ensemble machine learning models. *Informatica*. 44, 4 (Mar. 2021), 477–489. DOI:https://doi.org/10.31449/inf.v44i4.3159.

[5] Arif, M. et al. 2019. A Survey of Customer Review Helpfulness Prediction Techniques. *Advances in Intelligent Systems and Computing*. Springer International Publishing. 215–226.

[6] Bilal, M. et al. 2019. Profiling and predicting the cumulative helpfulness (Quality) of crowd-sourced reviews. *Information (Switzerland)*.

[7] Bilal, M. et al. 2021. Profiling reviewers' social network strength and predicting the "Helpfulness" of online customer reviews. *Electronic Commerce Research and Applications*. 45, (Jan. 2021), 101026. DOI:https://doi.org/10.1016/j.elerap.2020.101026.

[8] Chen, C. et al. 2019. Multi-domain gated CNn for review helpfulness prediction. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*. (2019), 2630–2636. DOI:https://doi.org/10.1145/3308558.3313587.

[9] Chua, A.Y.K. and Banerjee, S. 2016. Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality. *Computers in Human Behavior*. 54, (2016), 547–554. DOI:https://doi.org/10.1016/j.chb.2015.08.057.

[10] Dey, D. and Kumar, P. 2019. A novel approach to identify the determinants of online review helpfulness and predict the helpfulness score across product categories. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer. 365–388.

[11] Du, J. et al. 2021. Neighbor-aware review helpfulness prediction. *Decision Support Systems*. April (2021), 113581. DOI:https://doi.org/10.1016/j.dss.2021.113581.

[12] Enamul Haque, M. et al. 2018. Helpfulness prediction of online product reviews. *Proceedings of the ACM Symposium on Document Engineering 2018, DocEng 2018*. (2018). DOI:https://doi.org/10.1145/3209280.3229105.

[13] Eslami, S.P. et al. 2018. Which online reviews do consumers find most helpful? A multi-method investigation. *Decision Support Systems*. 113, (Sep. 2018), 32–42. DOI:https://doi.org/10.1016/J.DSS.2018.06.012.

[14] Fan, M. et al. 2019. Product-aware helpfulness prediction of online reviews. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*. 2, Ccl (2019), 2715–2721. DOI:https://doi.org/10.1145/3308558.3313523.

[15] Fink, L. et al. 2018. Longer online reviews are not necessarily better. *International Journal of Information Management*. 39, (Apr. 2018), 30–37. DOI:https://doi.org/10.1016/J.IJINFOMGT.2017.11.002.

[16] Ge, S. et al. 2019. Helpfulness-aware review based neural recommendation. *CCF Transactions on Pervasive Computing and Interaction*. 1, 4 (Dec. 2019), 285–295. DOI:https://doi.org/10.1007/s42486-019-00023-0.

[17] Hamad et al. 2018. Review helpfulness as a function of Linguistic Indicators. *IJCSNS International Journal of Computer Science and Network Security*. 18, 1 (2018), 234–240.

[18] Hong, H. et al. 2017. Understanding the determinants of online review helpfulness: A meta-analytic investigation. *Decision Support Systems*. 102, (2017), 1–11. DOI:https://doi.org/10.1016/j.dss.2017.06.007.

[19] Kaddoura, S. et al. 2022. A systematic review on machine learning models for online learning and examination systems. *PeerJ Computer Science*. 8, (May 2022), e986. DOI:https://doi.org/10.7717/PEERJ-CS.986.

[20] Kong, L. et al. 2022. Predicting Product Review Helpfulness - A Hybrid Method. *IEEE Transactions on Services Computing*. 15, 4 (2022), 2213–2225. DOI:https://doi.org/10.1109/TSC.2020.3041095.

[21] Krishnamoorthy, S. 2015. Linguistic features for review helpfulness prediction. *Expert Systems with Applications*. 42, 7 (May 2015), 3751–3759. DOI:https://doi.org/10.1016/J.ESWA.2014.12.044.

[22] Liu, A.X. et al. 2019. It's Not Just What You Say, But How You Say It: The Effect of Language Style Matching on Perceived Quality of Consumer Reviews. *Journal of Interactive Marketing*. 46, (May 2019), 70–86. DOI:https://doi.org/10.1016/J.INTMAR.2018.11.001.

[23] Luo, Y. and Xu, X. 2019. Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of yelp. *Sustainability (Switzerland)*. 11, 19 (2019). DOI:https://doi.org/10.3390/su11195254.

[24] Malik, M.S.I. 2020. Predicting users' review helpfulness: the role of significant review and

reviewer characteristics. *Soft Computing*. 24, 18 (Sep. 2020), 13913–13928. DOI:https://doi.org/10.1007/s00500-020-04767-1.

[25] Malik, M.S.I. and Hussain, A. 2018. An analysis of review content and reviewer variables that contribute to review helpfulness. *Information Processing and Management*. 54, 1 (2018), 88–104. DOI:https://doi.org/10.1016/j.ipm.2017.09.004.

[26] Malik, M.S.I. and Hussain, A. 2020. Exploring the influential reviewer, review and product determinants for review helpfulness. *Artificial Intelligence Review*. 53, 1 (2020), 407–427. DOI:https://doi.org/10.1007/s10462-018-9662-y.

[27] Mauro, N. et al. 2021. User and item-aware estimation of review helpfulness. *Information Processing and Management*.

[28] Mitra, S. and Jenamani, M. 2021. Helpfulness of online consumer reviews: A multi-perspective approach. *Information Processing and Management*. 58, 3 (2021), 102538. DOI:https://doi.org/10.1016/j.ipm.2021.102538.

[29] Ni, J. et al. 2020. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. (2020), 188–197. DOI:https://doi.org/10.18653/v1/d19-1018.

[30] Orimaye, S.O. et al. 2016. Learning Sentiment Dependent Bayesian Network Classifier for Online Product Reviews. *Informatica (Slovenia)*. 40, 2 (2016), 225–235.

[31] Salehan, M. and Kim, D.J. 2016. Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems*. 81, (Jan. 2016), 30–40. DOI:https://doi.org/10.1016/J.DSS.2015.10.006.

[32] Saumya, S. et al. 2018. Ranking online consumer reviews. *Electronic Commerce Research and Applications*. 29, (2018), 78–89. DOI:https://doi.org/10.1016/j.elerap.2018.03.008.

[33] Sidhu, R.K. et al. 2020. Machine learning based crop water demand forecasting using minimum climatological data. *Multimedia Tools and Applications*. 79, 19–20 (2020), 13109–13124. DOI:https://doi.org/10.1007/s11042-019-08533-w.

[34] Sun, X. et al. 2019. Helpfulness of online reviews: Examining review informativeness and classification thresholds by search products and experience products. *Decision Support Systems*. 124, (Sep. 2019), 113099. DOI:https://doi.org/10.1016/J.DSS.2019.113099.

[35] Wu, J. 2017. Review popularity and review helpfulness: A model for user review effectiveness. *Decision Support Systems*. 97, (2017), 92–103. DOI:https://doi.org/10.1016/j.dss.2017.03.008.

[36] Yenkikar, A. et al. 2022. Semantic relational machine learning model for sentiment analysis using cascade feature selection and heterogeneous classifier ensemble. *PeerJ Computer Science*. 8, (Sep. 2022), e1100. DOI:https://doi.org/10.7717/PEERJ-CS.1100.