KS Knowledge Engineering and Data Science (KEDS)
Vol 5, No 1, December 2022, pp. 101–108

pISSN 2597-4602
eISSN 2597-4637

Sentiment Analysis of Amazon Product Reviews using Supervised Machine Learning Techniques

Naveed Sultan *

Department of Information Technology, Khwaja Fareed University of Engineering and Information Technology, Abu Dhabi Rd, Rahim Yar Khan, Punjab, Pakistan

*naveedsultan587@gmail.com **

** corresponding author*

ARTICLE INFO

Article history:
Received 3 June 2022
Revised 10 July 2022
Accepted 14 August 2022
Published online 7 November 2022

Keywords:
Supervised machine learning
Random Forest Classification
Decision Tree
Support Vector Machine
K-Nearest Neighbor classification

ABSTRACT

Today, everything is sold online, and many individuals can post reviews about different products to show feedback. Serves as feedback for businesses regarding buyer reviews, performance, product quality, and seller service. The project focuses on buyer opinions based on Mobile Phone reviews. Sentiment analysis is the function of analyzing all these data, obtaining opinions about these products and services that classify them as positive, negative, or neutral. This insight can help companies improve their products and help potential buyers make the right decisions. Once the preprocessing is classified on a trained dataset, these reviews must be preprocessed to remove unwanted data such as stop words, verbs, pos tagging, punctuation, and attachments. Many techniques are present to perform such tasks, but in this article, we will use a model that will use different inspection machine techniques.

I. Introduction

People buy goods from various e-commerce websites as the world's commercial sites are practically online [1]. It is also a privileged condition where products are checked before the purchase. Consumers are more likely to buy a product through reviews. Internet retailers and distributors invite clients to express their thoughts on their merchandise. Millions of feedback on products, facilities, and places are produced daily online [2]. This makes the internet the primary source of a product or service's knowledge. Reviews, therefore, offer valuable feedback on a business, including its venue, pricing, and advice, allowing customers to consider every part of the business [3]. This is positive for consumers and encourages marketers to understand shoppers and their preferences that render their products.

When a company's amount of comments available rises, it gets more challenging for a potential consumer to decide whether or not to purchase it [4]. In this age of artificial intelligence, it takes time to polarize a sample into unique categories to read thousands of reviews and recognize a brand to consider its attractiveness among customers worldwide [5][6]. Today, studying data from actual customer reviews is an important field.

The author in [7] has worked in film reviews. Since vast repositories of online reviews are readily accessible, this domain is easy to work on. Also, with a machine-extractable ranking metric such as several ratings, reviewers usually summarize their overall sentiment, but they did not hand-label the data for implementing supervised learning and assessment. The Internet Movie Database (IMDb) is their database root, where the database includes only numeric values or scores. Ratings are collected randomly and grouped into three categories: positive, negative, or neutral. They focused only on finding the tendency of the emotion to be either positive or negative. The following three Naïve Bayes machine learning algorithms were used: Maximum Entropy Classification and Help Vector Machinery (SVM).

https://doi.org/10.17977/um018v5i12022p101-108

There is an emphasis in [8]. This is the definitive Flipkart feedback study using algorithms from the Bayes Naïve and Decision Tree. Using the product ratings and reviews of the single data set of Flipkart sellers and its classification, the subjectivity and objectivity, and that the buyer is negative to the positive meaning of the term. These assessments were, to a certain degree, positive and prospective both for your purchasers and for your providers. It is an observational research analyzing the efficacy of the semantic significance of the product evaluation categorization.

In [9], feedback from numerous e-shopping websites is evaluated. Analyzing ratings for online shopping sites is the primary goal of the framework. The ratings are categorized according to positive, negative, and neutral. Such findings help pick a specific e-shopping website based on the highest favorable reviews and scores. Firstly, the data collection of e-shopping websites providing ratings relevant to the services of individual websites is gathered. Then, add specific preprocessing methods to datasets to delete unwanted items and organize details correctly. After that, we use the POS tagger to assign tags according to the position of each phrase. To find the Score of each word, "sentiwordnet dictionary" is used. Sentiments then Positive, negative, and neutral are graded. In the graphical style, the comparison of the providers based on positive and negative feedback can be seen.

This paper aims to distinguish customers' positive and negative feedback of various products and develop a supervised learning model to polarize large quantities of reviews. Our dataset consists of feedback and ratings from consumers that we received from user reviews of Amazon products. Based on that, we extracted the features of our dataset and established several supervised models. Such models provide algorithms for supervised machine learning such as Naive bays, logistic regression, support vector machines, Ensemble Classification, Decision Tree, and K-nearest neighbor. At last, we will compare all the models and check each model's accuracy with the ROC curve, recall, and precision.

II. Methods

A. Data Preprocessing

We take the dataset from reviews of Amazon Products [3]. Our dataset has 483148 of the total reviews. In this case, the product name, Brand, price, rating, text of the review, and the review of the device's cast. We will review in the review column to better use the data for the first, as they are the most critical aspects of this project. We separate positive and negative reviews below. Figure 1 is for positive reviews, and it is for negative reviews.

	Reviews	Rating
0	I feel so LUCKY to have found this used (phone...	5
1	nice phone, nice up grade from my pantach revu...	4
2	Very pleased	5
3	It works good but it goes slow sometimes but i...	4
4	Great phone to replace my lost phone. The only...	4

Fig. 1. Data preprocessing

Besides the brief overview of the dataset, we have plotted a distribution of ratings concerning the number of reviews, and we also perform the task where it calculates the total number of reviews with ratings 5,4,3,2,1. it shows

There are five classes in our dataset, which is the rating starts from 1 to 5 stars, as well as the division among them the five classes have been wrong, which is a class 2 and 3 with a small amount of data, while grade 5 has more than 175000 reviews. Here is an example from our data set: a Revision of the text: "I am using this phone, this is amazing, Rating: '5'. The rating distribution of Amazon reviews can be seen in Figure 2.

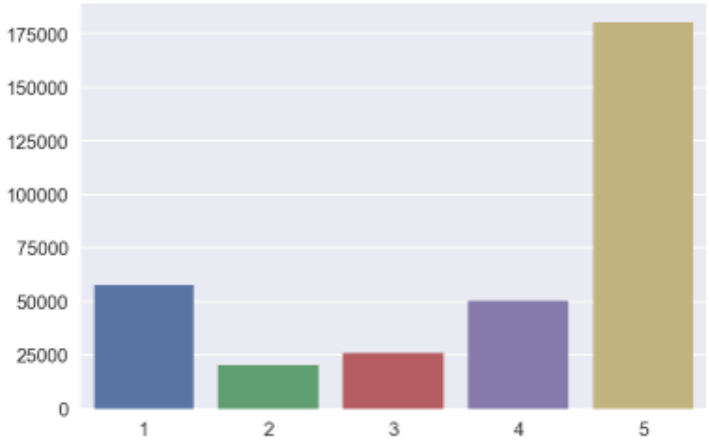

Fig. 2. Rating distribution of amazon reviews

For the research purpose of this project, we filtered the dataset with 16000 reviews and then again separated based on the review's rating.

B. Features

We have tried two types of features in our project. The first type is CountVectorizer [10]. The text must be analyzed to remove some terms to use textual data for predictive modeling, and it is also called the tokenization procedure. These words must then be encoded as integers or fluid-point values for machine algorithms as inputs. This procedure is known as function removal (or vectorization).

We use a Scikit learn library of CountVectorizer to convert a text collection into a vector of term/tokenization. This functionality makes it more flexible for text representation.

```
count_vector=CountVectorizer(stop_words="english")
```

The other method is TFIDF [11]. It is a statistical metric that assesses the significance of a word about a document in a collection of documents. This is because two components are multiplied: the number of times the term is in a document and the other way round the frequency of a document.

```
tfidf_vector = tfidfVectorizer(stop_words="English")
tfidf_vector.fit(X_train_data).
```

C. Classification

This research used six classification methods. The first is naïve bayes. The Naïve Bayes classification algorithm uses the alien of the theorem of Bayes to forecast the text tag based on the knowledge of its rules, terms, and circumstances [12]. It evaluates the chance of every tag being a text and then forecasts the time as likely as possible.

One of the most frequent tasks is the classification problems learning methods. In this approach, it is supposed that the x is dependent on the y, termed the assumption of Naïve Bayes. The calculation of naïve bayes as in (1).

$$P(x_1 \dots \dots \dots . x_k|y) = \prod_{i=1}^{k} p(x_i||y) \quad (1)$$

Second, utilized logistic regression to fix the binary classification problem using a classification technique in the classification of logistic regression, which utilizes a weighted combination of input and much effort [13]. The function Sigmoid transforms an actual number a to a number from 0 to 1.

A logistic regression classifier on Count Vectorizer and TFIDF features to compare it with rating accuracy. The default parameters that give us the accuracy of the results will be shown in the Results section. Logistic regression work with a sigmoid function, which predicts that the outcome values range from 0 to 1 or true false. The visualization of logistic regression can be seen in Figure 3.

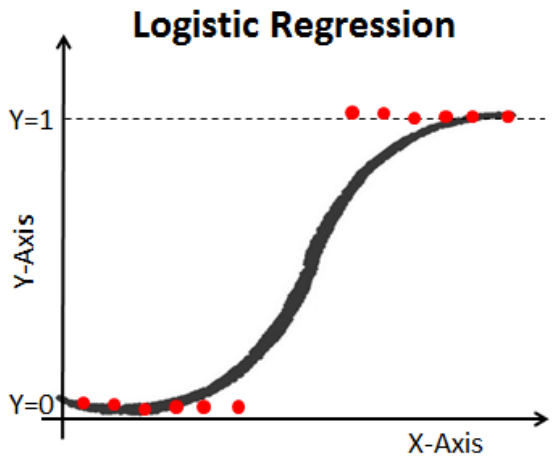

Fig. 3. Logistic regression

Third, a non-parametric classification procedure is the K-nearest neighbor (KNN). In recent years it has been frequently utilized. This approach is the closest neighbor of the input data to create a forecast for the first time for $K = n$. The great majority of the class's neighbors should then be mentioned. The distance between each neighbor and the distance Euclidean is a measure of the extent of similarity between the data points [14]. The equation of logistic regression as in (2).

$$f(x) = \frac{1}{K}(x + a)^n = \sum_{x \epsilon Nk(x)} y_i \quad (2)$$

Fourth, the Support Vector Machine (SVM) is a technique of classification that uses a small quantity of data to its best [15]. It is among the vectors belonging to a particular group or category and among those not belonging to the group.

Suppose, for example, two tags are available: costly and cheap, and the data contains two characteristics: x and y. It should be up to you to select which coordinates are more expensive and which are cheaper for each coordinate pair (x, y). In order to accomplish so, the SVM is to divide the two points, the so called border of decision, and, on the one hand, the group is so costly, and we cannot, on the other hand, reduce our costs.

Fifth, ensemble methods can create more than one model and then combine them to achieve better results [16]. Ensemble approaches are generally more precise than a single model [17]. This is also the case in several machine learning competitions, where the winning solutions are used in ensemble methods. The popular Netflix is ahead of the Competition, with the winner using a complex approach to implement a collaborative filtering algorithm. Here is the related code for this ensemble.

```
ess_model = RandomForestClassifier()
#Train Model
ess_model.fit(X_train_data_new,Y_train_data)
#Test Model
predictions["EssembleClasification"]= ess_model.predict(x_test_data_new)
```

The last is the decision tree. Decision tree is an algorithm of the supervised algorithm family of machine learning. It may be utilized both as a classification and regression problem [18]. The objective of the approach is to develop a model that predicts the value of a variable [19]. In order to resolve the problem of the leaf, the decision tree utilizes a tree representation to match a class label, and characteristics in the interior node of the tree are represented. The related code of decision tree as follows.

```
from sklearn import tree
tree_model = tree.DecisionTreeClassifier()
```

D. Evaluation Parameter

The methods or metrics we use to measure our project's evaluation are accuracy, precision, recall, and F1-score [20].

Precision predicts the percentage of positive reviews that use truly positive divided by the truly positive plus false positive as defined as in (3).

$$PR = \frac{tp}{tp+fp} \tag{3}$$

where tp is known as true positive and fp as false positive.

The recall measures the truly positive reviews divided by the total number of true positive and false positive reviews, as in (4).

$$RC = \frac{tp}{tp+fn} \tag{4}$$

where tp for true positive and fn for false negative

F1 Score is the combination of both precision and recalls, as in (5).

$$F1 - score = \frac{PR*RC}{PR+RC} \tag{5}$$

Accuracy measures the system's performance, the true positive and true negative reviews divided by the total number of actual, false positive, and false negative reviews, as in (6).

$$ACC = \frac{tp+tn}{tp+tn+fp+fn} \tag{6}$$

III. Results and Discussion

We divide the dataset of 483148 reviews into 80% of the training set and 20% of the testing set. After successfully training machine learning models, we used test data set to predict the model and test for accuracy. When the project was completed, we decided it was a significant activity that enabled us to reach our goal and gave us much confidence. We have designed a machine learning model that will help predict user review sentiments. This system can predict with different models' accuracy, which is quite valuable. Then the accuracy results are given in Table 1.

Table 1. The accuracy of count vectorizer and TFIDF model

Model	Accuracy	
	Count Vectorizer	***TFIDF***
Multinomial Naïve Bayes	0.924750	0.934750
Bernouli Naïve Bayes	0.819750	0.811625
Logistic Regression	0.952625	0.944750
KNN	0.898875	0.813750
SVM	0.749125	0.491625
Ensemble Classification	0.956750	0.960500
Decision Tree	0.938125	0.945500

The receiver operating curve (ROC) is a probability curve that indicates our binary classification based on the true and false-positive ratings. The area underneath the curve (AUC) is a metric of 0 to 1. The region underneath is the ROC curve. The ROC Curve of Ensemble Classification using TDIDF can be seen in Figure 4.

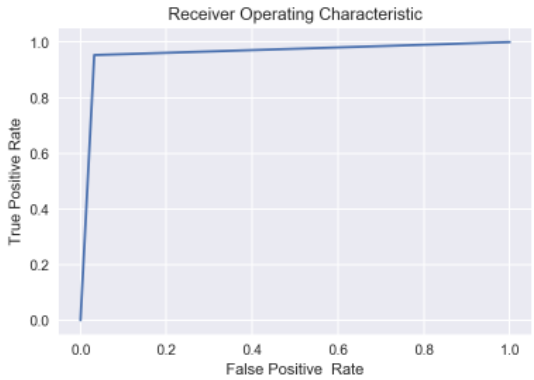

Fig. 4. ROC curve of ensemble classification using TFIDF

The above curve is only for Ensemble Classification using TFIDF techniques, and we also perform the same task for every model using TFIDF and Count Vector. We perform the following tasks with every model. These tasks were also performed with TFIDF and also with Count Vectorizer. The result of the evaluation can be seen in Table 2.

Table 2. The result evaluation

Features Model	Precision	Recall	F1-score
Count Vector	0.93	0.92	0.92
TFIDF	0.96	0.96	0.96

From Table 2, our model is quite successful as it produces 89-90 or more than 90% accuracy on test data set with different models and techniques, but it does not mean it can consistently produce such highly accurate results. There is a possibility that it can produce false results to some extent and can produce completely false results in some exceptions case. Positive reviews predictions must lie between the range of 0.5 and less than 1 and false reviews ranges from 0 to 0.5 but from the figure below, some false prediction of positive reviews represented pessimistically, and some pessimistic predictions represented positive ones. So there are some deficiencies which need to be resolved in future works. The actual and predicted output can be seen in Figure 5.

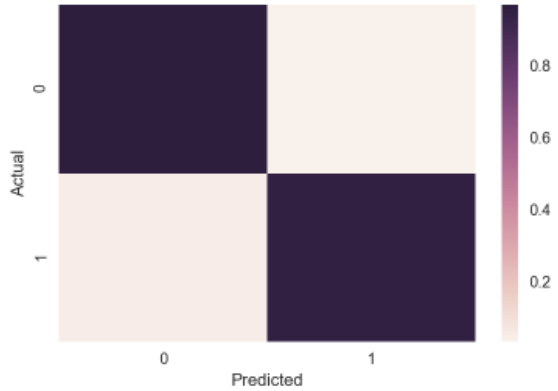

Fig. 5. Actual and predicted output

We live in a world of technology where artificial intelligence is a part of every system making it more autonomous and efficient. Nowadays, large ad networks and social or e-commerce businesses are implemented at a vast scale which uses targeted marketing and storing user data in a targeted manner by classifying user reviews in positive and negative using a system just like the system or algorithm we have developed using machine learning models. We also evaluated that combined or Ensemble machine learning models can produce more accurate and reasonable results than simple machine learning. At last, we compare all the models to check which model has the most fantastic accuracy, and our system is based on the GUI model, which performs the tasks in the following manners. The GUI model can be seen in Figure 6. The comparison results of the classification of all models in the system can be seen in Figure 7.

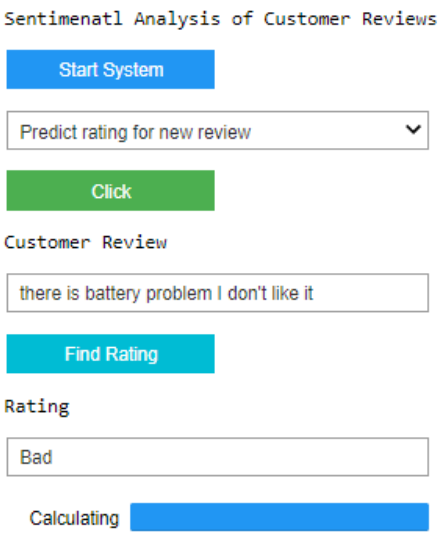

Fig. 6. System overview 01

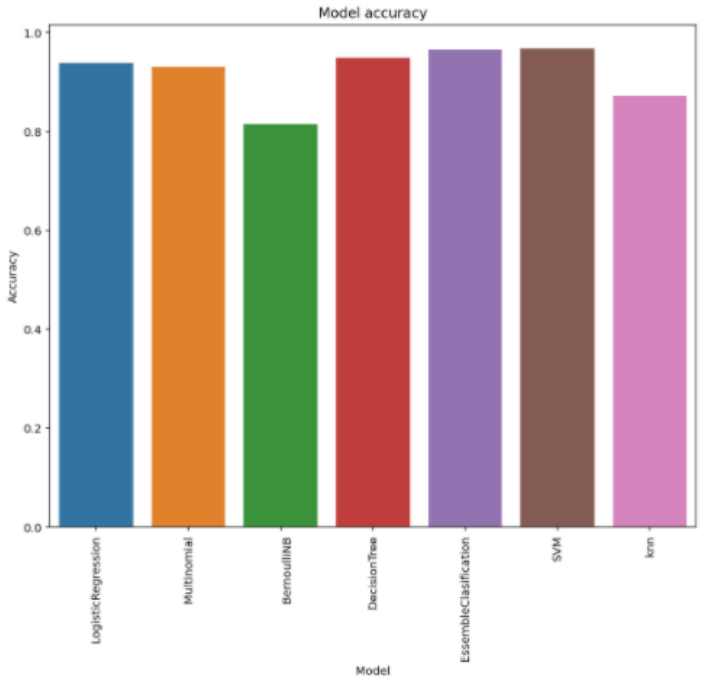

Fig. 7. Models comparison

IV. Conclusion

In conclusion, as we used two methods for different models, TFIDF and Count Vector, we used them with all the algorithms we mentioned in the model part, including Naive Bayes, SVM, KNN, Decision Tree, Logistic Regression, and Ensemble Classification. As we can see from the results, we have better accuracy on the test set with the following algorithms, Multinomial, Ensemble, and SVM Logistic Regression on both types of features. The same approach may be expanded to many more classification methods and utilizing a Neural network to decide whether the best classification for opinion mining and sentiment analysis will be chosen. One of the main features of this project, which remains a problem, is Problems Extraction from reviews. If this work is done in the future, it will benefit the suppliers or the company.

Declarations

Author contribution

All authors contributed equally as the primary contributor of this paper. All authors read and approved the final paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

Publisher's Note: Department of Electrical Engineering - Universitas Negeri Malang remains neutral with regard to jurisdictional claims and institutional affiliations.

References

[1] G. Taher, "E-Commerce: Advantages and Limitations," Int. J. Acad. Res. Accounting, Financ. Manag. Sci., vol. 11, no. 1, Feb. 2021.
[2] A. Datta, "The digital turn in postcolonial urbanism: Smart citizenship in the making of India's 100 smart cities," Trans. Inst. Br. Geogr., vol. 43, no. 3, pp. 405–419, Sep. 2018.
[3] A. S. Rathor, A. Agarwal, and P. Dimri, "Comparative Study of Machine Learning Approaches for Amazon Reviews," Procedia Comput. Sci., vol. 132, pp. 1552–1561, 2018.
[4] S. N. Ahmad and M. Laroche, "Analyzing electronic word of mouth: A social commerce construct," Int. J. Inf. Manage., vol. 37, no. 3, pp. 202–213, Jun. 2017.
[5] Z. Xiang, Q. Du, Y. Ma, and W. Fan, "A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism," Tour. Manag., vol. 58, pp. 51–65, Feb. 2017.
[6] J. Wang, M. D. Molina, and S. S. Sundar, "When expert recommendation contradicts peer opinion: Relative social influence of valence, group identity and artificial intelligence," Comput. Human Behav., vol. 107, p. 106278, Jun. 2020.
[7] Zhu Zhang, "Weighing Stars: Aggregating Online Product Reviews for Intelligent E-commerce Applications," IEEE Intell. Syst., vol. 23, no. 5, pp. 42–49, Sep. 2008.
[8] G. Kaur and A. Singla, "Sentimental analysis of Flipkart reviews using Naïve Bayes and decision tree algorithm," Int. J. Adv. Res. Comput. Eng. Technol., vol. 5, no. 1, pp. 148–153, 2016.
[9] U. R. Babu and N. Reddy, "Sentiment analysis of reviews for e-shopping websites," Int. j. eng. Comput. sci, vol. 6, no. 1, p. 19966, 2017.
[10] S. Khomsah and Agus Sasmito Aribowo, "Text-Preprocessing Model Youtube Comments in Indonesian," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 4, no. 4, pp. 648–654, Aug. 2020.
[11] A. I. Kadhim, "An Evaluation of Preprocessing Techniques for Text Classification," Int. J. Comput. Sci. Inf. Secur., vol. 16, no. 6, pp. 22–32, 2018.
[12] M. Castelli, L. Vanneschi, and Á. R. Largo, "Supervised learning: Classification," Encycl. Bioinforma. Comput. Biol. ABC Bioinforma., vol. 1–3, no. 2, pp. 342–349, 2018.
[13] M. Nabipour, P. Nayyeri, H. Jabani, S. S., and A. Mosavi, "Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis," IEEE Access, vol. 8, pp. 150199–150212, 2020.
[14] S. Hota and S. Pathak, "KNN classifier based approach for multi-class sentiment analysis of twitter data," Int. J. Eng. Technol, vol. 7, no. 3, pp. 1372–1375, 2018.
[15] D. A. Ragab, M. Sharkas, S. Marshall, and J. Ren, "Breast cancer detection using deep convolutional neural networks and support vector machines," PeerJ, vol. 7, p. e6201, Jan. 2019.
[16] O. Sagi and L. Rokach, "Ensemble learning: A survey," WIREs Data Min. Knowl. Discov., vol. 8, no. 4, Jul. 2018.
[17] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," Comput. Methods Programs Biomed., vol. 153, pp. 1–9, Jan. 2018.
[18] B. Choubin, E. Moradi, M. Golshan, J. Adamowski, F. Sajedi-Hosseini, and A. Mosavi, "An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines," Sci. Total Environ., vol. 651, pp. 2087–2096, Feb. 2019.
[19] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," Biomed. Signal Process. Control, vol. 52, pp. 456–462, Jul. 2019.
[20] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of Sentimental Reviews Using Machine Learning Techniques," Procedia Comput. Sci., vol. 57, pp. 821–829, 2015.