Sentiment Analysis on Amazon Reviews of Mobile Phones using Machine Learning

Shweta Singhal[1], Huda Lafta Majeed[2], Hassan Muayad Ibrahim[3], Nishtha Jatana[4], Charu Gupta[5], Agam Kumar[4], Bharti Suri[4], Oday Ali Hassen[6,*]

[1]Department of Information Technology Indira Gandhi Delhi Technical University for Women, New Delhi 110006, India

[2]Computer Science and Information Technology, University of Wasit, Al Kut 52001, Iraq

[3]University of Information Technology and Communication, Iraq

[4]Guru Gobind Singh Indraprastha University, Delhi, India

[5]Department of Computer Science and Engineering, Bhagwan Parshuram Institute of Technology, Delhi-85, India

[6]Ministry of Education, Wasit Education Directorate, Kut 52001, Iraq

Emails: miss.shweta.singhal@gmail.com; hulafta@uowasit.edu.iq; hassan.m@uoitc.edu.iq; nishtha.jatana@gmail.com; charu.wa1987@gmail.com; agamkumar1997@gmail.com; bhartisuri@gmail.com; oday123456789.oa@gmail.com

Abstract

The world is witnessing a boom in the digital age. Digital shops have literally landed into our homes. Almost any required product can now be purchased online via websites or mobile apps without having to step out. Due to online shopping, many customers rely on online reviews from other customers before making a purchase. Customer reviews are gaining more and more importance as they play a probably vital role in the sale and purchase of a product. Customer reviews also provide firsthand feedback coming directly from the customers themselves; this can benefit even the sellers in improving future sales. Analyzing the reviews can provide probable causes for failure or success of a product. Henceforth, the current paper presents the sentiment analysis of the reviews to better understand the feelings expressed by the customers. The very popular and widely used mobile phones were chosen as the product and Amazon was chosen as the digital seller for the current study. Initially, this work began with data preprocessing. Followed by data preprocessing, Bow and n-grams word embedding have been used to represent the clean reviews in vector representation, and then the features were derived. Finally, the performance of supervised machine learning classifiers such as Decision Tree, Naive Bayes, Random Forest, and SVM was empirically evaluated through accuracy, recall, f1-score, and precision. The results of empirical evaluation revealed that the Random Forest Classifier shows best performance with 97.48% accuracy.

Keywords: Sentiment Analysis; Phone Review; Machine Learning; Comparative Study; Random Forest

1. Introduction

Online selling and purchasing of various products have seen an enormous rise due to a paucity of time in this busy life. As a result, there is an advent of numerous e-shopping websites such as Amazon to meet customer requirements from almost anywhere, home/office/holidays etc. Customers are always on the lookout for good quality products at the cheapest possible price. As the physical or actual product cannot be checked for quality in an online purchase, thus customers rely on the reviews from other customers. These reviews affect the final decision of a customer to make the purchase or not. Better, analyze this, the need for sentiment analysis has been understood so that the popularity of a product can be easily reported among customers.

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

Sentiment Analysis comes under natural language processing (NLP) and is also known as opinion mining [1-7]. Several techniques have been applied in the area of Sentiment Analysis, that includes Recurrent Neural Networks [8], hybrid approaches [9-10], Ensembled approaches [11], deep learning [12]. A technique works on the text data to interpret the emotional tone (positive, negative, or neutral) of the author that is articulated in the text [13]. On an e-commerce website, there are options through which the customers give their opinion about the product [14]. Product sellers and new customers go through these reviews to know about the product's quality, popularity and other feedback. Understanding the customer sentiments have become extremely essential for the smooth running of any business. Customers can express their thoughts and feelings about a product and service much more freely than ever before. To understand the user experience, it is quite impractical for a human to go through every single line. Thus, the customer feedback can easily be analyzed automatically with the help of the latest techniques and technology [15-19].

Almost all purchasable goods can now be bought online. For the current study, we have chosen mobile phones and their reviews. The inevitability and abundance in mobile phone use by almost all human adults motivated this choice for our study. Before buying a mobile, the customers tend to check the sentiments of other customers using reviews. The reviews not only aid other customers in purchasing, rather they provide very useful insights to the product sellers too. The salespeople of the product would need to go through hundreds or thousands of product reviews for similar products. This task becomes very complicated and tedious to accomplish manually. Sometimes the rating score is good yet the sentiment of the review is seen as negative. In such a situation, the right sentiment can be ascertained only by understanding the review in more detail. Therefore, the purpose of this paper is to do a sentiment analysis of mobile phone reviews, in which this complex process can be simplified through machine learning.

Various e-commerce companies are in business nowadays. One of the oldest, most spread and widely used is Amazon. Henceforth, it has been considered for the current research. The authors in no manner promote or demote any e-commerce company. Amazon has been chosen for sheer research purposes and the abundance of available data. The sentiment analysis of an Amazon Unlocked Mobile Phones reviews dataset [20] is demonstrated in the current study. It involves classifying reviews as positive or negative using machine learning methods such as State Vector Machine (SVM), Decision Tree, Naive Bayes, and Random Forest [21]. Sentiment analysis is performed on the Amazon mobile phone reviews dataset, which has over 4 lakh reviews. The main goal of this study is to improve the data preprocessing steps and to compare the end performance of each model on the testing dataset by utilizing evaluation methods such as accuracy, recall, F1 score, precision, and confusion matrix [21].

Further, this paper has been section-wise organized as follows. In section 2, the work related to sentiment analysis of mobile phones or other products has been studied. After the study, the proposed methodology is explained in section 3. Along with this, data preprocessing and visualization have also been presented in this section. Thereafter, feature extraction for sentiment classification and its use to build different classification models are explained as well as the performance of the models is also given in section 4. Finally, the important findings found in this paper and future work are mentioned in Section 5.

2. Related Work

The process of buying products from customers on the e-commerce website and then giving reviews and ratings is a continuous process. The availability of data and the application of NLP techniques have attracted the attention of many researchers for sentiment analysis. We aim to do a sentiment analysis of mobile phone reviews from Amazon. The sentiment analysis studies already published relating to Amazon reviews or phone reviews have been studied in this paper.

There exists a lot of work [1-2], [4-5], [22-23] related to sentiment analysis on reviews of mobile phones. The corresponding data were collected and data preprocessing was initiated. Sentiment analysis was performed [1, 5] on unlocked mobile phones. Desai et al. [1] converted each review to lowercase, removed punctuation, removed stop words, and lemmatized, and tokenized in data preprocessing. Based on the rating, these data were divided into positive, neutral, and negative sentiments, but for training, positive and negative data were used. 1- and 2-star ratings were classified as negative, 3-star ratings as neutral, and 4-star and 5-star ratings were classified as positive sentiment. After data processing, 30,000 negative and 30,000 positive data were selected. The performance of machine learning and deep learning-based classifiers was compared, in which the deep learning model had the best performance at 98.51% and the Decision Tree at 95.1% in machine learning. In the pre-processing phase, [5] tokenized each review, removed stop words, lemmatized, converted to lowercase, and removed punctuation marks. They divided each review into positive, negative, or neutral classes based on rating. Ratings with 4 and 5 stars were made positive, ratings with 3 stars were made neutral and ratings with 1 and 2 stars were made negative. The number of neutral reviews was the lowest, which was 21000, so the researchers used this number for the balanced data of sentiments of different categories. They compared the performance of different machine learning and deep learning models with different feature extractions such as TF-IDF, bag-of-words, word2vec, and Glove. In which

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

the word2vec feature extraction technique gave a good performance with an accuracy of 92.72% for the CNN model. Similar work has been done on the mobile review data sets. Text data has been represented in numeric data by TF-IDF and then the performance of the model was achieved on cell phone data using different machine learning algorithms such as Naive Bayes, RNN, ANN, and SVM, in which the accuracy of RNN was 95.67% [2]. It was the best when the negation marking process was done with it and the effect of negation in sentences was detected. On the other hand, when the TF-IDF vectorizer was used with logistic regression, it had a good performance with 92% accuracy compared to Naïve Bayes and Random Forest [4]. To perform the sentiment analysis of data of 3 products from different mobile brands namely REDMI Note 3, APPLE IPHONE 5S, and SAMSUNG J7, 3 diverse machine learning algorithms Logistic Regression, Naive Bayes and SentiWordNet model were created and then the performance of these three models was measured in Recall, F1 score, and precision, in which the performance of the Naive Bayes classifier was found to be the best [22]. In addition to TF-IDF word embedding, after removing stop words (the word whose existence also has no importance in the analysis of our data), special characters, and punctuation as data preprocessing from the mobile review dataset [23], the text data was converted into a vector as a numeric representation using n-gram word embedding with a count-Vectorizer function that was imported from the Scikit-learn library. The vectorized text data is used to generate a prediction model using the Naive Bayes classifier, which gives a rating of 1 to 5 as the output with an accuracy of 74%. It gives overall 96% accuracy when 1- and 5-star ratings are predicted [23]. In this, the sentiment polarity score was determined with the help of the TextBlob library from the text data. The sentiment polarity score was then normalized to map the sentiment polarity score to a rating of 1 to 5.

Apart from mobile reviews, sentiment analysis of different products [24 - 25] was also studied. Hawlader et al. [24] performed sentiment analysis on the Amazon electronic product review dataset and for analysis, using Word2Vec, BoW, and TF-IDF in feature extraction and MLP classifier, SVM, Random Forest, Naive Bayes, and Decision Tree for classification. The researcher did a comparative study of its performance and assessed that the MLP classifier performed well using the BoW as feature extraction, which had an accuracy of 92% [11]. On the other hand, sentiment analysis of Amazon fine food reviews was done by different feature extraction and different supervised machine learning classifiers [25]. Using SVM, Random Forest, Naive Bayes, and KNN classifiers as well as TF-IDF, bag-of-words, and word2vec in feature extraction, reviews were categorized into positive and negative sentiments [25] and finally the researchers concluded that the accuracy of SVM with TF-IDF was the best among all the models with 94%.

Recently, the performance of sentiment analysis has been assessed using hybrid models [26 - 28]. Goswami et al. [26] studied the interaction between a user's emotional states and a mobile or computer device based on the emotional attributes of a user on the IMDb dataset. Hardness factor, positive or negative thinking factor, stress factor, negative emotion factor, and positive emotion factor, etc. used 5 as emotional attributes. For sentiment analysis, 8 datasets from different fields were used to build and estimate the hybrid deep model. This study is very helpful in making the system aware of emotional knowledge and providing the user with a good interaction system. Chauhan et al. [27] used Twitter data that was extracted using "Amazon" and "Hachette" tags. It did feature extraction from N-Gram word embeddings followed by 4 data preprocessing steps. Random forest and Naive Bayes were used to build the hybrid model and then compared the performance of this model with the model of Naive Bayes alone, in which the performance of the hybrid model was found to be very good. Through Table 1, the study of the paper related to Sentiment Analysis is shown in brief.

Table 1: Tabular Representation of literature review

Ref No	Year	Publisher	Types of paper	Model Used	Datasets	Classification Performance Measures
[1]	2021	IEEE	Comparisn	Decision Tree, BERT	Mobile	Accuracy: 95.1%, 98.51%
[2]	2021	ScienceDirect	New Technology	RNN	Cell phone	Accuracy: 95.67%
[3]	2020	IEEE	Comparisn	Convolutional neural network, NB	Books	Accuracy: 68%, 46.3%
[4]	2020	Springer	Comparisn	Random forest, Naive Bayes classifier and Logistic regression	Mobile	Accuracy: 92%, 91%, 93%
[5]	2019	IJACSA	Comparisn	word2vec+CNN	Mobile	Accuracy: 92.72**%**
[6]	2019	IEEE	Comparisn	The R script was built as a computed field Classification	Books	NA

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

				in Tableau, and it was applied using the classify polarity function.		
[7]	2018	IEEE	Comparisn	Support vector Machine Classifier (SVC)	1.Electronics reviews, 2. Cell Phone and Accessories Reviews and Musical Instruments product reviews	Accuracy: 94.02%.
[8]	2017	IEEE	Comparisn	LSTM, GRU	Health and Personal Care product	Accuracy: 78.10%, 83.90%
[9]	2017	Riga: University of Latvia	New Technology	SVM, introduced new method	Product Review	Accuracy: 71.33%, 72.00%
[10]	2016	IEEE	Comparisn	Naïve Bayes' Logistic Regression	APPLE IPHONE 5S, SAMSUNG J7, REDMI NOTE 3	**Naïve Bayes:** Recall =0.870 Precision =0.675, F-Measure =0.760 and **Logistic Regression:** Recall =0.778 Precision =0.713 F-Measure =0.744
[11]	2016	Springer	Comparisn	SentiME,	Product Reviews	Precision =0.85686 Recall=0.90541 F1 score =0.88046
[12]	2015	IEEE	Comparisn	Deep Neural Network, SVM	Electronic products	Accuracy: 90%, 85%
[13]	2015	Springer	Comparisn	SVM, LSTM	beauty, book, electronic, and home products	Accuracy: 80.95%, 87.6%
[14]	2015	IJCSIT	Comparisn	Sentiment Polarity Methodology	iPhone5	Not Used
[15]	2014	IEEE	New Technology	Existing Method, Proposed Method	Digital camera	Accuracy: 74.1 95%, 76.02%
[16]	2014	ScienceDirect	New Technology	SVM, Proposed Method	Books	Accuracy: 88.43%, 87.25%
[17]	2013	ACM	Comparisn	Boosted SVM	movie reviews	Accuracy: 93%
[18]	2012	ACM	Comparisn	Opinion word extraction and aggregation enhanced with	movie reviews	Accuracy: 90.15%
[19]	2012	Springer	Comparisn	A Lexicon-based Classification	Reviews of hotels	Precision values are 84% and 92% for both positive and negative reviews.

3. Methodology

Sentiment analysis, also known as Opinion Mining, is a type of classification problem. In this, someone's idea, which is articulated in the text, is divided into neutral, positive, or negative sentiments. A classification algorithm

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

[21] (e.g. SVM, Naïve Bayes) is used for classification. Here, many classification algorithms based on machine learning have been used. The methodology is explained below in Fig. 1.

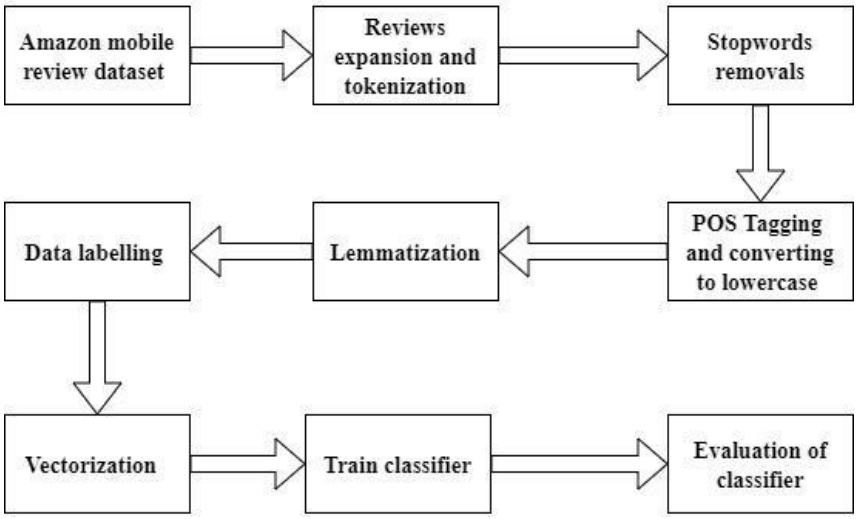

Figure 1. Proposed Methodology for the Sentiment Analysis on Amazon Reviews

In the above Fig 1, loading of the amazon mobile review dataset, expansion, and tokenization of reviews, stop words removals, and post tagging and converting to lowercase are used in data preprocessing steps. Data preprocessing is mentioned in detail in the next subsection 3.3. Vectorization [21] is used in feature selection. The process of vectorization, also known as word embedding, is used to represent text data in numeric data separately. TF-IDF, Word2vec, Bag of Words (BoW), N-Gram, and other standard algorithms are used in word embedding [21]. A trained classifier is used for the model building. For this, the supervised machine learning algorithm in which SVM, Decision Tree, Naive Bayes, and Random Forest have been used and models have been prepared through it. It's just as crucial to evaluate a machine learning model as to develop this model. The model being built will run on previously uncovered data. To construct a robust model, a thorough and versatile evaluation is needed. Evaluation of the classifier is used to know the classification capability of the model after it is created by utilizing evaluation methods such as accuracy, recall, F1 score, precision, and confusion matrix [21]. The confusion matrix is a table that classifies the model's predictions to show how well the classifier performs. The matrix is built for a set of test data for which the true values have already been determined. Simply said, the number of relevant and irrelevant forecasts is compared to the true values. It also checks the classification model's accuracy by evaluating some parameters. The following are the variables:

- True Positives (TP): TP is defined as the number of samples that are predicted to be positive-class by a classifier and are positive-class.
- True Negatives (TN): TN is the number of samples that a classifier predicts as a negative class and to which it belongs.
- False Positives (FP): FP is defined as the number of samples that are projected to be negative but belong to the positive class.
- False Negatives (FN): FN is the number of times a classifier predicts positive-class samples, but those samples are negative-class samples.

3.1 Feature Extraction

One of the things that must be finished before implementing a classifier is feature extraction. The removal of inappropriate and superfluous words is the most crucial task in feature extraction for greater classification results. In the context of text data, features are mined with the following techniques [21]:

- Bag of words

This is one of the easiest word embedding techniques that work by frequency count. In this, all the words in the corpus are collected, and then by counting the frequency of each unique word, it is converted into a vector. In this, the order of the word does not matter, the disadvantage of which is that it is difficult to capture the semantic meaning in it. Apart from this, there is also the problem of sparsity.

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

- N-Grams

It is a word embedding technique and is commonly used. Because it can tell the semantics of a sentence and its implementation is also easy. Here, N denotes the sequence of n nearby items. When the value of n is 1, it is called a unigram, a bigram for the value of n is 2, a trigram for the value of n is 3, and so on. For example, the n-gram of the sentence "This is a phone" for a different value of n would be as follows.

n=1 (unigrams)	[‘This’, ‘is’, ‘a’, ‘phone’]
n=2 (bigrams)	[‘This is’, ‘is a’, ‘a phone’]
n=3 (trigrams)	[‘This is a’, ‘is a phone’]

- TF-IDF

It is used to give weight to a word. The logic is that if a word is appearing more often in a sentence and less frequently in the entire corpus, then its weight is given more and if it is appearing more often in the corpus then its weight is given less. Two terms are used to give the weighting, TF, and IDF.

Let,

A = Number of occurrences of term t in document d

B = Total number of terms in document d

C = Total number of documents in the corpus

D = Number of documents with the term t in them

$$TF = \frac{A}{B} \quad \text{……. (1)}$$

$$IDF = \frac{C}{D} \quad \text{……. (2)}$$

$$\text{Weight of term t in document d} = TF*IDF \quad \text{……. (3)}$$

- WORD2VEC

It is a word embedding technique developed by Google, which works on deep learning. In this, the semantic meaning of the word is captured, as Happy and Joy are the same word. Its advantage is that it generates a dense matrix whereas the other techniques like BoW, TF-IDF, and N-Grams generate a sparse matrix with many of zero entries.

3.2 Classification Algorithms

The main goal of this step is to make a categorization model utilizing a variety of machine-learning algorithms with the help of the sklearn library [28]. The data is separated into train and test sets after feature extraction. The training data is used to train the model for categorizing the provided review into positive or negative. The test data is used to determine the accuracy of the classification model and to evaluate it. Four prominent supervised classifiers will be employed in this project, which are:

- Naive Bayes
- Random forest
- Decision Tree and
- Support Vector Machine (SVM).

3.2.1 Naive Bayes

One of the simplest classification methods based on the Bayes theorem is the Naive Bayes algorithm. The Bayes theorem discusses the possibility of reason. It is predicated on the notion that the characteristics of the data points are independent of one another. In practice, the characteristics of a data point can rely on one another. It is termed Naive Bayes for this reason. The most typical applications of this strategy are for binary and multi-class classification problems.

3.2.2 Random Forest

Random Forest is merely an ensemble method for solving regression and classification problems that combines a number of decision trees with a method called bagging. The fundamental idea is to combine several decision trees

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

to determine the outcome rather than relying just on individual decision trees. Random Forest uses numerous decision trees as a basic learning model, and the outcome is decided by majority voting.

3.2.3 Decision Tree

Decision tree algorithms are a type of supervised machine learning in which data is continually divided depending on parameters using a greedy approach. The decision tree is explained through decision nodes and leaves. Decision nodes segregate the data based on decisions or outcomes, which are represented by leaves.

3.2.4 Support Vector Machine (SVM)

SVM works on labeled data, which is a machine-learning algorithm. It works on both classification and regression. In classification, there is a need for a decision boundary that helps to separate one class from another. Decision boundaries may be linear or nonlinear. If data is linearly separable then a linear decision boundary is used whereas if data is not linearly separable then a nonlinear decision boundary is used. In SVM, data is classified using a hyperplane. Based on the number of features, the dimension of the hyperplane is decided. If there are only 2 features x1 and X2 then the decision boundary will be a line. If there are 3 features X1, X2 and X2 then the decision boundary will be a plan. Generally, if there are n features then the dimension of the decision boundary will be n-1. Many hyperplanes possibly separate positive class and negative class.

After understanding the data pre-processing, feature selection, and classification techniques, the procedure for sentiment analysis of Amazon phone review is explained below systematically.

3.3 Procedure

The procedure for sentiment analysis of Amazon phone reviews is explained systematically.

Step 1: load the dataset, remove the empty review row, and expand the shortened form of the review

Step 2: Tokenize the review into words using the nltk library.

Step 3: Create the stop words list for the English language using nltk library and add punctuation and some modal auxiliary verbs in the stop words list that are not available in the stop word list before this.

Step 4: Handle the negation words [not, no] by removing them from the stop words list and creating the final list of stop words.

Step 5: Remove the stop words, special characters, and numbers from the tokenized words without considering case sensitivity.

Step 6: Getting POS (Part of Speech) tags of the tokenized words and then converting the tokenized words into lowercase.

Step 7: Lemmatize the tokenized words using POS tags and join those words to make a clean review sentence.

Step 8: Assign sentiment type to each review using rating. If a rating has 1 and 2 stars then assign -1(negative sentiment) and 1(positive sentiment) for the 4 and 5 stars rating whereas assigning 0(neutral sentiment) for 3 stars rating.

Step 9: Divide the data into train test split and convert the review into a vector using the CountVectorizer function of the sklearn library [28] and extract features using the N-gram word embedding technique from the train data

Step 10: Train the classifier using train data to develop the prediction model using a supervised machine-learning algorithm for the extracted feature and sentiment types

Step 11: Evaluate the model performance using the assessment parameters such as precision, recall, accuracy, and f1 score [27], etc.

3.3 Dataset

The dataset is of Unlocked Mobile Phones [20], in which the verified customer on the Amazon website has given the rating and review of the phone. The dataset used is available to all in the public domain on Kaggle [22]. This dataset has 413840 reviews of different products from different brands of mobile phones. A sample from the dataset is shown in Fig. 2. There are 6 attributes which means 6 columns in the dataset which are as follows.

- Product Name: It states the name of the product.
- Brand Name: It states the name of the brand.
- Price: It shows the price of the product.
- Rating: It shows the rating shared by the customer.

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

- Reviews: This column gives a review of the product.
- Review Votes: Tells how many customers have been helpful through a review.

	Product Name	Brand Name	Price	Rating	Reviews	Review Votes
0	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	5	I feel so LUCKY to have found this used (phone...	1.0
1	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	4	nice phone, nice up grade from my pantach revu...	0.0
2	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	5	Very pleased	0.0
3	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	4	It works good but it goes slow sometimes but i...	0.0
4	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	4	Great phone to replace my lost phone. The only...	0.0

Figure 2. Sample from the dataset

3.4 Data Preprocessing

In data processing, incomplete, duplicate, and noisy data acquired in the actual world is turned gradually into a useful and effective format. Here, the review is pre-processed systematically in the following manner.

Step 1 Handling null values: It is important to find out which column has null values and then handle those null values appropriately. It has been found that 65171 in the brand name, 5933 in the price, 62 in the review, and 2296 in the review votes were null values. Since interest is in dividing the only and only reviews into positive and negative sentiment, the row in which the only and only review was null has been eliminated and the number of such rows was only 62. After removing 62 rows out of the total of 413840, there were 413778 rows left.

Step 2 Expansion of shortened form of a word: Whatever was in the shorthand form has been expanded in the review text to capture the correct sentiment of the word that comes after the word ‘not’ later as if the word good is a positive sentiment, then the same not good can be classified as negative sentiment.

Step 3 Removing stop words and punctuation: In text sentiment analysis, it is necessary to remove those words which do not strengthen the analysis, such words are called stop words. For this, the stop words of English words have been taken using nltk and Modal auxiliary verbs have been added that were not already there in this list such as 'could', 'might', 'must', 'need', and 'would'. As for the word not, no was already on the list of stop words, it was removed from the list and simultaneously added punctuation to the list. This is how a list of stop words has been prepared. The stopwords set is problem specific, it is not a universal list and needs to be removed from the text in the analysis so that a good feature set can be extracted. Apart from this, any special symbols and digits other than text have been removed.

Step 4 Tokenization and post tagging: After this, the review is divided into small word units, this process is called tokenization. The words in the stopword were in lowercase, so the tokenized review text also had to be converted to lowercase. Before converting to lowercase, the pos tag of that word is required because after converting to lowercase the actual tag of a word can change. Pos tags are needed in lemmatization.

Step 5 Lemmatization: Lemmatization is nothing but the procedure of bringing the root form of a word so that different forms of a word can be considered into a single item. With the help of lemmatization, words have been converted to the base form.

In the end, out of the 6 columns, kept the rating and review column and removed the rest. Along with this, a new column was added named ‘Clean Reviews’, in which the lemmatized word was joined and inserted as values. Which is presented in Fig. 3.

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

Rating	Reviews	Clean Reviews
5	I feel so LUCKY to have found this used (phone...	feel lucky found use phone u not use hard phon...
4	nice phone, nice up grade from my pantach revu...	nice phone nice grade pantach revue clean set ...
5	Very pleased	pleased
4	It works good but it goes slow sometimes but i...	work good go slow sometimes good phone love
4	Great phone to replace my lost phone. The only...	great phone replace lose phone thing volume bu...

Figure 3. Sample of the dataset after cleaning.

	Rating	Reviews	Clean Reviews
310	5	A+++++++++	NaN
763	4	😊	NaN
1075	5	100%	NaN
1193	4	😊	NaN
2020	5	a+	NaN
...	...	...	...
412043	5	A+	NaN
412160	5	A+	NaN
412307	5	A+	NaN
412484	5	A+	NaN
412601	5	A+	NaN

1018 rows × 3 columns

Figure 4. Data whose value become null after cleaning

After cleaning the review, once again whether any value is null in the clean review column has been checked. It was observed that there are 1018 null values in the Clean Reviews column. The values that became null after cleaning have been checked again. It was observed that some reviews had emojis and some had grading as presented in Fig. 4. It was not considered to do text-based sentiment analysis. As such, grading itself is a rating and not a review. 1018 rows out of 413,778 rows that had null values have been removed and the remaining 412760 rows have been used further.

3.5 Data visualization and analysis

Data representation helps a lot in data analysis. Tableau [29] and Matplotlib have been used for data visualization. The top 10 brands based on the best-selling products from the dataset have been selected. Which is shown below through a bar graph in Fig 5. In this, the phones of the Samsung brand were the most sold.

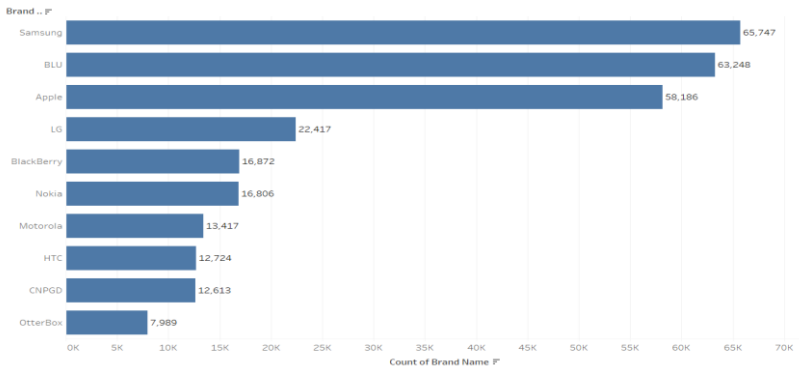

Figure 5. Top 10 brands of mobile phones based on the product sales

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

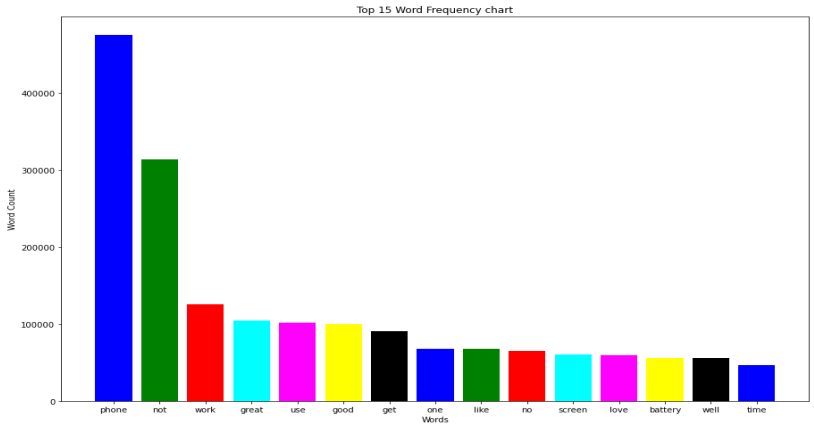

Figure 6. Top 15 frequently occurring words

From all the clean reviews, a list of all the words called corpus was made. It does not contain unique words. The word frequency for unique words has been found. This helped us in getting the number of times a word has occurred. 'Mobile' word came up most often (475189 times) in the corpus, followed by no word. The top 15-word frequencies are shown in Fig 6.

The length of each clean review was obtained and from the analysis, it was seen that the mean length of the review was 110.501614. In 75% of the data, the length of the review was 114, with the minimum review length being 1 and the maximum review length 15312. From this, it can be observed that if someone wants to get feedback from the customer about a product, then it would be appropriate to take the capacity of the textbox up to 300 characters. Now based on rating, the three groups of sentiment have been made. 1- and 2-star ratings are made negative, 3-star ratings are neutral and 4- and 5-star ratings are made positive. Negative sentiment is shown as -1, neutral sentiment as 0, and positive sentiment as 1 and it is shown in Fig 7. The clean dataset had 97025-row negative sentiment, 31725-row neutral sentiment, and 284010-row positive sentiment out of 412760 rows as shown in Fig.8 In this way, two new features of Review Length and Sentiment were added. The updated data was saved on the disk.

Rating	Clean Reviews	Review Length	Sentiment
5	feel lucky found use phone u not use hard phon...	178	1
4	nice phone nice grade pantach revue clean set ...	123	1
5	pleased	7	1
4	work good go slow sometimes good phone love	36	1
4	great phone replace lose phone thing volume bu...	105	1
1	already phone problem know state use dang not ...	176	-1
2	charge port loose get solder need new battery ...	97	-1
2	phone look good not stay charge buy new batter...	93	-1
5	originally use samsung galaxy sprint want retu...	387	1
3	battery life great responsive touch issue some...	104	0

Figure 7. Dataset with sentiment and review length

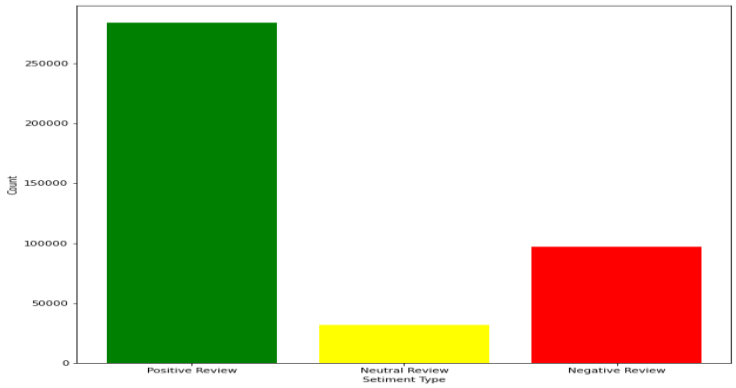

Figure 8. Bar chart of different sentiment types

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

4. Experimental setup and Results

In this section, the classification model has been created with the help of a supervised machine-learning algorithm. For this, the cleaned dataset from the disk has been imported with the help of the pandas' library [31]. As in the cleaned dataset, there are 3 types of sentiments positive, neutral, and negative. The neutral sentiment data has been removed, as this data was not very significant to training the model. The data in the Neutral category was also less as compared to the other classes. Now there were only two types of sentiments, positive and negative. It has been seen that the number of reviews with negative sentiment was 97025 and the positive sentiment was 204010, which is not evenly distributed. The model should not be trained with biased data in any way. Any biased data should be discarded, to design a good model that could classify both sentiments correctly. For this, all the data of negative sentiment whose number was 97025 has been taken and the same amount of data i.e., 97025 data from positive sentiment has been taken. Thus, 194050 data (97025 negative and 97025 positive) have been taken to discard bias. After this, the data was divided into train and test data. 30% of data i.e., 58215 data have been used in testing, and 70% of data i.e., 135835 in training. The following steps were then followed.

- Feature Extraction
- Model design
- Evaluation of the model and analysis of the result

4.1 Feature extraction

It is needed to encode the text data into numeric data before inputting it into the machine-learning model so that the machine can understand and operate on the numeric data. The pixel value of the image is given from the image data to the machine and it does not require a separate representation. Nevertheless, in the context of text data, it needs to represent text data separately in numeric data and this process is called Vectorization or Word embedding. There are some common techniques in word embedding like Bag of words (BoW), N-Gram, TF-IDF, Word2vec [21], etc. which have their advantages and disadvantages. In this paper, n-gram word embedding has been used because it can tell the semantics of a sentence and its implementation is also easy. Here, N denotes the sequence of n nearby items. When the value of n is 1, it is called a unigram, a bigram for the value of n is 2, a trigram for the value of n is 3, and so on. For example, the n-gram of the sentence "This is a phone" for a different value of n would be as follows.

n=1 (unigrams) [‘This’, ‘is’, ‘a’, ‘phone’]

n=2 (bigrams) [‘This is’, ‘is a’, ‘a phone’]

n=3 (trigrams)[‘This is a’, ‘is a phone’]

It was needed to build the feature in such a way that the model could predict the correct semantic of the word even when it is made negative, for example, "good", "not good", or "not good phone". For this, a combination of unigrams, bigrams, and trigrams have been used along with BoW for feature extraction. In this, 3500 features were taken for which the max_features hyperparameter of CountVectorizer was set to 3500, and the ngram_range hyperparameter was kept at (1,3).

4.2 Model Design

After feature extraction, 4 models have been built using a Decision Tree, Naive Bayes, Random Forest, and SVM classifier [21] to perform sentiment classification. The model has been trained with 135835 training data. After the model has been trained, the confusion matrix, accuracy, precision, recall, and f1 score [21] are used as the evaluation parameters. 58215 testing data was used for model evaluation.

4.3 Evaluation of the model and analysis of the result

In this section, the performance of different classifiers has been assessed. The confusion matrix, accuracy, precision, recall, and f1-score have been calculated for Decision Tree, Naive Bayes, Random Forest, and SVM classifiers. The confusion matrix of all the classifiers is shown in Fig 9.

Table 2 and Table 3 show precision, recall, and f1-score for both positive and negative sentiment respectively.

The features assigned to it during training affect the model’s success. The performance of the model is related to the choice of the right feature. The data pre-processing was done well. The expansion of the text made it easier to understand the negative term, which helped the model to correctly predict the type of emotion. The pos tag was taken before converting the text to lowercase so that changing the text from uppercase to lowercase would not affect the pos tag. As “Running” is used as a noun in the sentence, “Running is a good exercise”. If it is changed to the lower case, then its pos tag will change which has the effect of lemmatization. That is why after detecting the pos tag, the text was converted to lowercase and then lemmatized based on the pos tag. For feature selection,

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

a combination of unigram, bigram, and trigram was used, resulting in a model with good accuracy. In the n-gram word embedding, there is a sparsity issue. After feature extraction, work on feature selection is also required which will be done in the future. As a result, the number of features will be reduced, and the least amount of memory will be required for the vector matrix. The accuracy of all the models was greater than 90% as shown in Fig 10. It can be observed that the Random Forest classifier outperforms the other three classifiers with an accuracy of 97.48%, while the Naive Bayes classifier performs the worst with an accuracy of 91.92%.

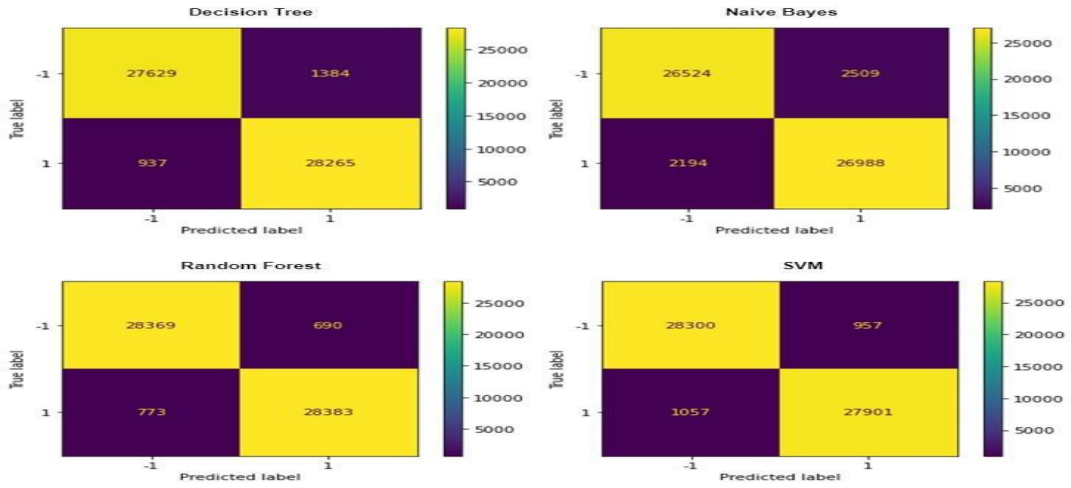

Figure 9. Confusion Matrix of the different model

Table 2: Result Table of the different classifiers for Positive sentiment

Classifier	Precision	Recall	F1-Score
Decision Tree	0.95	0.97	0.96
Random Forest	**0.98**	**0.97**	**0.97**
SVM	0.97	0.96	0.97
Naïve Bayes	0.91	0.92	0.92

Table 3: Result Table of the different classifiers for Negative sentiment

Classifier	Precision	Recall	F1-Score
Decision Tree	0.97	0.95	0.96
Random Forest	**0.97**	**0.98**	**0.97**
SVM	0.96	0.97	0.97
Naïve Bayes	0.92	0.91	0.92

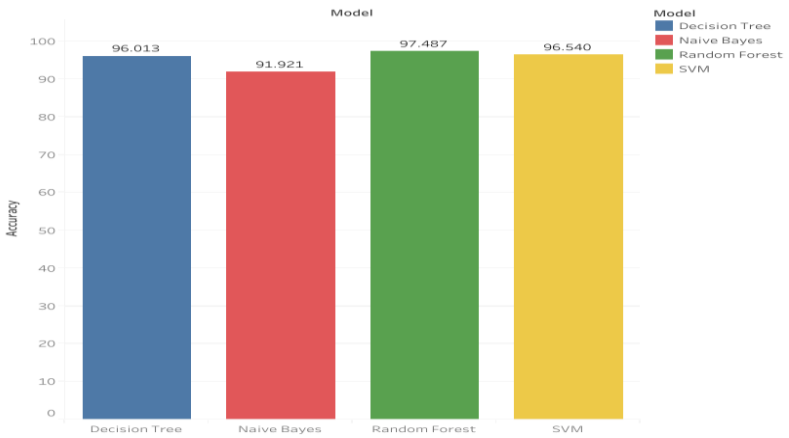

Figure 10. Accuracy of the different models.

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

5. Conclusion and Future work

The proposed work, implements sentiment analysis on customer reviews from Amazon for Mobile Phones using machine learning. The performance of the model is related to the features given to it for training. There is a proportional relationship between the selection of the right feature and the performance of the model. Data pre-processing was done well and promising results were obtained. All the models had an accuracy of more than 90%. 4 models have been built using different machine learning classifiers such as Decision Tree, Naive Bayes, Random Forest, and SVM classifier, and their performance was empirically evaluated. A combination of unigram, bigram, and trigram has been used for feature selection, which resulted in yielding an accuracy of 91.92% in case of Naive Bayes, 96.01% in Decision Tree, 97.49% in Random Forest, and 96.54% in SVM. Thereby, it can be concluded that the Random Forest classifier gives the best performance among all the 4 classifiers while the Naive Bayes presents the lowest accuracy. The frequency of each unique word in the cleaned review was found, and it has been observed that 'phone', 'not', 'work', 'great', and 'use' were the 5 most frequently used words. In this work, the performance of 4 supervised machine learning models is studied for sentiment analysis of the amazon mobile review, while there are many more models of machine learning and deep learning, for evaluation of their performance, which may be empirically evaluated with the present study in the future. Customers can share their feelings through emojis or special characters in some cases, which will also be taken care of at the time of data processing in future work. Apart from this, the performance of different models with different feature extraction techniques may also be evaluated and compared with the performance of the current model. There was the problem of sparsity in the n-gram word embedding. For this, future work may include working on feature selection after feature extraction, which might result in reducing the number of features and using the least amount of memory for the vector matrix. This may result in providing useful information for the sellers and websites to improve their sales.

Funding Statement: No funds were taken from any organization for the work conducted.

Conflicts of Interest: No Conflict of interest to disclose.

References

[1] P. Mukherjee, Y. Badr, S. Doppalapudi, S. M. Srinivasan, R. S. Sangwan, and R. Sharma, "Effect of negation in sentences on sentiment analysis and polarity detection," Procedia Computer Science, vol. 185, no. 1, pp. 370–379, 2021. DOI: https://doi.org/10.1016/j.procs.2021.05.044.

[2] S. Tammina and S. Annareddy, "Sentiment analysis on customer reviews using convolutional neural network," in 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020. DOI: https://doi.org/10.1109/ICCCI48352.2020.9104057.

[3] Banerjee, N. Intwala, and V. Sawant, "Sentiment analysis of Amazon mobile reviews," in 2nd International Conference on Information and Communication Technology for Sustainable Development (ICT4SD), Samarinda, Indonesia, 2019. DOI: https://doi.org/10.1007/978-981-13-7166-0_14.

[4] S. A. Aljuhani and N. S. Alghamdi, "A comparison of sentiment analysis methods on Amazon reviews of mobile phones," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 6, pp. 608–617, 2019. DOI: https://doi.org/10.14569/IJACSA.2019.0100678.

[5] Almjawel et al., "Sentiment analysis and visualization of Amazon books' reviews," in 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, 2019. DOI: https://doi.org/10.1109/CAIS.2019.8769467.

[6] T. U. Haque, N. N. Saber, and F. M. Shah, "Sentiment analysis on large-scale Amazon product reviews," in 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, Thailand, 2018. DOI: https://doi.org/10.1109/ICIRD.2018.8376357.

[7] K. Baktha and B. K. Tripathy, "Investigation of recurrent neural networks in the field of sentiment analysis," in 2017 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2017. DOI: https://doi.org/10.1109/ICCSP.2017.8286386.

[8] K. Korovkinas, P. Danėnas, and G. Garšva, "SVM and Naïve Bayes classification ensemble method for sentiment analysis," Baltic Journal of Modern Computing, vol. 5, no. 4, pp. 398–409, 2017.

[9] K. S. Kumar, J. Desai, and J. Majumdar, "Opinion mining and sentiment analysis on online customer review," in 2016 International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, India, 2016. DOI: https://doi.org/10.1109/ICCIC.2016.7919614.

[10] Z. Hu, J. Hu, W. Ding, and X. Zheng, "Review sentiment analysis based on deep learning," in 2015 12th International Conference on e-Business Engineering (ICEBE), Beijing, China, 2015. DOI: https://doi.org/10.1109/ICEBE.2015.26.

[11] X. Fang and J. Zhan, "Sentiment analysis using product review data," Journal of Big Data, vol. 2, no. 1, pp. 1–14, 2015. DOI: https://doi.org/10.1186/s40537-015-0015-2.

[12] Bhatt et al., "Amazon review classification and sentiment analysis," International Journal of Computer

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024

Science and Information Technologies, vol. 6, no. 6, pp. 5107–5110, 2015.

[13] J. Bhaskar, K. Sruthi, and P. Nedungadi, "Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers," in 2014 International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, India, 2014. DOI: https://doi.org/10.1109/ICRAIE.2014.6909196.

[14] E. Fersini, E. Messina, and F. A. Pozzi, "Sentiment analysis: Bayesian ensemble learning," Decision Support Systems, vol. 68, no. 1, pp. 26–38, 2014.

[15] Sharma and S. Dey, "A boosted SVM-based sentiment analysis approach for online opinionated text," in 2013 Research in Adaptive and Convergent Systems (RACS), Montreal, Canada, 2013. DOI: https://doi.org/10.1145/2513228.2513296.

[16] Srivastava et al., "Sentiment analysis of Twitter data: A hybrid approach," International Journal of Healthcare Information Systems and Informatics (IJHISI), vol. 14, no. 2, pp. 1–16, 2019. DOI: https://doi.org/10.4018/IJHISI.20190401.oa01.

[17] Gallagher, E. Furey, and K. Curran, "The application of sentiment analysis and text analytics to customer experience reviews," International Journal of Data Warehousing and Mining (IJDWM), vol. 15, no. 4, pp. 21–47, 2019. DOI: https://doi.org/10.4018/IJDWM.20190401.oa01.

[18] K. Kowsari et al., "Text classification algorithms: A survey," Information, vol. 10, no. 4, p. 150, 2019. DOI: https://doi.org/10.3390/info10040150.

[19] S. Yarkareddy et al., "Sentiment analysis of Amazon fine food reviews," in 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2022. DOI: https://doi.org/10.1109/ICSSIT.2022.9715981.

[20] Goswami et al., "Sentiment analysis of statements on social and electronic media using machine and deep learning classifiers," Computational Intelligence and Neuroscience, preprint, 2022. DOI: https://doi.org/10.1155/2022/8439092.

[21] J. Sangeetha and U. Kumaran, "Sentiment analysis of Amazon user reviews using a hybrid approach," Measurement: Sensors, vol. 27, p. 100790, 2023. DOI: https://doi.org/10.1016/j.measen.2023.100790.

[22] M. F. Harunasir et al., "Sentiment analysis of Amazon product reviews by supervised machine learning models," Journal of Advances in Information Technology, vol. 14, no. 4, pp. 857–862, 2023. DOI: https://doi.org/10.12720/jait.14.4.857-862.

[23] Gitanshu Chauhan et al., "Amazon product reviews sentimental analysis using machine learning," in 2024 IEEE International Conference on Computing, Power, and Communication Technologies (IC2PCT), 2024. DOI: https://doi.org/10.1109/IC2PCT.2024.9900625.

[24] H. Ali et al., "Analyzing Amazon product sentiment: A comparative study of machine and deep learning, and transformer-based techniques," Electronics, vol. 13, no. 7, p. 1305, 2024. DOI: https://doi.org/10.3390/electronics13071305.

[25] M. Hawlader et al., "Amazon product reviews: Sentiment analysis using supervised learning algorithms," in 2021 International Conference on Electronics, Communications, and Information Technology (ICECIT), Khulna, Bangladesh, 2021. DOI: https://doi.org/10.1109/ICECIT.2021.9659335.

[26] H. Ali, E. Hashmi, Y. Y. Yildirim, and S. Shaikh, "Analyzing Amazon product sentiment: A comparative study of machine and deep learning, and transformer-based techniques," Electronics, vol. 13, no. 7, p. 1305, 2024. DOI: https://doi.org/10.3390/electronics13071305.

[27] J. Sangeetha and U. Kumaran, "Sentiment analysis of Amazon user reviews using a hybrid approach," Measurement: Sensors, vol. 27, p. 100790, 2023. DOI: https://doi.org/10.1016/j.measen.2023.100790.

[28] M. F. Harunasir, N. Palanichamy, S. C. Haw, and K. W. Ng, "Sentiment analysis of Amazon product reviews by supervised machine learning models," Journal of Advances in Information Technology, vol. 14, no. 4, pp. 857–862, 2023. DOI: https://doi.org/10.12720/jait.14.4.857-862.

[29] G. Chauhan, A. Sharma, and N. Dwivedi, "Amazon product reviews sentiment analysis using machine learning," in 2024 IEEE International Conference on Computing, Power, and Communication Technologies (IC2PCT), 2024. DOI: https://doi.org/10.1109/IC2PCT.2024.9900625.

DOI: https://doi.org/10.54216/FPA.180110
Received: July 01, 2024 Revised: September 25, 2024 Accepted: December 26, 2024