Journal of Computer and Communications, 2025, 13(1), 136-151
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Analysing Effectiveness of Sentiments in Social Media Data Using Machine Learning Techniques

Thambusamy Velmurugan[1], Mohandas Archana[2], Ajith Singh Nongmaithem[3]

[1]PG & Research Department of Computer Science, D.G.Vaishnav College, Chennai, India
[2]PG Department of IT & BCA, D.G.Vaishnav College, Chennai, India
[3]Department of Computer Science, Souht East Manipur College, Manipur, India
Email: velmurugan_dgvc@yahoo.co.in, archanadgvc@gmail.com, ajithex@gmail.com

How to cite this paper: Velmurugan, T., Archana, M. and Ajith Singh, N. (2025) Analysing Effectiveness of Sentiments in Social Media Data Using Machine Learning Techniques. *Journal of Computer and Communications*, **13**, 136-151.
https://doi.org/10.4236/jcc.2025.131010

Received: January 8, 2025
Accepted: January 28, 2025
Published: January 31, 2025

Abstract

Every second, a large volume of useful data is created in social media about the various kind of online purchases and in another forms of reviews. Particularly, purchased products review data is enormously growing in different database repositories every day. Most of the review data are useful to new customers for theier further purchases as well as existing companies to view customers feedback about various products. Data Mining and Machine Leaning techniques are familiar to analyse such kind of data to visualise and know the potential use of the purchased items through online. The customers are making quality of products through their sentiments about the purchased items from different online companies. In this research work, it is analysed sentiments of Headphone review data, which is collected from online repositories. For the analysis of Headphone review data, some of the Machine Learning techniques like Support Vector Machines, Naive Bayes, Decision Trees and Random Forest Algorithms and a Hybrid method are applied to find the quality via the customers' sentiments. The accuracy and performance of the taken algorithms are also analysed based on the three types of sentiments such as positive, negative and neutral.

Keywords

Support Vector Machine, Random Forest Algorithm, Naive Bayes Algorithm, Machine Learning Techniques, Decision Tree Algorithm

1. Introduction

Social media and digital platforms produce massive amounts of data every second due to the growth of online shopping and e-commerce, gathering consumer reviews and opinions about a wide variety of products. These reviews, which offer

important insights into customer sentiment and product quality, are increasingly being kept in online repositories and databases. This review data is an essential tool for businesses looking to better understand customer perceptions and enhance their products, as well as for potential customers. Effective analysis of this data necessitates sophisticated data mining and machine learning techniques, which can reveal patterns and trends in user feedback that would be challenging to find by hand. One popular use of data mining and machine learning is sentiment analysis, which aims to evaluate the thoughts and feelings that customers express in their reviews. Businesses can determine the perceived quality and usability of their products by analyzing customer sentiment, which is frequently divided into three categories: positive, negative, and neutral.

In order to automatically classify and predict sentiments based on review data, Machine Learning Techniques like Support Vector Machines (SVM), Naive Bayes (NB), and Random Forest (RF) are crucial to this analysis. In order to categorize the sentiments that customers express, to apply a number of machine learning algorithms to sentiment analysis of headphone review data that was gathered from online repositories.

The architecture for processing and evaluating customer reviews to ascertain sentiment polarity, particularly in the Boat Headphone dataset gathered from Flipkart, is depicted in **Figure 1**. Customer reviews are the first step in the process,

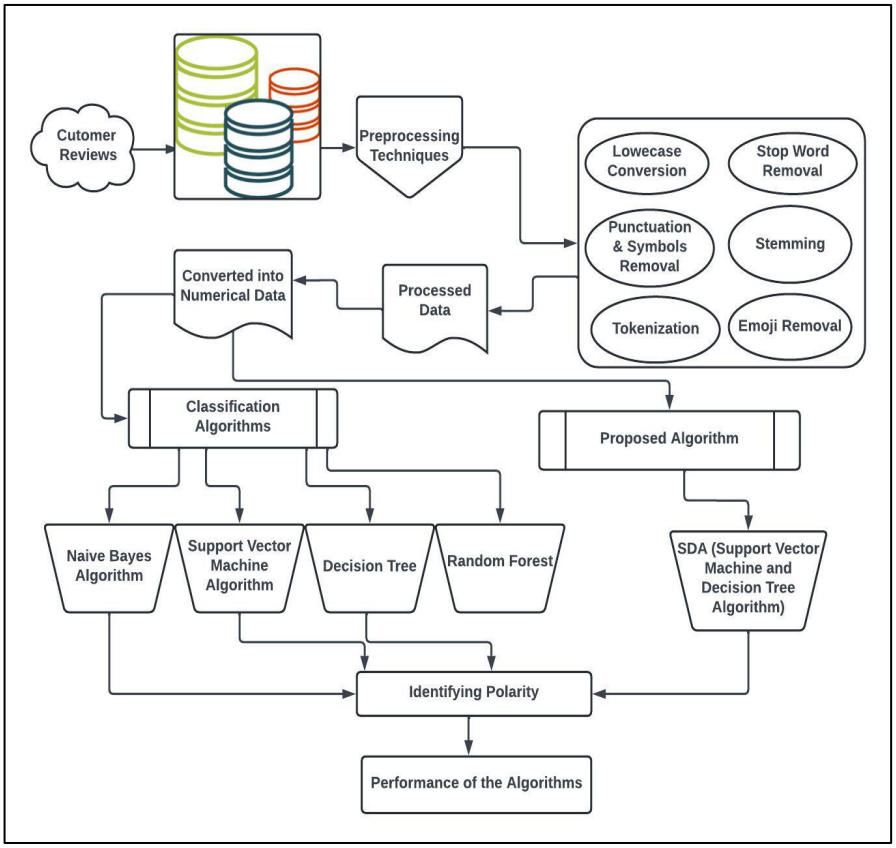

Figure 1. Architecture of research work.

and they are prepared by using preprocessing techniques. Lowercase conversion, stop word removal, punctuation and symbol removal, stemming, tokenization, and emoji removal are some of these methods. Following preprocessing, the data is converted into a numerical format so that machine learning algorithms can use it.

The data is fed into different classification algorithms for sentiment analysis after it has been processed. Among these algorithms are Random Forest, Decision Tree, Support Vector Machine (SVM), and Naive Bayes. A Proposed Algorithm, in this case a hybrid model known as SDA (Support Vector Machine and Decision Tree Algorithm), is presented alongside these conventional algorithms. Every algorithm determines whether customer reviews are positive, negative, or neutral by classifying their polarity.

The organization of the research work is chapter 2 discusses the literature review, chapter 3 describes the dataset, chapter 4 discusses the Materials and Methods, chapter 5 presents the experimental results, and chapter 6 discusses the conclusion are as follows.

2. Literature Survey

A literature review offers a thorough summary of the body of knowledge and theoretical frameworks pertaining to a particular topic, which forms the critical basis of scholarly research. As a result, a thorough literature review is crucial for determining the significance and novelty of new research as well as providing a critical lens through which current knowledge is evaluated and expanded upon.

A research work carried out by Loukili *et al.* in [1], in which that the artificial intelligence methods like Machine Learning and Natural Language Processing determine the results of various algorithms, including KNN, Random Forest, Logistic Regression, and CatBoost Classifier, indicate that LR is the model with the highest accuracy, scoring a 0.900 (or 90%). Another research carried out by Mujawar *et al.* in [2], in which that the sentiment analysis methods work when used on user reviews of wireless earphones from the Indonesian online retailer Tokopedia. The results show that the Naïve Bayes classifier's superior performance across several evaluation metrics, it is determined to be the best method overall.

A research paper titled as "A combined approach of sentimental analysis using machine learning techniques", done by Gupta *et al.* in [3], in which accuracy of more than 78%, the Random Forest classifier is shown to be the best-performing approach among the models tested and the most useful model for sentiment analysis in this work. Another research work carried out by Elangovan, Durai, and Varatharaj Subedha in [4], in which the Deep Belief Network (DBN) is used for sentiment classification in the suggested technique. The APGWO-DLSA approach proved to be the most effective method in the research after a series of tests showed its superior performance, reaching a maximum accuracy of 94.77% on the Cell Phones and Accessories (CPAA) dataset and 85.31% on the Amazon Products (AP) dataset.

A research work titled as "Sentiment analysis and fake amazon reviews classification using SVM supervised machine learning model" carried out by Tabany, Myasar, and Meriem Gueffal [5], in which that the SVM model acquires 70% of accuracy and it is superior to Naive Bayes, Logistic Regression, and Random Forest classifiers. The SVM's performance was further enhanced through hyperparameter tuning, which led to 93% sentiment analysis accuracy.

Reviews of the literature give a succinct overview of previous studies and shed light on the state of knowledge today. It helps to direct the development of research questions and assist in identifying research gaps. They determine the significance and background of the new research by examining earlier studies. They also bolster methodological decisions and raise the research's legitimacy.

3. Description of the Dataset

The Boat_Headphone Flipkart dataset 9977 instances, which was obtained from Kaggle and has attributes like user reviews and ratings, is shown in **Figure 2**. The dataset is split into training and testing subsets in order to assess the effectiveness of sentiment prediction models. 2019 examples make up the testing set, which is used to evaluate how well the model generalizes to previously undiscovered data. This division enables a thorough assessment of the model's efficacy and accuracy in gauging sentiments from user reviews and ratings.

review	rating
It was nice produt. I like it's design a lot. It's easy to carry. And. Looked stylish.READ MORE	5
awesome sound....very pretty to see this nd the sound quality was too good I wish to take this product loved this product 😍😍😍READ MORE	5
awesome sound quality. pros 7-8 hrs of battery life (including 45 mins approx call time)Awesome sound output. Bass and treble are really very clear without equaliser. With equaliser, sound wary depends on the handset sound quality.Weightless to carry and in head tooMic is good, but in traffic it is not too good (3.25/5)3.5mm Option is really important to mention. Really expecting other leading brands to implement this.ConsVery tight in ears. adjusters are ok .. this ll be very tight...READ MORE	4
I think it is such a good product not only as per the quality but also the design is quite good . I m using this product from January ... In this pandamic situation it has became the most useful and helpful . Overall the bass and the sound quality is pretty good and another thing that will give you such a sigh of relief that it will provide a wire that will help you in case of lacking charges.READ MORE	5
awesome bass sound quality very good bettary long life and I have a purchase Rs.999 only really grateful product don't forget to likeREAD MORE	5
Awsome sound powerful bass battery backup is also excellent and i loved bass the most and I'm huge lover of music and the most bass design and build is also very nice,and this was my first time when i bought and headphone or and electronics item and glad to say on the first time only i loved it 😍 😍 😍 it's a very excellent product. Thanks boat for making such a nice product keep making such and the most Thanks to flipkart team for delivering such a nice product by risking there lif...READ MORE	5
This product sound is clear and excellent bass. Obviously this is a good product valuable from money. so guys any no daut this is a really good product.....READ MORE	4
Should u buy this---Pros:-1. Sound quality and build quality is awesome2. Bluetooth connectivity is average3. Bass is clear and High 😍4. Battery Backup is very good5. U can use this as wire Headphone when u have no charge on headphone..(This is good)Cron:-1. If u use it for too long u can feel the pain in ur ear2. its on the ear not over the ear...if u need over the ear at this price go for Moto puls maxFinal opinion:---If u want good sound quality.. good bass and a wireless head...READ MORE	4
First of all, I want to talk about sound quality. The sound quality is best for this price segment. Bass is really punchy, mids are also good, but on high volumes, the sound cracks a bit, otherwise it is really nice. 9/10.Then come the build quality. The build quality too is amazing. You cant get better headphones than these. They are sturdy and very light, made out of very good quality plastic. 9.5/10.The thing I didn't like was that they are a bit tight. Sometimes, they become uncomforta...READ MORE	5
Good looking Super Fine clear Sound and power full bassREAD MORE	5

Figure 2. Sample Dataset of Boat Headphone Reviews

4. Methods and Materials

A systematic approach to sentiment analysis in customer reviews is used in this research Methods and Materials, with a focus on the Boat Headphone dataset that was gathered from Flipkart.

4.1. Preprocessing Methods

Text mining relies heavily on preprocessing techniques because they improve the quality and suitability of the data for analysis [6]. These methods help to eliminate noise and inconsistencies from text by standardizing and cleaning it, making it possible to derive more precise and insightful conclusions. **Figure 3** illustrates the steps involved in preparing the text data for further analysis.

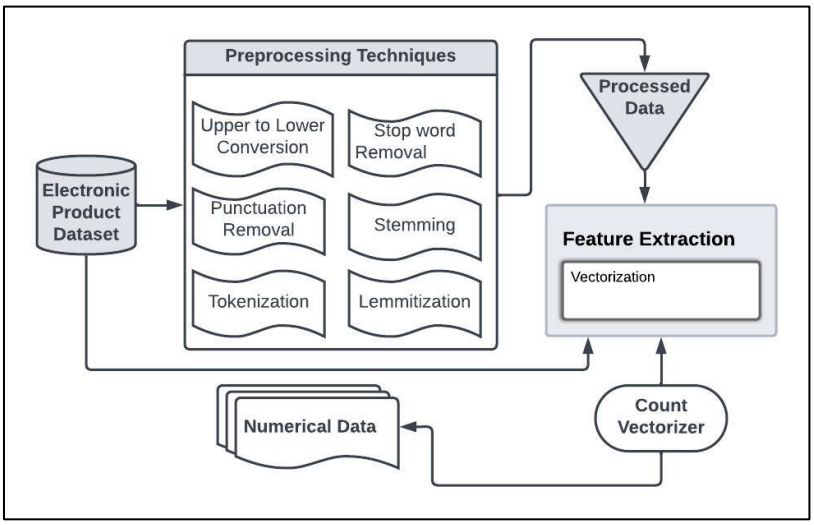

Figure 3. Preprocessing techniques.

***Lowercase Review*:** In this stage, all of the reviews' text is changed to lowercase. It lessens variability and enhances consistency in text processing by helping to standardize the text and ensuring that terms like "Excellent" and "excellent" are treated as the same term by the analysis.

$$f(x) = ax^2 + bx + c \tag{1}$$

Here $f(x)$ is a function of x, and a, b and c are constants. This is the quadratic equation where the lowercase letters

***Stopwords*:** Stopwords are frequently eliminated from texts because they don't significantly add meaning to sentiment analysis. Examples of these words are "and," "the," and "is." Eliminating these stopwords helps the model concentrate on more significant terms by lowering noise in the data.

$$D' = D \setminus S \tag{2}$$

where D' is the resulting document after removing all words that belong to the stopword set S. \ denotes the set difference operation, remove from D' all the elements are also in S. This yields document $D' = \{\omega' \in D \mid \omega' \notin S\}$, containing only the words from D that are not in the stopword set S.

***Review of the Tokenized Text*:** Tokenization divides the text into tokens. This process is essential to convert continuous text into tokens that can be examined and used as input for machine learning models.

T be a text sequence of characters $\{C_1, C_2, ..., C_n\}$

(T) be the tokenization function that splits T into words or meaningful units.

$$f_{token}(T) = \{\omega_1, \omega_2, \ldots, \omega_n\} \tag{3}$$

Review of stemmed words: A stemmed word is reduced to its base or root form. As an illustration, "running" could be stemmed to "run." By combining various word forms into a single representation, this technique can help the model identify and analyze related terms more effectively [7].

The stemming process applied to the entire set of words W is then

$$W' = \{f_{stem}(\omega_1), f_{stem}(\omega_2), \ldots, f_{stem}(\omega_n)\} = \{\omega_1, \omega_2, \ldots, \omega_n\} \tag{4}$$

Here W' represents the set of stemmed words, where each word ω_i has been transformed into root from ω_i by the function f_{stem}.

Lemmatized Review: By taking into account the context and meaning of the word, lemmatization reduces words to their base or root form more precisely than stemming does. Lemmatized terms like "better" would be "good." Text analysis and sentiment prediction are improved by this method's more accurate text normalization.

Let

$$W = \{\omega_1, \omega_2, \ldots, \omega_n\} \tag{5}$$

For each ω_i in W

$$f_{lemma}(\omega_i, POS(\omega_i)) = \omega_i \tag{6}$$

Applying lemmatization to the entire set of words W yields

$$\begin{aligned} W' &= \{f_{lemma}(\omega_1, POS(\omega_1)), f_{lemma}(\omega_2, POS(\omega_2)), \ldots, f_{lemma}(\omega_n, POS(\omega_n))\} \\ &= \{\omega_1, \omega_2, \ldots, \omega_n\} \end{aligned} \tag{7}$$

Together, these preprocessing procedures standardize, clean, and condense the text data into a format that is suitable for sentiment analysis and other natural language processing applications. The hybrid and Machine Learning approaches were used in this research to assess the effectiveness of the methods and forecast the sentiments expressed in customer reviews. These strategies include more sophisticated approaches like hybrid models that combine several techniques, as well as more conventional ones like Support Vector Machines and Naïve Bayes algorithms. The objective was to evaluate these techniques' performance through comparative analysis and ascertain how well they classified sentiments.

4.2. Machine Learning Algorithms

By identifying patterns in the data and categorizing text into positive, negative, and neutral categories, machine learning algorithms are essential for predicting polarities in sentiment analysis tasks. Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) are frequently employed algorithms for this purpose. Each of these algorithms makes use of a different

strategy: DT and RF create decision rules based on feature values, SVM determines the best hyperplane for classification, and NB employs probabilistic techniques.

4.2.1. Naïve Bayes Algorithm

Based on the premise that the features used for classification are conditionally independent given the class label, Naïve Bayes is a probabilistic classification algorithm. Because of its simplicity, it can function well in high-dimensional spaces and with little data [8]-[10]. Because of its efficiency and effectiveness in handling large datasets, Naïve Bayes often delivers robust performance in various applications, like text classification and spam filtering, despite its "naïve" assumption that rarely holds true in practice.

The Naïve Bayes algorithm aims to find the class C that maximizes the posterior probability $P(c \mid X)$ using Bayes theorem

$$P(c \mid X) = \frac{P(X \mid c).P(c)}{P(X)} \tag{8}$$

since $P(X)$ is constant for all classes

$$c = \arg\max \mathsf{C}_{\neq C} P(X \setminus c).P(c) \tag{9}$$

To compute $P(X \mid c)$, the Naive Bayes assumption assumes that the features are conditionally independent given the class.

$$P(X \mid c) = \prod_{i=1}^{n} P(x_i \mid c) \tag{10}$$

Its primary benefits are its simplicity of use and speed in producing probabilistic predictions, which make it a preferred option for a variety of classification tasks.

Thus, the equation for predicting the class c is

$$c = \arg\max P(c) \prod_{i=1}^{n} P(x_i \mid c) \tag{11}$$

This equation forms the Naïve Bayes classifier.

4.2.2. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates data points of different classes in a high-dimensional space. The optimal hyperplane maximizes the margin, or distance, between the closest data points of each class, known as support vectors [11] [12]. SVM is effective in handling both linear and non-linear classification problems through the use of kernel functions, which map data into higher dimensions to make it linearly separable [13] [14].

The distance between the hyperplane and the closest data points is called the margin

$$y_i (\omega.x_i + b) \geq 1 \quad \text{for all} \quad i \tag{12}$$

Thus SVM optimization problem can be formulated as follows,

$$\min \frac{1}{2}\|\omega\|^2 \tag{13}$$

Subject to

$$y_i\left(\omega.x_i + b\right) \geq 1, \text{ for all } \; i \tag{14}$$

Once ω and b are determined, a new data points x can be classified based on the sign

$$f\left(x\right) = \omega.x + b \tag{15}$$

The predicted class for x

$$Class = sign\left(f\left(x\right)\right) = \begin{cases} +1 \, if \; f\left(x\right) > 0 \\ -1 \, fi \; f\left(x\right) < 0 \end{cases} \tag{16}$$

In this way, SVM classifies new data points by finding which side of the hyperplane.

4.2.3. Decision Tree Algorithm

A supervised learning technique used for both regression and classification problems are the decision tree algorithm. It builds a model in the shape of a tree structure, with each internal node standing for a feature-based decision, each branch for the decision's result, and each leaf node for the final classification or prediction [15]-[17].

For a node with classes $\{C_1, C_2, ..., C_k\}$ and probabilities $P\left(c_i\right)$ for each class c_i.

$$G = 1 - \sum_{i=1}^{k} P(c_i)^2 \tag{17}$$

Measure the randomness in the information being processed. For a node with classes $\{C_1, C_2, ..., C_k\}$

$$H = 1 - \sum_{i=1}^{k} P\left(c_i\right) \log_2 P\left(c_i\right) \tag{18}$$

For a dataset D with entropy $H\left(D\right)$ and a split on feature A resulting in subsets $D_1, D_2, \ldots, D_n$ the information gain IG for feature A is

$$IG\left(D, A\right) = H\left(D\right) - \sum_{j=1}^{n} \frac{D_j}{D} H\left(D_j\right) \tag{19}$$

Recursively dividing the dataset into subsets based on feature values that best separate the data in accordance with a criterion—such as information gain or Gini impurity—builds the tree. Decision trees are helpful for comprehending the decision-making process because they are simple to interpret and visualize.

4.2.4. SDA Hybrid Algorithm (Hybrid Algorithm)

The Novel Random Decision Algorithm, also known as the SDA Hybrid Algorithm, was created to improve sentiment analysis of customer reviews. To increase classification accuracy, this technique combines elements with decision-making procedures [18]. In order to capture various facets of the sentiment landscape, the SDA Hybrid Algorithm combines multiple decision trees, each trained on

randomly selected subsets of the data. With the use of both traditional algorithms' and structured decision-making, SDA seeks to build a strong model that can manage the complexity and variability of customer reviews. This hybrid method provides a more accurate and nuanced sentiment prediction from customer feedback by reducing over fitting and enhancing generalization [19] [20].

Apply Decision Tree to segment X based on optimal splits

$$S = \{S_1, S_2, \ldots, S_k\} = f_{DT}(X) \tag{20}$$

where each $S_i \subset X$ represents a subset of data points

$$y_i = f_{SVM}(S_i) \tag{21}$$

The overall prediction for the input features X is the combination of predictions from subset

$$y = \bigcup_{i=1}^{k} y_i \tag{22}$$

5. Results and Discussions

The experimental findings in this research shed light on how well different hybrid and machine learning approaches perform in predicting sentiment from customer reviews [15] [16].

Figure 4 represents the outcomes of various preprocessing techniques applied

review	review_lower	review_wo_punct	review_wo_stop	review_token	review_lemmatized
sound quality good enough considering that I got it for 1000. some might find it too tight over the ears usually ones with bigger heads it's fine for me though.Didn't feel deep bass but rocking nonetheless.READ MORE	sound quality good enough considering that i got it for 1000. some might find it too tight over the ears usually ones with bigger heads it's fine for me though.didn't feel deep bass but rocking nonetheless.read more	sound quality good enough considering that i got it for 1000 some might find it too tight over the ears usually ones with bigger heads its fine for me thoughdidnt feel deep bass but rocking nonethelessread more	sound quality good enough considering got 1000 might find tight ears usually ones bigger heads fine thoughdidnt feel deep bass rocking nonethelessread	['sound', 'quality', 'good', 'enough', 'considering', 'got', '1000', 'might', 'find', 'tight', 'ears', 'usually', 'ones', 'bigger', 'heads', 'fine', 'thoughdidnt', 'feel', 'deep', 'bass', 'rocking', 'nonethelessread']	['sound', 'quality', 'good', 'enough', 'considering', 'got', '1000', 'might', 'find', 'tight', 'ear', 'usually', 'one', 'bigger', 'head', 'fine', 'thoughdidnt', 'feel', 'deep', 'bass', 'rocking', 'nonethelessread']
Very good product.i got it as price 1275 RS with phone pay discount. Product is awesome..loud bass with clarity. Fit to our ear. But it Bluetooth range is smaller than said in description i.e. 6 to 8 m. Other good delivery n packaging very good. battery backup also awesome I am not charging since bought this product.just checked that charging issue find or not. Great product....go for it.READ MORE	very good product.i got it as price 1275 rs with phone pay discount. product is awesome..loud bass with clarity. fit to our ear. but it bluetooth range is smaller than said in description i.e. 6 to 8 m. other good delivery n packaging very good. battery backup also awesome i am not charging since bought this product.just checked that charging issue find or not. great product....go for it.read more	very good producti got it as price 1275 rs with phone pay discount product is awesomeloud bass with clarity fit to our ear but it bluetooth range is smaller than said in description ie 6 to 8 m other good delivery n packaging very good battery backup also awesome i am not charging since bought this productjust checked that charging issue find or not great productgo for itread more	good producti got price 1275 rs phone pay discount product awesomeloud bass clarity fit ear bluetooth range smaller said description ie 6 8 good delivery n packaging good battery backup also awesome charging since bought productjust checked charging issue find great productgo itread	['good', 'producti', 'got', 'price', '1275', 'rs', 'phone', 'pay', 'discount', 'product', 'awesomeloud', 'bass', 'clarity', 'fit', 'ear', 'bluetooth', 'range', 'smaller', 'said', 'description', 'ie', '6', '8', 'good', 'delivery', 'n', 'packaging', 'good', 'battery', 'backup', 'also', 'awesome', 'charging', 'since', 'bought', 'productjust', 'checked', 'charging', 'issue', 'find', 'great', 'productgo', 'itread']	['good', 'producti', 'got', 'price', '1275', 'r', 'phone', 'pay', 'discount', 'product', 'awesomeloud', 'bass', 'clarity', 'fit', 'ear', 'bluetooth', 'range', 'smaller', 'said', 'description', 'ie', '6', '8', 'good', 'delivery', 'n', 'packaging', 'good', 'battery', 'backup', 'also', 'awesome', 'charging', 'since', 'bought', 'productjust', 'checked', 'charging', 'issue', 'find', 'great', 'productgo', 'itread']
Pros-- Excellent Battery Backup (nearly 7 hrs)- Good Bass - All in One - Wired and Wireless, Built in Mic, Play/Pause, Receive/End Call- Good Build Quality- Worth the Price...I got it at ₹699(₹300 Cashback from PhonePe) only...Cons-- Bass is nothing until using Bass Boost.- Worst Sound Quality when connecting to 3.5 mm jack- Need to Speak loudly	pros-- excellent battery backup (nearly 7 hrs)- good bass - all in one - wired and wireless, built in mic, play/pause, receive/end call- good build quality- worth the price...i got it at ₹699(₹300 cashback from phonepe) only...cons-- bass is nothing until using bass boost.- worst sound quality when connecting to 3.5 mm jack- need to speak	pros excellent battery backup nearly 7 hrs good bass all in one wired and wireless built in mic playpause receiveend call good build quality worth the pricei got it at ₹699₹300 cashback from phonepe onlycons bass is nothing until using bass boost worst sound quality when connecting to 35 mm jack need to speak loudly while	pros excellent battery backup nearly 7 hrs good bass one wired wireless built mic playpause receiveend call good build quality worth pricei got ₹699₹300 cashback phonepe onlycons bass nothing using bass boost worst sound quality connecting 35 mm jack need speak	['pros', 'excellent', 'battery', 'backup', 'nearly', '7', 'hrs', 'good', 'bass', 'one', 'wired', 'wireless', 'built', 'mic', 'playpause', 'receiveend', 'call', 'good', 'build', 'quality', 'worth', 'pricei', 'got', '₹699₹300', 'cashback', 'phonepe', 'onlycons', 'bass', 'nothing', 'using', 'bass', 'boost', 'worst', 'sound', 'quality', 'connecting', '35', 'mm', 'jack', 'need', 'speak',	['pro', 'excellent', 'battery', 'backup', 'nearly', '7', 'hr', 'good', 'bass', 'one', 'wired', 'wireless', 'built', 'mic', 'playpause', 'receiveend', 'call', 'good', 'build', 'quality', 'worth', 'pricei', 'got', '₹699₹300', 'cashback', 'phonepe', 'onlycons', 'bass', 'nothing', 'using', 'bass', 'boost', 'worst', 'sound', 'quality', 'connecting', '35', 'mm', 'jack', 'need', 'speak',

Figure 4. Results of preprocessing techniques.

to the text data. It highlights the impact of each technique on the quality of the data before it is fed into analytical models. The preprocessing steps listed—such as lowercasing, cleaning, stopword removal, tokenization, stemming, and lemmatization—are evaluated based on specific metrics, which might include text clarity, data consistency, and model performance improvements.

Table 1 displays the word count by using vectorization method, one can see how frequently particular words appear in the dataset and how frequently they appear in customer reviews. Words with corresponding counts, such as "good," "sound," "product," and "quality," are listed in the table. For instance, "good" is the most frequently occurring word with 4276 occurrences. It is followed by "sound" with 2827 occurrences and "product" with 2658.

Table 1. Frequency of words in customer reviews

Words	Count	Words	Count
Good	5325	**Best**	1143
Sound	2334	**Read**	1023
Product	2764	**Product**	1032
quality	2643	**Awesome**	976
Ear	912	**Battery**	956
Price	730	**Headphone**	1674

While Naïve Bayes shows competitive performance, particularly in the Positive category, it performs less well when it comes to classifying sentiments that are neutral. Understanding the terms that customer use most frequently in their reviews can be greatly aided by looking at this figure, which offers insights into important themes and sentiments shown in **Figure 5**.

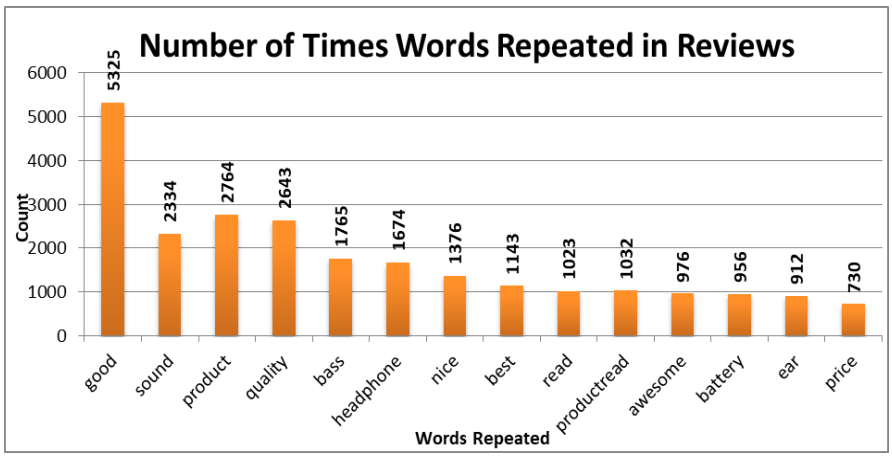

Figure 5. Frequency of words repeated in reviews.

These steps were improved performance and accuracy in sentiment analysis by preparing the dataset for the application of algorithms such as Naive Bayes, SVM,

Random Forest, and hybrid models.

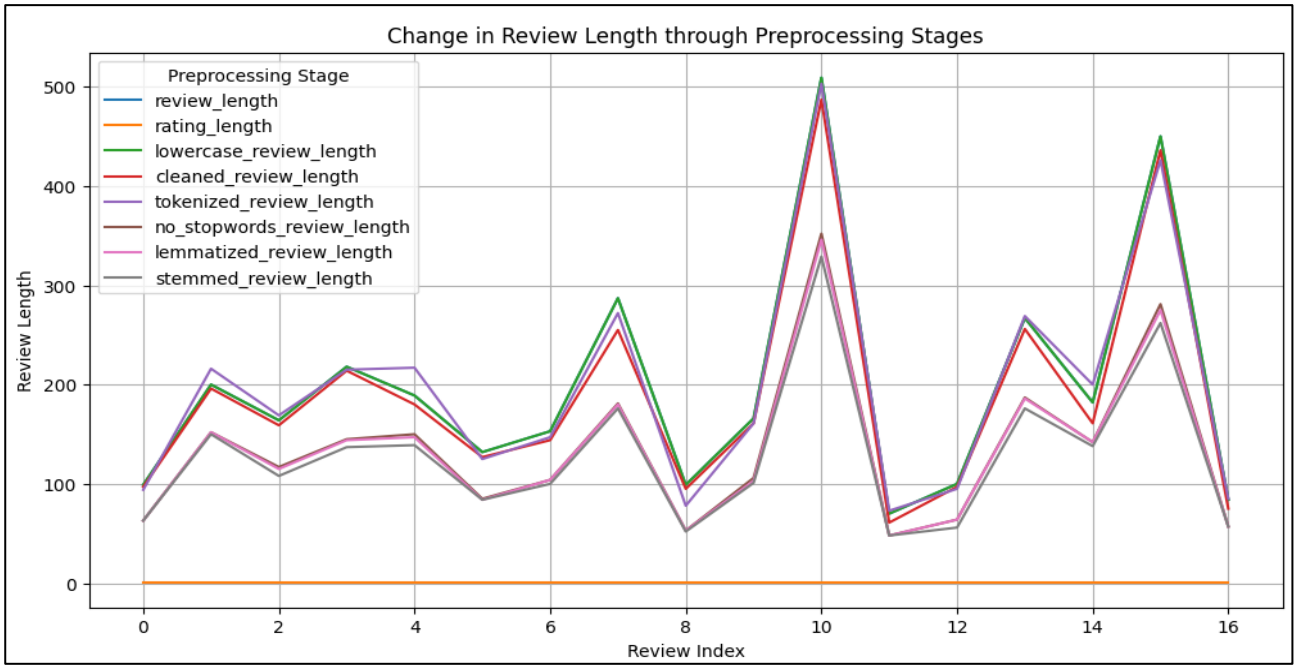

Figure 6. Length of reviews in different phase.

The lengths of the reviews before lowercasing, cleaning, tokenization, stopword removal, lemmatization, and stemming are displayed in the columns along with the rating and original review lengths. The impact of each preprocessing step on the text data is displayed in **Figure 6**.

Table 2 represents the review text's initial length in terms of characters or words before any preprocessing was done. These metrics enable a thorough comparison of each preprocessing method's efficacy. It aids in quantifying the amount of redundancy or noise eliminated at each step. The text is optimized for tokenization, feature extraction, and classification algorithms by offering insights into how the dataset is transformed for machine learning models.

Table 3 represents a comparison of the sentiment polarity identification performance of four distinct classification algorithms: Decision Tree, Naïve Bayes, Support Vector Machine (SVM), and SDA (a hybrid algorithm). Precision, recall, and F1-score for three sentiment categories—Negative, Neutral, and Positive—as well as the support (number of samples) for each category are used to assess each algorithm's performance.

The existing algorithms and hybrid algorithm for sentiments extraction form the customer reviews using Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and the SDA hybrid algorithm counts (positive, neutral, and negative) are displayed in **Figure 7**.

In determining the sentiment polarity (negative, neutral, and positive) of a dataset, this table shows the effectiveness of a number of machine learning algorithms, including Decision Tree, Naïve Bayes, Random Forest, Support Vector

Table 2. Length of reviews in each preprocessing phase.

reviewlength	ratinglength	Lowercase review length	Cleaned review length	Tokenized review length	Stopwords Review length	lemmatized review_length	Stemmed review length
99	1	99	97	94	63	63	63
200	1	200	196	216	152	152	150
164	1	164	159	169	117	115	108
218	1	218	214	215	145	144	137
189	1	189	180	217	150	147	139
132	1	132	127	125	85	84	84
153	1	153	144	147	104	104	100
287	1	287	255	272	181	180	176
99	1	99	95	78	53	53	52
166	1	166	161	161	106	103	101
509	1	509	487	503	352	346	329
70	1	70	61	73	48	48	48
100	1	100	97	95	64	64	56
267	1	267	256	269	187	186	176
182	1	182	161	200	142	142	138
450	1	450	436	426	281	275	262
84	1	84	75	85	57	57	57

Table 3. Polarities predicted by algortihms.

Sentiments	NB	RF	DT	SVM	SDA
Positive	1743	1801	1754	1758	1832
Neutral	264	207	238	251	165
Negative	12	11	27	10	22

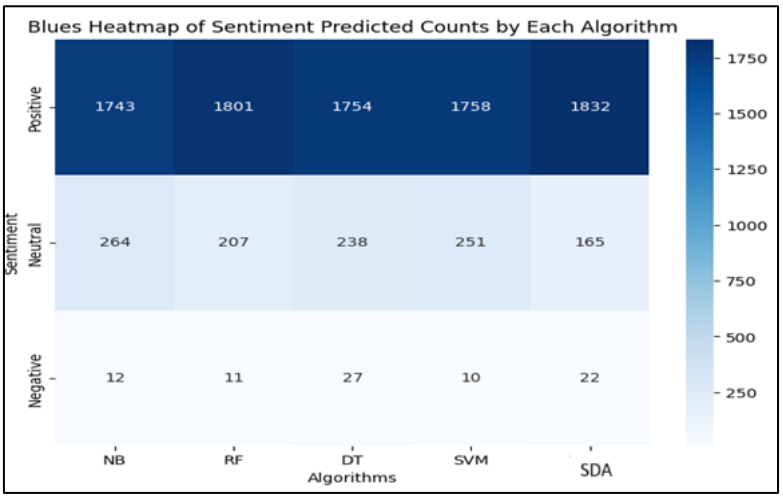

Figure 7. Polarities Predicted by Algorithms.

Table 4. Performance analysis of algorithms.

Decision Tree Algorithm			
Polarity Identification	**precision**	**recall**	**f1-score**
Negative	0.9	0.75	0.82
Neutral	0.75	0.6	0.67
Positive	0.85	0.93	0.89
Naïve Bayes			
Negative	0.85	0.7	0.77
Neutral	0.75	0.6	0.67
Positive	0.9	0.95	0.92
Random Forest			
Negative	0.83	0.5	0.68
Neutral	0.78	0.7	0.73
Positive	0.87	0.93	0.9
Support Vector Machine			
Negative	0.85	0.73	0.79
Neutral	0.72	0.57	0.64
Positive	0.85	0.89	0.87
SDA Algorithm (Hybrid Algorithm)			
Negative	0.98	0.98	0.98
Neutral	0.95	0.9	0.92
Positive	0.96	0.96	0.96

Machine, and the hybrid SDA algorithm (Support Vector Machine + Decision Tree). The precision, recall, F1-score, and support metrics are used to evaluate each model's performance, as shown in **Table 4**. Considered the other algorithms, the SDA (hybrid) algorithm is the best model for identifying sentiment polarity in this dataset shown in **Figure 8**. All sentiment classes receives nearly flawless scores on every metric, indicating that it strikes the best balance between accuracy and dependability when determining sentiment polarities.

SVM and Naïve Bayes perform reasonably well, while Random Forest and Decision Tree have lower recall, particularly for negative and neutral sentiments. The accuracy of several algorithms in a classification task is shown in the **Table 5**. With an accuracy of 72%, the Naïve Bayes algorithm performed moderately. With an accuracy of 82%, the Support Vector Machine (SVM) outperformed the others, indicating a strong capacity for accurate data classification.

The decision tree algorithm yields 78.20% accuracy. It was marginally less compared with SVM, but it was still quite good. At an amazing accuracy of 96.01%, the suggested approach—dubbed SDA—significantly outperformed the other algorithms.

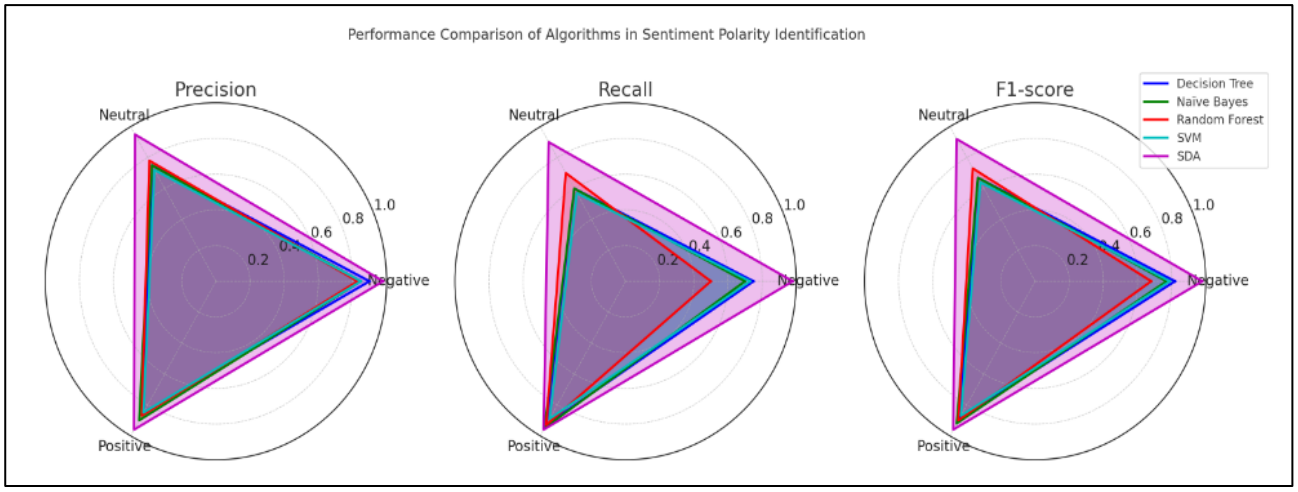

Figure 8. Comparison of Predicting Sentiments by Algorithms

Table 5. Accuracy of algorithms

Algorithms	Accuracy %
Naïve Bayes	72.0
Support Vector Machine	82.0
Random Forest	84.1
Decision Tree	78.2
SDA (Proposed method)	96.0

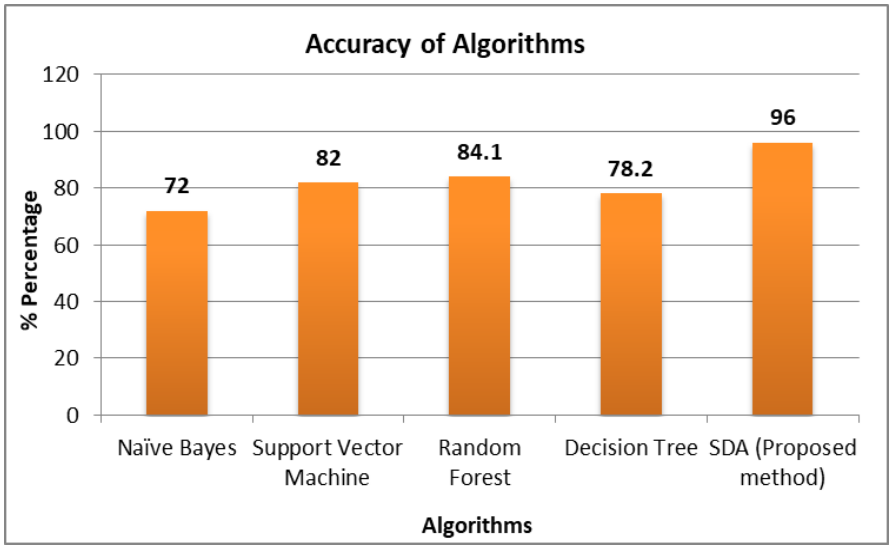

Figure 9. Accuracy of algorithms

The substantially of higher accuracy is 96.01%, the SDA (Proposed method) outperformed the traditional algorithms by a wide margin. **Figure 9** shows that the suggested approach is very good at identifying the underlying patterns in the data, which results in predictions that are more accurate.

6. Conclusion

Customer review data posted in social media are usefull to know the products quality and further analysis. In order to find the performance of machine learning algorithms, the boat headphone reviews are given as input to thee chosen algorithms. The existing machine learning algorithms Naïve Bayes, Support Vector Machine, Random Forest and Decision Tree and a hybrid algorithm namely SDA are utilised to find the performance and efficiency of the algorithms in terms of its precision, recall, f1-score as well as the accuracy. The polarities of the chosen dataset are identified in order to find the sentiments of the customer reviews. From the experimental results of this approach, it is found that the performance of the hybrid algorithm is better than the other existing algorithms. Hence, it is concluded that the hybrid algorithm yields better results and analysed the customer reviews for sentiments. In future some of the other machine learning algorithms are applied in the same procedure to find the sentiments as well as accuracy using different kinds of customer reviews data.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Manal, L., Messaoudi, F. and El Ghazi, M. (2023) Sentiment Analysis of Product Reviews for E-Commerce Recommendation Based on Machine Learning. *International Journal of Advances in Soft Computing & Its Applications*, **15**, 1-13. https://doi.org/10.15849/IJASCA.230320.01

[2] Adel Ramadhan, F., Permana Ruslan, R.R. and Zahra, A. (2023) Sentiment Analysis of E-Commerce Product Reviews for Content Interaction Using Machine Learning. *Cakrawala Repositori IMWI*, **6**, 207-220. https://doi.org/10.52851/cakrawala.v6i1.219

[3] Gupta, K., Jiwani, N. and Afreen, N. (2023) A Combined Approach of Sentimental Analysis Using Machine Learning Techniques. *Revue d'Intelligence Artificielle*, **37**, 1-6. https://doi.org/10.18280/ria.370101

[4] Elangovan, D. and Subedha, V. (2023) Adaptive Particle Grey Wolf Optimizer with Deep Learning-Based Sentiment Analysis on Online Product Reviews. *Engineering, Technology & Applied Science Research*, **13**, 10989-10993. https://doi.org/10.48084/etasr.5787

[5] Tabany, M. and Gueffal, M. (2024) Sentiment Analysis and Fake Amazon Reviews Classification Using SVM Supervised Machine Learning Model. *Journal of Advances in Information Technology*, **15**, 49-58. https://doi.org/10.12720/jait.15.1.49-58

[6] Başarslan, M.S. and Kayaalp, F. (2023) Sentiment Analysis with Ensemble and Machine Learning Methods in Multi-Domain Datasets. *Turkish Journal of Engineering*, **7**, 141-148. https://doi.org/10.31127/tuje.1079698

[7] Mujawar Sofiya, S. and Bhaladhare, P.R. (2023) An Aspect-Based Multi-label Sentiment Analysis using Improved BERT System. *International Journal of Intelligent Systems and Applications in Engineering*, **1**, 228-235.

[8] Dey, S., Wasif, S., Tonmoy, D.S., Sultana, S., Sarkar, J. and Dey, M. (2020) A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment

Analysis on Amazon Product Reviews. 2020 *International Conference on Contemporary Computing and Applications* (*IC3A*), Lucknow, 5-7 February 2020, 217-220. https://doi.org/10.1109/ic3a48958.2020.233300

[9] Mubarok, M.S., Adiwijaya, and Aldhi, M.D. (2017) Aspect-Based Sentiment Analysis to Review Products Using Naïve Bayes. *AIP Conference Proceedings*, Surabaya, 23 November 2016, Article 020060. https://doi.org/10.1063/1.4994463

[10] Singla, Z., Randhawa, S. and Jain, S. (2017) Sentiment Analysis of Customer Product Reviews Using Machine Learning. 2017 *International Conference on Intelligent Computing and Control* (*I2C2*), Coimbatore, 23-24 June 2017, 1-5. https://doi.org/10.1109/i2c2.2017.8321910

[11] Brownfield, S. and Zhou, J. (2020) Sentiment Analysis of Amazon Product Reviews. In: *Advances in Intelligent Systems and Computing*, Springer, 739-750. https://doi.org/10.1007/978-3-030-63319-6_68

[12] Aashutosh, B., Patel, A., Chheda, H. and Gawande, K. (2015) Amazon Review Classification and Sentiment Analysis. *International Journal of Computer Science and Information Technologies*, **6**, 5107-5110.

[13] Rathor, A.S., Agarwal, A. and Dimri, P. (2018) Comparative Study of Machine Learning Approaches for Amazon Reviews. *Procedia Computer Science*, **132**, 1552-1561. https://doi.org/10.1016/j.procs.2018.05.119

[14] Abraham, M.P. (2020) Feature Based Sentiment Analysis of Mobile Product Reviews Using Machine Learning Techniques. *International Journal of Advanced Trends in Computer Science and Engineering*, **9**, 2289-2296. https://doi.org/10.30534/ijatcse/2020/210922020

[15] Momina, S., Awan, S.M., Hussain, N. and Gondal, Z.A. (2019) Sentiment Analysis on Mobile Phone Reviews Using Supervised Learning Techniques. *International Journal of Modern Education & Computer Science*, **7**, 32-43.

[16] Salem, M.A.M. and Maghari, A.Y.A. (2020) Sentiment Analysis of Mobile Phone Products Reviews Using Classification Algorithms. 2020 *International Conference on Promising Electronic Technologies* (*ICPET*), Jerusalem, 16-17 December 2020, 84-88. https://doi.org/10.1109/icpet51420.2020.00024

[17] Yiran, Y. and Srivastava, S. (2019) Aspect-Based Sentiment Analysis on Mobile Phone Reviews with LDA. *Proceedings of the* 2019 4*th International Conference on Machine Learning Technologies*, Nanchang, 21-23 June 2019, 101-105. https://doi.org/10.1145/3340997.3341012

[18] Kumari, U., Sharma, A.K. and Soni, D. (2017) Sentiment Analysis of Smart Phone Product Review Using SVM Classification Technique. 2017 *International Conference on Energy, Communication, Data Analytics and Soft Computing* (*ICECDS*), Chennai, 1-2 August 2017, 1469-1474. https://doi.org/10.1109/icecds.2017.8389689

[19] Dubey, T. and Jain, A. (2019) Sentiment Analysis of Keenly Intellective Smart Phone Product Review Utilizing SVM Classification Technique. 2019 10*th International Conference on Computing, Communication and Networking Technologies* (*ICCCNT*), Kanpur, 6-8 July 2019, 1-8. https://doi.org/10.1109/icccnt45670.2019.8944795

[20] Nandi, B., Ghanti, M. and Paul, S. (2017) Text Based Sentiment Analysis. 2017 *International Conference on Inventive Computing and Informatics* (*ICICI*), Coimbatore, 23-24 November 2017, 9-13. https://doi.org/10.1109/icici.2017.8365326