Journal of Big Data

RESEARCH Open Access

Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics

Kiran Chaudhary[1], Mansaf Alam[2], Mabrook S. Al-Rakhami[3*] and Abdu Gumaei[3]

*Correspondence:
malrakhami@ksu.edu.sa
[3] Research Chair of Pervasive and Mobile Computing, Information Systems Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
Full list of author information is available at the end of the article

Abstract

Social media is popular in our society right now. People are using social media platforms to purchase various products. We collected the data from various social media platforms. We analyzed the data for prediction of the consumer behavior on the social media platform. We considered the consumer data from Facebook, Twitter, Linked In and YouTube, Instagram, and Pinterest, etc. There are diverse and high-speed, high volume data which are coming from social media platform, so we used predictive big data analytics. In this paper, we have used the concept of big data technology to process data and analyze it to predict consumer behavior on social media. We have analyzed consumer behavior on social media platforms based on some parameters and criteria. We analyzed the consumer perception, attitude towards the social media platform. To get good quality of result, we pre-process data using various data preprocessing to detect outlier, noises, error, and duplicate record. We developed mathematical modeling using machine learning to predict consumer behavior on the social media platform. This model is a predictive model for predicting consumer behavior on the social media platform. 80% of data are used for training purposes and 20% for testing.

Keywords: Big data analytics, Predictive, Consumer perception, Social media, Data analytics, Consumer behaviour

Introduction

The easy way to promote the product to everyone is through the social media platform. In this paper, predictive analytics is used to find consumer behavior on the social media platform. We have proposed a mathematical and machine learning-based predictive model to find the consumer behavior towards products on the social media platform. We have validated the model; the description is given in the result and discussion section. The highest accuracy on validation of data is 98% and the transition from Interest to Instagram is 99.51%.

Social media

Social media is websites and packages that can be premeditated to let human beings go halves content bits and pieces speedy, efficiently, and in actual time. It can share images, reviews, activities, etc. in actual time and has converted the mode we go to and, as well, the mode we do an activity. Stores that use social media as an essential part of their advertising technique usually see the quantifiable penalty. But the input to successful social media is to no longer treat it as a further accessory but to indulgence it with equal care, admiration, and curiosity to you, do the entire you're advertising and marketing efforts. Marketing on social media means the use of dissimilar platforms for connecting it with your customers to shape your brand, growth of income, and density of website visitors. This needs publishing notable contents on your profiles on social media so that it can be more attentive and attractive to your followers, effects of your studying, and jogging social media classified ads. The predominant social media structures are "LinkedIn", "Facebook", "Instagram", "Twitter", "Pinterest", and "YouTube". Various examples of social media along with descriptions are given in Table 1.

In a research work [1] the researchers have discussed big data from social media that has been used as significant to important acumen into person behavior and broadly analyzed by research scholars. The researchers in their study [2], have described that how big data analytics and machine learning algorithms can help monitor social media and recognize consumers view of lavishness hotels from beginning to end, the new visual data analysis, and spin into an improved managing brand strategy for the managers of comfort hotel. In the study [3], authors have collected 8434 start-up firm's data from twitter, they have generated features based on social media, and developed a model based on machine learning for predicting the engagement level of each firm in social media. The outcome of this work describes that deep learning gives the best accuracy in engagement level prediction. The outcome of this paper also tells that tweets number by the company retweets number received, and the likes number received to have the most significance in finding social media marketing behavior usefulness. While the growing curiosity in social media and content of user-generated (UGC) on websites, such as YouTube, Facebook, and LinkedIn, social media users are seen to be contributing to the content of marketing. In this paper, we have the customer perception and assertiveness towards social media by using big data analytics. Figure 1 explains that the consumer perception and consumer attitude makes the consumer behavior which t impacts on social media. Big data analytics are used to study the overall impact of consumer behavior on social media.

We have taken data of consumers from various social media platforms. The data taken from social media platforms were dirty so we cleaned the data with data preprocessing techniques to make the data clean and quality. We have developed a model for prediction, consumer behavior towards the product. We have taken 80% data for training and 20% data for testing. We have also developed a mathematical model for consumer behavior prediction on the social media platform. The main purpose of this paper is to help the product seller to know the consumer behavior on various social media platforms regarding the product. The paper is organized as follows. The introduction is given in "Introduction" section. "Literature review" section describes the literature review and recent work. The preliminary and problem formulation is given in "Preliminary" section.

Table 1 Various examples of Social Media

S.No.	Social Media	Examples	Description
1	"Social Networks"	"Facebook, LinkedIn"	A "social networking website" is a social media page on the web which allows customers or users to knob up with persons who have analogous pastimes and upbringings. "Facebook"," Twitter," and "Instagram "are the three most popular. These three are popular instances of a "social network website"
2	"Bookmarking Sites"	"Pinterest, Flipboard, Diggs"	"Bookmarking" websites permit consumers or users to shop and position hyperlinks to somewhat online resources quantity and websites, a distinguished function of those websites which have the capability of the customers to "tag" hyperlinks, that makes them fewer difficult to look at, and continuously, the proportion with their supporters, is a popular instance of a "bookmarking "webpage
3	"Social News"	"Digg"	A "social news" website permits their customers to submission of news hyperlinks and different pieces of stuff to outside articles. Customers that continue to choose on giving matters and the gadgets with the very best quantity of vote are the maxima displayed. An amazing instance of a "social news" web page is read it
4	"Media Sharing"	"Pinterest, YouTube, Video"	Websites of "Media sharing" that permit consumers to distribute amazing media sorts along with two leading, first ones is photo sharing and second is video hosting websites. The maximum of those websites is furthermore stipulated social capabilities, impartial to the capability for making profiles and the choice of remarking at the uploaded images or motion pictures. Those structures habitually inspire consumer-generated material where every person can create, curate, and share the creativity that speaks approximately them or spark conversation
5	"Microblogging"	"Twitter, Facebook"	Those are simply websites that permit the consumer or users to submit their short-written stories, which can incorporate links to creation and provider websites, as well as hyperlinks to different "social media" websites. These are then published on the 'walls' of every person who has subscribed to that consumer's account. The maximum typically used microblogging website is Twitter
6	"Blog comments and forums"		The forum of the internet is a website that lets customers or users engage in discussions via posting and replying to messages of the community. A weblog comment on a website is an identical issue moreover being petite is additionally focused on it. The remarks are normally concentrated throughout the specific attached blog challenge

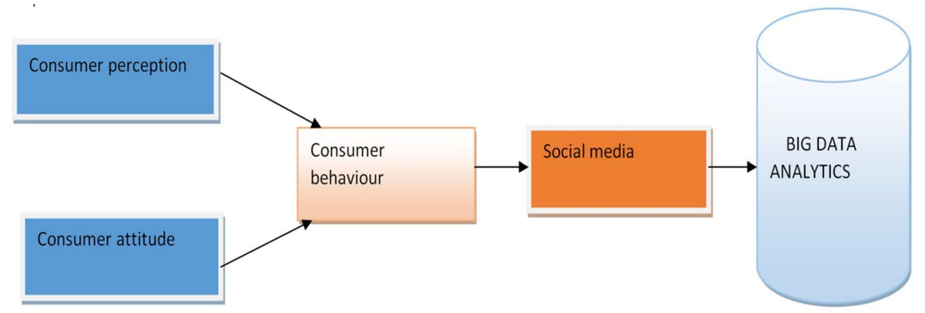

Fig. 1 Consumer perception and attitude

"Big data analytics for social media consumers" section presents big data analytics for social media consumers. "Social media consumer behaviour model" section describes the social media consumer behaviour model. "Data pre-processing" section deals with data pre-processing. The result and discussions are given in "Result and Discussion" section. The paper is finally concluded in "Conclusion" section.

Motivation

We were motivated by an article on a machine learning-based approach to enhancing social media marketing [4], in which authors have used machine learning techniques for social media marketing. They have proposed machine learning integrated social media marketing implementation and performance analysis. The weka tool is used in this work to do the data analytics. We have used python to do analytics in our paper.

Literature review

The research work [5] investigates the various predictors of helpfulness and readership of online consumer reviews using a sentiment mining approach of big data analytics. It concludes that the length and longevity of online consumer reviews have a positive relationship with its readership and helpfulness. The study [6] emphasizes big data analytics related to unstructured data that form at least 95% of the big data. They have also reviewed analytics technique for audio, video, text, and social media data and they propose and invent new tools and techniques for predictive analytics for structured data, big data are noisy, unreadable, and interrelated so there is a need to develop a new statistical technique for the data provided by social media text, audio and video [7], this paper focuses on understanding the consumer behavior in relation with different types of social media. Primarily the big data focus on the psychological aspect of predicting the consumer needs rather than understanding. In this paper, it is studied that by knowing the consumer preferences, predicting the consumer behavior and what they will buy next after purchasing the goods that helps to understand consumer perception about the brand and how to improve target advertising.

The researcher in their work [8], they have proposed, that by predicting the consumer personality from social media the products are recommended and it was also proposed the framework personality-based product recommender to analyses the personality. The framework in this paper is built on the basis of five-factor personality theory of [9] and

[10]. Social media platforms are not created equal presented by sumo heavy industries [29], according to it the active users on the various platforms are given in Table 2. The maximum users are active on Instagram and the lowest users are active on Facebook.

The research [11], in this paper they investigate the positive and negative sentiments of peoples and the reason behind their sentiments via brand authenticity. They use database of 2204 coded tweet for analysis of brand authenticity and sentiment polarity. The author examines the tweets qualitatively to know about the sentiments related to brand authenticity then this can quantitatively create a framework in that they forecast both the authenticity of brand dimensions and their polarity of sentimentality. They classified tweets based on various categories like quality, commitment, heritage uniqueness and symbolism. Latent semantic analysis (LSA) is used to extract common words in each category and the result shows higher accuracy for brand authenticity dimension prediction and their sentiment polarity. This research paper [12] is focused on the data available on social media and on the basis of reviewed papers available. They studied by highlighting the various state of art techniques and quality attributes that help in analyzing the performance of social media.

The research paper [13], the authors studied the data discovery phases, gathering, and training. In this paper, they understand the problem faced by the researcher in social media analytics i.e. problem faced before the data in analysis and also discussed a solution to these problems. They discussed the various challenges faced by the researcher while doing the studies. It can be solved by using three steps social media analysis, that is the volume of data are expecting, determine the most important part of your research, infrastructure to manage the volume and format of data, and if there is unstructured data how to extract the structured information from it. The study [14] focused on the information from the software in terms of the amount of work and studies of social media for dynamically expanded between 2009 and 2016 social network, online media, and online systems have grown and information system has also grown in the field of the human and medical sphere. In the social sciences, the information system has been more relevant to the needs of society and the issues related to these are continuously explained. The researcher in their research [15] they have focused on the events that involve the public voting. They used the data of twitter in 2015 and 2016 to examine the relationship between the voting of the audience regarding euro vision song contest and predictors based on quantity and emotions and then compared the result of using data before and during an event. They analyzed the volume of tweets and express sentiments to examine the relationship. Various research studies have used twitter as an information source to investigate the consumer's opinion regarding brands [16]; the social media

Table 2 Social Media Platforms

S. No.	Platforms	Monthly active users
1	Facebook	2 Billion
2	Twitter	328 Million
3	Instagram	800 Million
4	Pinterst	200 Million
5	Snapchat	375 Million

platform twitter is a rapid and useful mode for the company to find out, that how consumers feel regarding their business and managers [17]. Emotion is the positive expression or negative feelings of a consumer by social media with a definitive purpose [18]. The researchers demonstrated their work [19] that how knowledge from user-generated data helps in understanding and improvement in their supply chain. In this study [20] the authors have explained, that this study is expressive where an attempt is made to discovered the elements of flipkart and amazon through that the respondents are satisfied. A research work [21] in which authors have developed a score covering the position effect of social media inside. They have explained, that the score will help to analyze the inside effects on persons and company. The outcome of the research work [22] gives the social facilitation inspiration, participating and socializing inspiration, and the information inspiration that surely pressures customers' common attitudes in the direction of social networking sites and had a well-built outcome on their attitude in the direction of marketers' social networking sites. According to finance online review for business [23], the social selling is increasing that is given in Table 3.

Preliminary

The linear predictor used for prediction of social media consumer behavior. The linear predictor basic form j data point for j = 1, 2, …, n is

$$f\left(j\right) = \alpha_0 + \alpha_1 x_{j1} + \alpha_2 x_{j2} + \cdots + \alpha_m x_{jm}$$

where x_{jq} for l = 1,2, …, m is the q-th instructive variable value for data point j, andα_0,α_1, α_2 , …, α_m are the coefficients representative of the relation result of a specific informative variable on the output, the coefficients α_0,α_1, α_2 , …, α_m are accumulated into a single vector α of size $m+1$. For every data point j, an additional explanatory pseudo-variable x_{j0} is added, along with a fixed value of 1, equivalent to the intercept coefficientα_0. The resultant informative variables x_{j0} (=1), x_{j1} ……,x_{jm} are then congregated into a single vector x_j of size $m+1$. In the case of vector notation, it is inscribed the linear predictor function given as $f\left(j\right) = \alpha x_j$ by applying the dot product of two vectors. The matrix representation of it is shown as $f\left(j\right) = \alpha^T x_j + x_j^T \alpha$ where α and x_j are supposed to be (m + 1) by -1 column vectors, α^T is vector transpose α, $\alpha^T x_j$ represent matrix multiplication between (m + 1) row vector and (m + 1) column vector. In this case, for each data point j, a set of explanatory variables is created as $x_{j1} = x_j$

Table 3 Social Selling

S. No.	Parameters	Percentage (%)
1	Lead development	69
2	Account research	65
3	Call preparation	60
4	Stack holder research	59
5	Awareness building	50
6	Social brand building	41
7	Contents creation	25
8	Competitive analysis	17

$,x_{j2} = x_j^2, \ldots\ldots\ldots\ldots x_{jm} = x_j^m$ The radial basis functions (RBF's), which is expended to compute selected changed version of the distance to selected fixed point: Ø(x;c) = Ø(||x-c||) = Ø(∅($\sqrt{(\text{x1} - \text{c1})^2 + \ldots\ldots\ldots\ldots + (\text{xk} - \text{ck})^2}$ for k-dimensional result value, the Gaussian RBF, which has the equivalent functional form as the normal distribution $\varnothing(x; c) = e^{-b(||x-c||)^2}$ that drops off quickly which is the distance from $\boldsymbol{c}$ increases. The notation and its descriptions are given in Table 4.

Problem formulation

We have removed the dirtiness from the data to make the quality data. To get the quality result we have to make the quality data by removing the outliers, noises, errors from the data using tools and techniques. We have used regression analysis to remove the noises from the data. The linear regression analysis for YouTube and Facebook is given in Eq. 1. The ∈ symbol is representing an error in the data.

$$y_{ut} = a_{utfb} + a_{utfb}x_{fb} + \in \tag{1}$$

Let a_{utfb} is intercept and b_{utfb} coefficient for YouTube and Facebook. y_{ut} represent the Likes/Followers/Visits/Downloads of a product on YouTube while x_{fb} represent the like/download on Facebook. The intercept for YouTube and Facebook is given in Eq. (2) and coefficient in Eq. (3)

$$a_{utfb} = \frac{(\sum y_{ut})(\sum x_{fb}{}^2) - ((\sum x_{fb})(\sum x_{fb}y_{ut})}{n(\sum x_{fb}{}^2 - (\sum x_{fb})^2} \tag{2}$$

$$b_{utfb} = \frac{(n\sum x_{fb}y_{ut}) - (\sum x_{fb})(\sum x_{fb})}{n(\sum x_{fb}{}^2 - (\sum x_{fb}{}^2)} \tag{3}$$

We can put the value of a_{utfb} and b_{utfb} from Eq. (2) and (3) into Eq. (1). We get the Eq. (4)

Table 4 Notations and Descriptions

Notations	Descriptions
α^T	Transpose of vector $\boldsymbol{\alpha}$
a_{utfb}	Intercept for YouTube and Facebook
b_{utfb}	The coefficient for YouTube and Facebook
∈	Error in data
a_{LiTw}	Intercept for LinkedIn and Twitter
b_{LiTw}	The coefficient for LinkedIn and Twitter
y_{ut}	Likes/Followers/Visits/Downloads on YouTube
x_{fb}	Likes/Followers/Visits/Downloads on Facebook
y_{Li}	Likes/Followers/Visits/Downloads on LinkedIn
a_{IgPi}	Intercept for Instagram and Pinterest
b_{IgPi}	The coefficient for Instagram and Pinterest
y_{Ig}	Likes/Followers/Visits/Downloads on Instagram
x_{Pi}	Likes/Followers/Visits/Downloads on Pinterest

$$y_{ut} = \left[\frac{(\sum y_{ut})(\sum {x_{fb}}^2) - ((\sum x_{fb})(\sum x_{fb}y_{ut})}{n(\sum {x_{fb}}^2 - (\sum x_{fb})^2}\right] + \left[\frac{(n\sum x_{fb}y_{ut}) - (\sum x_{fb})(\sum x_{fb})}{n(\sum {x_{fb}}^2 - (\sum {x_{fb}}^2)}\right] + \in \tag{4}$$

The $\mathrm{a_{LiTw}}$ *and* $\mathrm{b_{LiTw}}$ are intercept and coefficient for LinkedIn and Twitter respectively for Likes/Followers/Visits/Downloads of the product by users. $\mathrm{y_{Li}}$ show the like/download of product on LinkedIn and $\mathrm{x_{Tw}}$ give the like/download on Twitter in Eqs. (5) and (6).

$$a_{LiTw} = \frac{(\sum y_{Li})(\sum {x_{Tw}}^2) - ((\sum x_{Tw})(\sum x_{Tw}y_{Li})}{n(\sum {x_{Tw}}^2 - (\sum x_{Tw})^2} \tag{5}$$

$$b_{LiTw} = \frac{(n\sum x_{Tw}y_{Li}) - (\sum x_{Tw})(\sum x_{Tw})}{n(\sum {x_{Tw}}^2 - (\sum {x_{Tw}}^2)} \tag{6}$$

Modify the Eq. (1) by replacing y_{ut} with $\mathrm{y_{Li}}$ *and* x_{fb} with $\mathrm{x_{Tw}}$ we get Eq. (7).

$$y_{Li} = a_{LiTw} + b_{LiTw}x_{Tw} \tag{7}$$

Substitute $\mathrm{a_{LiTw}}$ and $\mathrm{b_{LiTw}}$ in the above equation, we get the following Eq. (8)

$$y_{Li} = \left[\frac{(\sum y_{Li})(\sum {x_{Tw}}^2) - ((\sum x_{Tw})(\sum x_{Tw}y_{Li})}{n(\sum {x_{Tw}}^2 - (\sum x_{Tw})^2}\right] + \left[\frac{(n\sum x_{Tw}y_{Li}) - (\sum x_{Tw})(\sum x_{Tw})}{n(\sum {x_{Tw}}^2 - (\sum {x_{Tw}}^2)}\right] + \in \tag{8}$$

The $\mathrm{a_{IgPi}}$ and $\mathrm{b_{IgPi}}$ are intercept and coefficient for Instagram and Pinterest respectively for Likes/Followers/Visits/Downloads of the product by users. $\mathrm{y_{Ig}}$ is the like/download of product on Instagram and $\mathrm{x_{Pi}}$ is the like/download on Pinterest in Eqs. (9) and (10).

$$a_{IgPi} = \frac{(\sum y_{Ig})(\sum {x_{Pi}}^2) - ((\sum x_{Pi})(\sum x_{Pi}y_{Ig})}{n(\sum {x_{Pi}}^2 - (\sum x_{Pi})^2} \tag{9}$$

$$b_{IgPi} = \frac{(n\sum x_{Pi}y_{Ig}) - (\sum x_{Pi})(\sum x_{Pi})}{n(\sum {x_{Pi}}^2 - (\sum {x_{Pi}}^2)} \tag{10}$$

Replace y_{ut} with y_{Ig} *and* x_{fb} with x_{Pi} in Eq. (1) to get Eq. (11).

$$y_{Ig} = \frac{(\sum y_{Ig})(\sum {x_{Pi}}^2) - ((\sum x_{Pi})(\sum x_{Pi}y_{Ig})}{n(\sum {x_{Pi}}^2 - (\sum x_{Pi})^2} + \frac{(n\sum x_{Pi}y_{Ig}) - (\sum x_{Pi})(\sum x_{Pi})}{n(\sum {x_{Pi}}^2 - (\sum {x_{Pi}}^2)} + \in \tag{11}$$

The results to be forecasted are presumed to be random variables, the instructive variables themselves. We are having fixed values, and anyone random variables are supposed to be restricted on them. As we see the result, the data analysis changes the informative variables in random ways, incorporating making several copies of a given informative variable; each changed using a dissimilar function. We used usual techniques to make different informative variables in the usage of interface variables by fascinating products of two existing informative variables.

We have fixed a set of nonlinear functions which are used to change the data point, value(s) using basis functions. Polynomial regression which is using a linear predictor function to fit a random degree polynomial relationship between YouTube, Facebook and LinkedIn, Twitter sets of data points, by adding various instructive variables equivalent to numerous influences of the persisting informative variable.

Big data analytics for social media consumers

Analytics of marketing depends on big data in expressions of future predictions of the customers behavior. Therefore several companies invest in the tools of big data solutions to supervise the experience of customers in social media.

The most common benefits of big data analytics for social media marketing [24] are given below.

i. *Omni channel sources:* The strategy of artificial Intelligence allows data processing, that is coming from different channels. Several business websites advise sign-ups via Google or Facebook accounts, so this enables marketers to collect information related to customers from social media activity such as the history of browsing mobile applications, desktop and storages on Cloud.

ii. *Real-time interaction:* The activity of the users on social media like ads clicked, visited pages and followed, posted comments, saved links, and added friends are the primary technique to a successful study of the market. There is no other outlet that can give a more updated and precise picture of market demand.

iii. *Target clients:* Such as other business initiatives, social media marketing is predicted to extend income. Therefore, knowing your targeted viewers means the entire thing. Machine learning solutions achieve faraway beyond and provides the chance to require out precious insights from individual information, many photos, music preferences, locations, and many other social network activities.

iv. *Future predictions:* The approach of big data and predictive analytics in social media make it possible to enhance deciding on the idea of history. The business based on data tends to succeed enormously as computers can provide forthcoming consumer choices. Though interests and habits change with time, generally, they continue to be related. Once a social network users buys something, there's an excellent possibility of selecting similar products.

v. *Security issues:* With the success of social media and private information being put on a show, privacy is the whole thing for patrons, strange though it'd sound. Whereas this feature still leaves much to be preferred, the volume of enterprises considers security issues to be the main concern. Vendors of data altogether with marketers and business owners are obliged to supply data security from leaks to third-party hands without consumers' permission. Big data solutions put forward alternative behavior of protection, for example, expression and voice recognition, permission, enroll notifications, etc.

vi. *Campaign evaluation:* Big data analytics makes it possible to successfully observe the go up and down dynamics of ROI metrics. As a result, marketers can put insights on how flourishing a social media campaign was. Predictive analytical tools perform extremely, when it involves anticipating what products and services consumers want. Measuring user behavior across a variety of social media channels, namely, their interaction and reply to online ads can converse volumes about consumer behavior and their shopping preferences.

vii. *Reasonable prices:* Pricing decisions are over and over again provoking from time to time because several parameters must be kept in mind. Usually, it starts with product cost, competition issues, market demand, positive revenue, currency, and inflation levels and finishes up with an overall economic situation within the world. A strong big data strategy via social media shouldn't only include paying piles of cash to your Instagram influencers, but also communicating together with your loyal customers, say, through A/B testing or online surveys, to realize what proportion they are able to spend on your products. All this will help marketers to adjust prices in a more flexible and accurate manner that meets customer expectations.

The big data analytics technique is most popular right now in the real world. Most of the people in the real world are using big data analytics techniques for their research to analyze their data in real-world, that are coming in different type into existence with different properties like high volume, velocity, variety, and volume, such data is termed as big data, such type of data cannot be analyzed by using traditional analytics technique. The big data analytics technique is helpful to analyze such type of data. The consumer behavior data are which are coming in high volume right now into existence. The variety of consumer behavior data are coming in variety into the existence. Consumer behavior is very important on social media, so the big data analytics technique is helpful in behavior prediction from social media.

Big data analytics is considered a complex process of examining a varied and large set of data or big data, to know the hidden information, such as hidden patterns, unknown correlations, consumer perception, and customer preferences, that assists organizations to take proper decisions. There are five dimensions of big data management volume, velocity, variety, value, and veracity, which are depicted in Fig. 2.

Social media consumer behaviour model

In this model, we are predicting consumer behavior on social media like Facebook, YouTube, LinkedIn, and Twitter, etc. We predicted the behavior of consumers by doing big data analytics. We have developed a social media consumer behavior model. The framework of the model is given in Fig. 3. In this model, data have been considered from the sources Facebook, YouTube, LinkedIn, and Twitter. The data are cleaned by removing noises, errors, duplication, and outliers to make the quality data. In research study [25], they have measures consumer assignation with social media, somewhere consumer assignation integrates consumer answers to communications of marketing, in this work the author's debates that convinced motivations for social media usage assist as predecessors to general attitudes concerning social networking sites, that successively affects attitudes toward sellers' social networking sites. A study [9] in which it is given that the factors such as the creative idea, the actual social part being focused on, the particular social media platform, and comparison with the trademark may be an influence on consumers' attitude toward CSR as well as their assignation with CSR communication in social media.

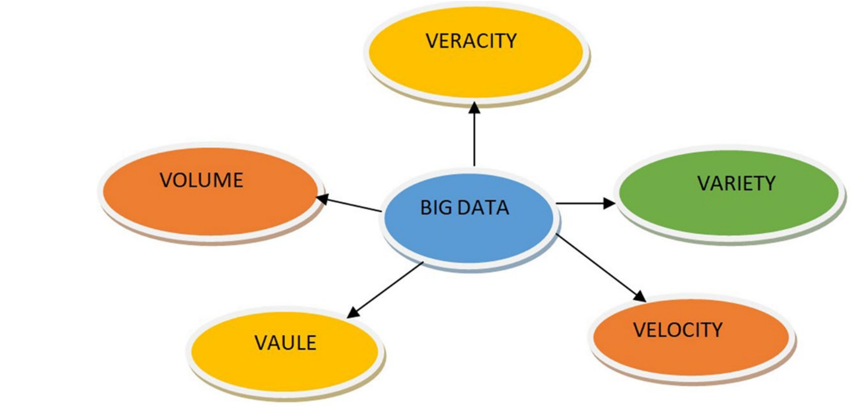

Fig. 2 Characteristics of big data

The cleaned data are divided into two parts. We have taken 80% data for training our model and 20% for testing the model which is depicted in Fig. 4. The output of the model is validated and got the final result. Let consider data source is ds and Facebook, YouTube, LinkedIn and Twitter represent f, y, l, t respectively. The sources of data are given in the equation given $\mathbf{ds} = \sum(\boldsymbol{f}, \boldsymbol{y}, \boldsymbol{l}, \boldsymbol{t})$. Customer behavior model makes a group of customers based on common behavior among the customer in direction to find out how related customers will act in alike situations. The machine learning algorithms are used to process the data for the prediction of consumer behavior on social media. The supervisor learning algorithms are used for prediction. The supervised learning algorithms are used because of the level given in the dataset.

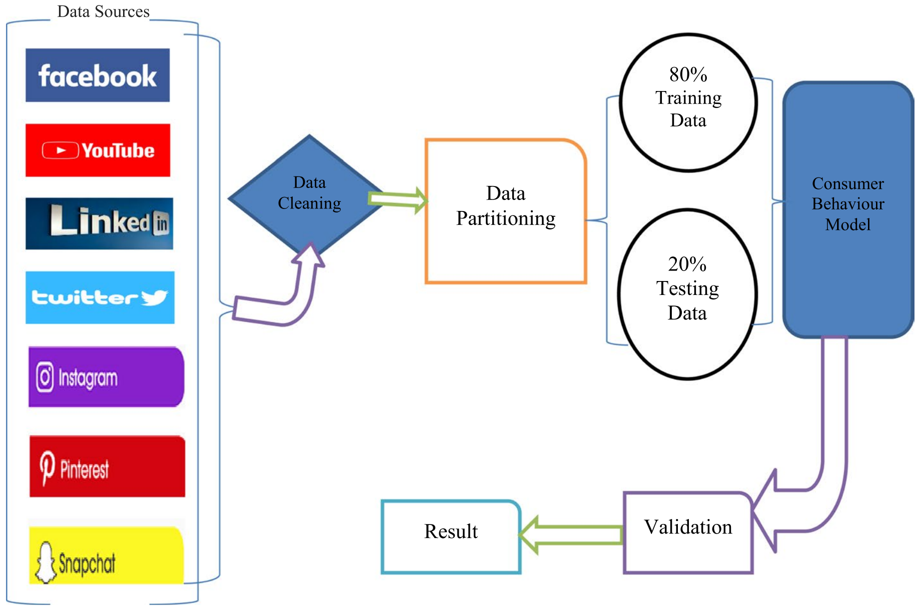

Fig. 3 Consumer behavior model

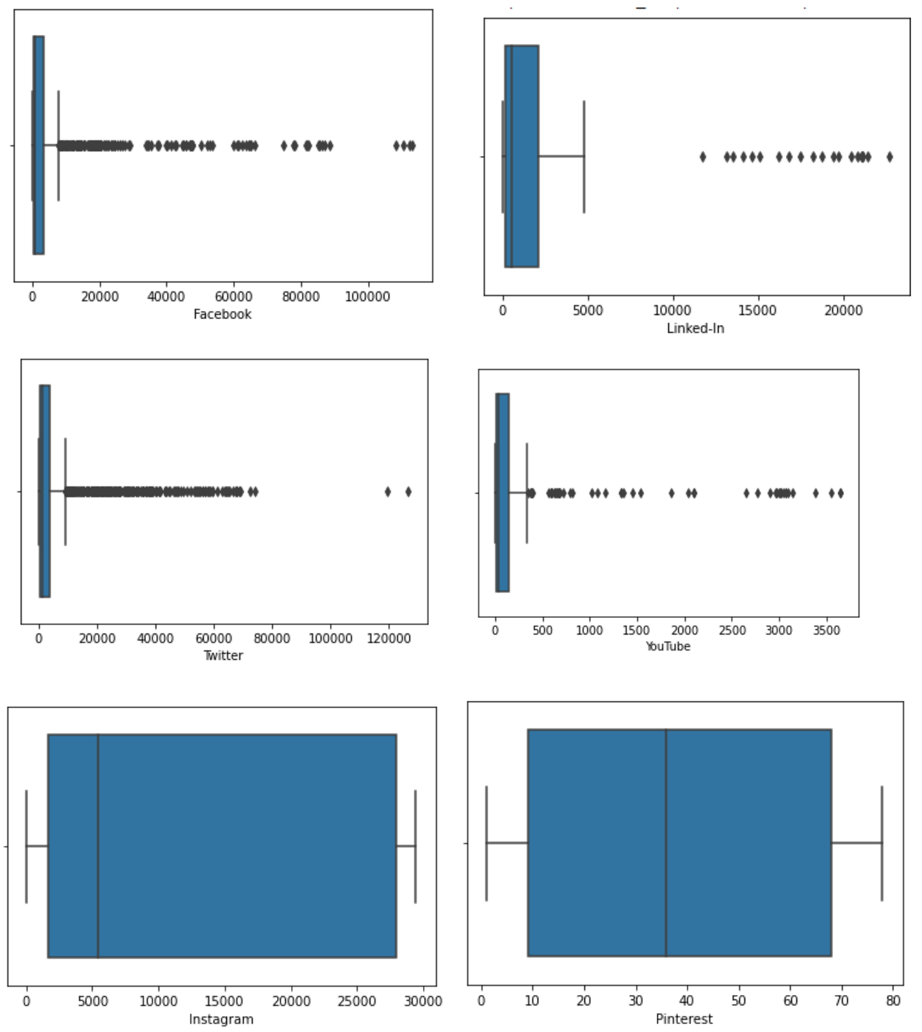

Fig. 4 Outlier detection

Data pre-processing

The dataset contains a total of 5279 records. We cleaned the data by removing the missing value of the attributes in the dataset. Out of 5279 records, 3962 are cleaned records after removing the missing values. There are four attributes in the dataset namely agency, platform, URL, sampled date, and Likes/Followers/Visits/Downloads. The set of the platform is defined f, which is f = {Facebook, Instagram, Linked-In, Twitter, YouTube, Pinterest,}. The shape of a dataset is given in Table 5.

The data are pre-processed to detect outliers from data. The outlier's detection for social media platforms Facebook, Instagram, Linked-In, Twitter, YouTube, and Pinterest are given in Fig. 4.

We have removed the outlier from the data to make the quality data. The data that behavior different from other data in the data set is removed. The outlier detections from Facebook, Linked In, Twitter, and YouTube data are shown in Fig. 4. We know that the Outliers are very dangerous. The outliers can strongly influence the result of a model. Frequently, the researchers assess the outliers to find whether every exacting evidence is

Table 5 Dataset Shape

The shape of dataset: (5279, 5)	Count
Agency	0
Platform	0
URL	140
Date sampled	0
Likes/followers/visits/downloads	1301
Dtype	int64

Table 6 Preprocessing Techniques

Data cleaning	Data transformation	Data reduction
Missing data	Normalization	Data cube aggregation
Noisy data	Attribute selection	Attribute subset selection
Duplicate record	Discretization	Numerosity reduction

the result of an error in the collection of data or an exceptional occurrence that should be taken into consideration for data processing. The various techniques used in preprocessing is given in Table 6

We removed the missing data by simply ignoring the missing rows in the dataset. The noisy data is removed with the help of regression. The duplicate records are removed from the dataset by using python expression. We have made the quality data with the help of pre-processing technique to trained and test our model.

Result and discussion

We have applied various functions on social media data such as Facebook, LinkedIn, Twitter, Instagram, Pinterest, and YouTube and investigated Likes, Followers, Visited, and Download from all platforms. We have computed count, means, standard deviation, min, for all platforms which are given in Table 7. The researcher's work [26], they have shown that organizations can mine business intelligence from social media data behavior on significant business application, assessing brand behavior. Specifically, they developed a text analytics framework that assimilates different separate social media data sources that consumers, employees, and organizations generated to measure brand behavior.

There is also computed 25%, 50%, 75%, and Maximum for Likes, Followers, Visited, Downloaded for Facebook, LinkedIn, Twitter, YouTube, Instagram and Pinterest. The count is the highest in Facebook and lowest in LinkedIn. The highest mean is 6619.079313 for Facebook among Facebook, LinkedIn, Twitter, and YouTube, and the lowest mean 240.758242 for YouTube. The highest standard deviation is 13660.290499 which belong to Twitter and the lowest standard deviation is 631.486554 of YouTube. The deviation of consumer behaviour from LinkedIn to Facebook is $\{\text{LinkedIn}_{\text{std}}\ \text{Facbook}_{\text{std}}\} = \{6273.373210, 12172.251382\}$ and similarly for others. Let CB be consumer behavior therefore

Table 7 Functions for All Platform

Functions	Likes/ followers/ visits/ downloads all platform	Facebook	Linked-In	Twitter	YouTube	Instagram	Pinterest
Count	3.962000e+03	1488.000000	134.000000	1223.000000	455.000000	11.000000	15.000000
Mean	1.856503e+04	4617.489247	3443.895522	6619.079313	240.758242	12237.909091	37.666667
Std	1.436232e+05	12172.251382	6273.373210	13660.290499	631.486554	13250.201217	32.480030
Min	1.000000e+00	3.000000	26.000000	1.000000	1.000000	22.000000	1.000000
25%	2.000000e+02	289.000000	151.250000	388.500000	11.000000	1720.000000	9.000000
50%	7.205000e+02	717.000000	540.000000	1348.000000	42.000000	5450.000000	36.000000
75%	3.433000e+03	3373.000000	2100.250000	3907.500000	142.000000	28007.000000	68.000000
Max	2.785806e+06	113264.000000	22691.000000	126661.000000	3649.000000	29448.000000	78.000000

$$\mathrm{CB}(\mathrm{LinkedIn}_{\mathrm{std}} \rightarrow \mathrm{Facbooks}) = \frac{\mathrm{Facebook}_{\mathrm{std}} - \mathrm{LinkedIn}_{\mathrm{std}}}{\mathrm{Facebook}_{\mathrm{std}} + \mathrm{LinkedIn}_{\mathrm{std}}} \times 100 = 48.46\%$$

$$\mathrm{CB}(\mathrm{LinkedIn}_{\mathrm{std}} \rightarrow \mathrm{Twitter}_{\mathrm{std}}) = \frac{\mathrm{Twitter}_{\mathrm{std}} - \mathrm{LinkedIn}_{\mathrm{std}}}{\mathrm{Twitter}_{\mathrm{std}} + \mathrm{LinkedIn}_{\mathrm{std}}} \times 100 = 37.06\%$$

$$\mathrm{CB}(\mathrm{YouTube}_{\mathrm{std}} \rightarrow \mathrm{LinkedIn}_{\mathrm{std}}) = \frac{\mathrm{LinkedIn}_{\mathrm{std}} - \mathrm{YouTube}_{\mathrm{std}}}{\mathrm{LinkedIn}_{\mathrm{std}} + \mathrm{YouTube}_{\mathrm{std}}} \times 100 = 81.71\%$$

$$\mathrm{CB}(\mathrm{Facebook}_{\mathrm{std}} \rightarrow \mathrm{Twitter}_{\mathrm{std}}) = \frac{\mathrm{Twitter}_{\mathrm{std}} - \mathrm{Facebook}_{\mathrm{std}}}{\mathrm{Twitter}_{\mathrm{std}} + \mathrm{Facebook}_{\mathrm{std}}} \times 100 = 12.22\%$$

$$\mathrm{CB}(\mathrm{YouTube}_{\mathrm{std}} \rightarrow \mathrm{Facebook}_{\mathrm{std}}) = \frac{\mathrm{Facebook}_{\mathrm{std}} - \mathrm{YouTube}_{\mathrm{std}}}{\mathrm{Facebook}_{\mathrm{std}} + \mathrm{YouTube}_{\mathrm{std}}} \times 100 = 90.14\%$$

$$\mathrm{CB}\left(\mathrm{Pinterest}_{\mathrm{std}} \rightarrow \mathrm{Instagram}_{\mathrm{std}}\right) = \frac{\mathrm{Instagram}_{\mathrm{std}} - \mathrm{Pinterest}_{\mathrm{std}}}{\mathrm{Instagram}_{\mathrm{std}} + \mathrm{Pinterest}_{\mathrm{std}}} \times 100 = 99.51\%$$

The consumer behavior deviation from one social media platform to another is given in Table 8

The consumer behavior is highly deviated from Pinterst to Instagram, which is 99.51%, and the lowest deviation from Facebook to Twitter that is 12.22%. The Density of Facebook, LinkedIn, Twitter, and YouTube are given in Fig. 5.

The numbers of unique value in each data source's columns are given in Table 9. The highest unique value of Likes/Followers/Visits/Downloads is 2158 and the lowest is 20 for Platform and date sampled.

The Likes/Followers/Visits/Downloads on social media platforms namely Facebook, LinkedIn, Twitter, YouTube, Instagram, and Pinterest relationship are given in Fig. 6.

Creating data features

In our machine learning-based social media consumer behavior model maps a data inputs set are given in Table 10. The purpose of creating data feature for our model is to learn a

Table 8 Consumer Behavior Deviation .

S.No.	Consumer behavior deviation	Percentage deviation
1	CB($\text{LinkedIn}_{std} \rightarrow \text{Facbook}_{std}$)	48.46
2	CB($\text{LinkedIn}_{std} \rightarrow \text{Twitter}_{std}$)	37.06
3	CB($\text{YouTube}_{std} \rightarrow \text{LinkedIn}_{std}$)	81.71
4	CB($\text{Facebook}_{std} \rightarrow \text{Twitter}_{std}$)	12.22
5	CB($\text{YouTube}_{std} \rightarrow \text{Facebook}_{std}$)	90.14
6	CB($\text{Pinterst}_{std} \rightarrow \text{Instagram}_{std}$)	99.51

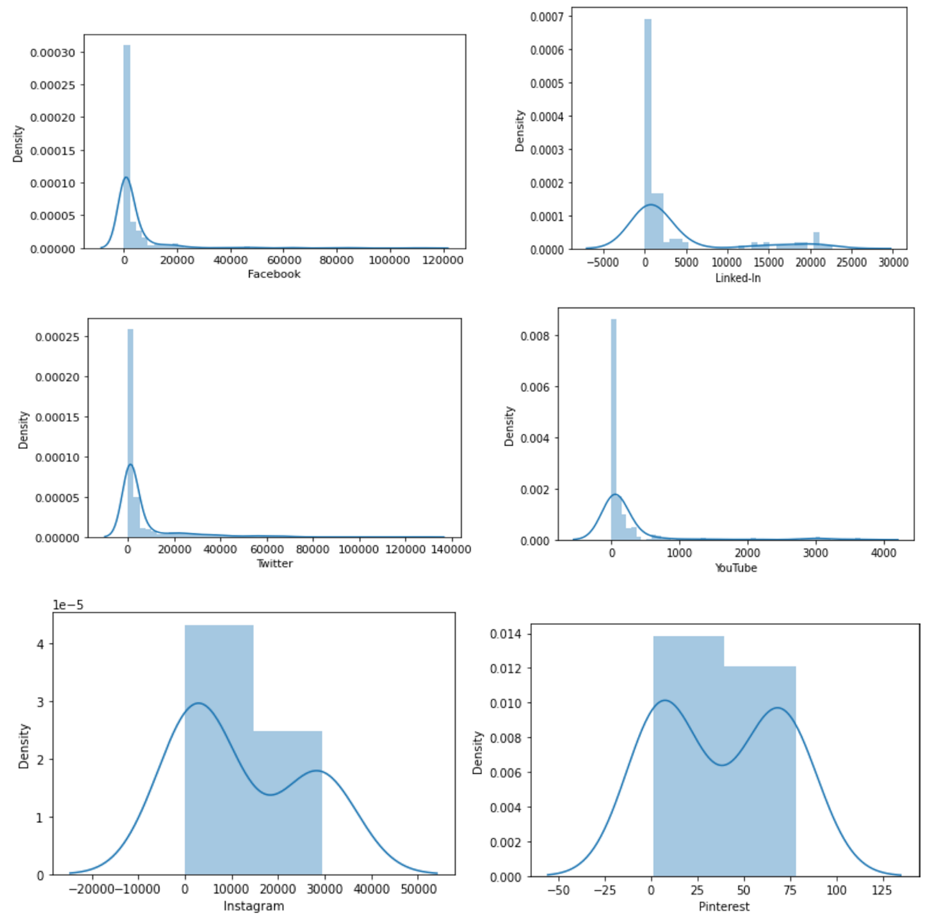

Fig. 5 Density verse platform

pattern of consumer behavior in terms of likes, followers, visited and downloaded or characterizing between the inputs and target Facebook, Twitter, LinkedIn, and YouTube, so that new data is given to the our model, where target is unidentified, our model can accurately predict the target consumer behavior on social media platforms such as Facebook, Twitter, LinkedIn, and YouTube. We have considered four social networking sites namely Facebook, LinkedIn, Twitter, and YouTube from our created data features.

Considering the Dtype parameter for classification, based on these parameters float64, object, datetime64 [ns], and int64 were used to make four classes namely c1, c2, c3, and c4.

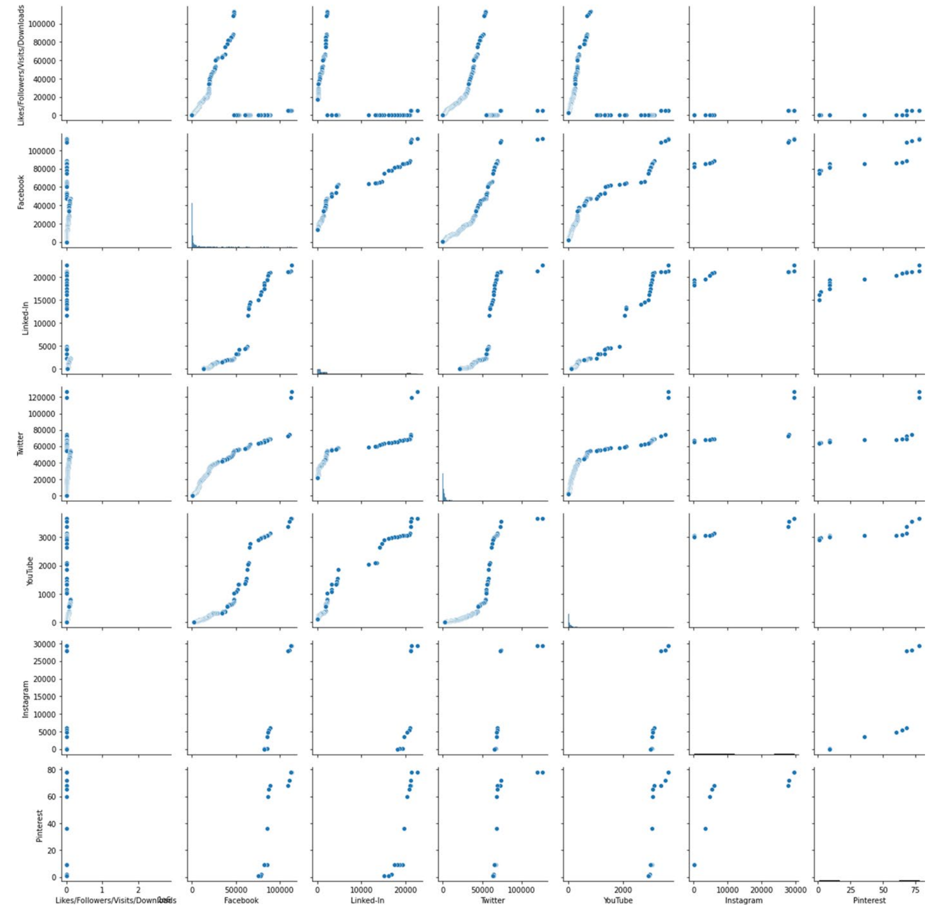

Fig. 6 Relationship between two variables

$$\mathbf{c1}=\left\{\begin{array}{c|c} \text{Facebook} & 1488 \\ \text{Linked - In} & 134 \\ \text{Twitter} & 1223 \\ \text{YouTubeok} & 455 \\ \text{Instagram} & 11 \\ \text{Pinterest} & 15 \end{array}\right\}\ \mathbf{c2}=\left\{\begin{array}{c|c} \text{Agency} & 3962 \\ \text{Platform} & 3962 \\ \text{URL} & 3880 \end{array}\right\}$$

$$\mathbf{c3}=\left\{\begin{array}{c|c} \text{Date Sampled} & 3962 \\ \text{date} & 3962 \end{array}\right\}\ \mathbf{c4}=\left\{\begin{array}{c|c} \text{dayofweekb} & 3962 \\ \text{quarter} & 3962 \\ \text{month} & 3962 \\ \text{year} & 3962 \\ \text{dayofyear} & 3962 \\ \text{dayofmonth} & 3962 \\ \text{weekofyear} & 3962 \end{array}\right\}$$

Model comparison

We have considered various models based on root mean square error and accuracy on validation data which is given in Table 11. The Linear Regression, Decision Tree Regressor, Random Forest Regressor, Extra Tree Regressor, Ada Boost Regressor, XGB

Table 9 Unique Values in Data Source

Data sources Sources	Unique value
Agency	139
Platform	20
Url	250
Date Sampled	20
Likes/Followers/Visits/Downloads	2158
Facebook	1114
Linked-In	112
Twitter	1072
YouTube	204

Regressor, and Bagging Regressor model are used in our problem. The Model LR, DTR, FRF, ETR, ABR, XGB, and BR are representing Linear Regression, Decision Tree

Table 10 Data Input Set

#	Column	Non-null	Count	Dtype
0	Agency	3962	Non-null	Object
1	Platform	3962	Non-null	Object
2	URL	3880	Non-null	Object
3	Date Sampled	3962	Non-null	Datetime64[ns]
4	Facebook	1488	Non-null	Float64
5	Linked-In	134	Non-null	Float64
6	Twitter	1223	Non-null	Float64
7	YouTube	455	Non-null	Float64
8	Instagram	11	Non-null	Float64
9	Pinterest	15	Non-null	Float64
10	Date	3962	Non-null	Datetime64[ns]
11	Dayofweek	3962	Non-null	int64
12	Quarter	3962	Non-null	int64
13	Month	3962	Non-null	int64
14	Year	3962	Non-null	int64
15	Dayofyear	3962	Non-null	int64
16	Dayofmonth	3962	Non-null	int64
17	Weekofyear	3962	Non-null	int64

Table 11 Root Mean Squire and Accuracy on Validation Data

S.No.	Models	Root mean square error	Accuracy on validation data
1	LR	156556.45293730905	0.022388308037899925
2	DTR	20691.78703623191	0.9829226515205226
3	RFR	25373.003231378854	0.9743215863591789
4	ETR	21133.91540130951	0.982185059749982
5	ABR	35891.80415270158	0.9486175250138749
6	XGB	29863.671852777836	0.9644277897507563
7	BR	23921.588532478676	0.9771753318099993

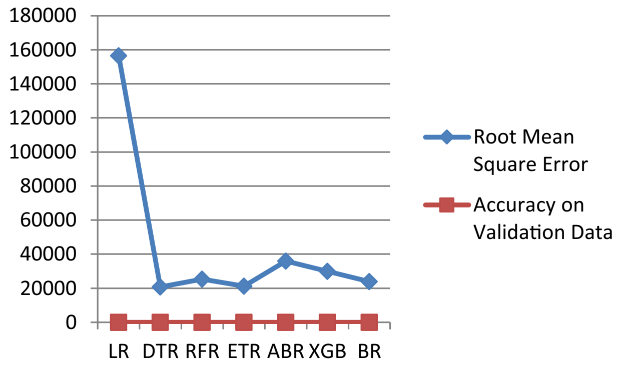

Fig. 7 Graphical representation of models

Regressor, Random Forest Regressor, Extra Tree Regressor, Ada Boost Regressor, XGB Regressor, and Bagging Regressor respectively.

The highest root mean square error in linear regression model, while highest accuracy on validation of data in Decision Tree Regressor. The lowest root square mean error is 20691.78703623191, in the decision tree Regressor model, while 0.0223883080378999925 is the lowest accuracy on validation data in the linear regression model. The lowest root mean square error and highest accuracy on validation data in Decision Tree Regressor, so this model is best for our problem to predict the consumer behavior on social media. In a research study [27] the researchers have concentrated on predicting the user status in the second-hand market Wallapop based merely on Twitter profiles of users. The study [28] in which authors explained the base for developing upcoming churn prediction model which will be helpful in the informed decision-making process. The graphical representation of various models is given in Fig. 7.

The red color in the graph represents accuracy on validation of data for various models namely Linear Regression, Decision Tree Regressor, Random Forest Regressor, Extra Tree Regressor, Ada Boost Regressor, XGB Regressor, and Bagging Regressor which are represented as LR, DTR, RFR, ETR, ABR, XGB, and BR. The blue color represents the root Mean Square error for these models.

Conclusion

We predicted consumer behavior from the social media data like Facebook, YouTube, LinkedIn, Twitter, Instagram, and Pinterest. This model is helpful for businesses to predict consumer behavior about the product using social media data. The decision tree is the best model for consumer behavior prediction on social media. The highest consumer deviation 99.51% from one social media to another and the minimum is 12.22%. The highest root means square error is 156556.45293730905 among all and the minimum is 20691.78703623191. The maximum accuracy in all is 0.9829226515205226 and minimum 0.022388308037899925. We have used the machine learning technique to predict consumer behavior on social media with the use of mathematical concepts using Big Data Analytics. These models predict consumer behavior of various platform based on consumers likes, followers, download, etc. The limitation of this model is that it will not work on daily basis consumer data. If this model is used on daily basis data, the result will be very poor.

Acknowledgements

We would like to thanks Deanship of Scientific Research, Research Chair of Pervasive and Mobile Computing, King Saud University, Riyadh, Saudi Arabia for providing financial support regarding this work.

Authors' contributions

KC: She has conceptualized the concept idea and collected the dataset. She has also prepared the literature review and contributed to the business concept as well as logic. MA: He did the experimental work using tools and techniques. He contributed to the result discussion. ASA-R: He worked on the methodology of the problem and organization of this paper. He also helps in experiment. AG: He contributed to developing the model and validation of it. He also contributed to language editing and writing. All authors read and approved the final manuscript.

Funding

This work is financially supported by the Deanship of Scientific Research, Research Chair of Pervasive and Mobile Computing, King Saud University, Riyadh, Saudi Arabia.

Availability of data and materials

Dataset https://drive.google.com/file/d/1WEFDyrLUL5H5HfOqOv0nFW-IxGp-_lBl/view?usp=sharing

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Commerce, Shivaji College, University of Delhi, New Delhi, India. [2]Department of Computer Science, Big Data, Cloud Computing and IoT Laboratory, Jamia Millia Islamia, New Delhi, India. [3]Research Chair of Pervasive and Mobile Computing, Information Systems Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia.

Received: 15 February 2021 Accepted: 10 May 2021
Published online: 25 May 2021

References

1. Tufekci Z. Big questions for social media big data: representativeness, validity and other methodological pitfalls. In: Eighth international AAAI conference on weblogs and social media. 2014.
2. Giglio S, Pantano E, Bilotta E, Melewar TC. Branding luxury hotels: evidence from the analysis of consumers'"big" visual data on TripAdvisor. J Bus Res. 2020;119:495–501.
3. Jung SH, Jeong YJ. Twitter data analytical methodology development for prediction of start-up firms' social media marketing level. Technol Soc. 2020;63:101409.
4. Arasu BS, Seelan BJB, Thamaraiselvan N. A machine learning-based approach to enhancing social media marketing. Comput Electr Eng. 2020;86:106723.
5. Salehan M, Kim DJ. Predicting the performance of online consumer reviews: a sentiment mining approach to big data analytics. Decis Support Syst. 2016;81:30–40.
6. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manag. 2015;35(2):137–44.
7. Matz SC, Netzer O. Using big data as a window into consumers' psychology. Curr Opin Behav Sci. 2017;18:7–12.
8. Buettner R. Predicting user behavior in electronic markets based on personality-mining in large online social networks. Electron Mark. 2017;27(3):247–65.
9. Chu SC, Chen HT, Gan C. Consumers' engagement with corporate social responsibility (CSR) communication in social media: evidence from China and the United States. J Bus Res. 2020;110:260–71.
10. Costa PT. McCrae RR: Revised NEO Personality Inventory (NEO PIR) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual. *Odessa: Psychological Assessment Resources.*1992.
11. Shirdastian H, Laroche M, Richard M-oO. Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. Int J Inform Manag. 2019;48:291–307.
12. Ghani, NA, et al. Social media big data analytics: A survey. Comput Hum Behav. 2019; 101:417–28.
13. Stieglitz S, Mirbabaie M, Ross B, Neuberger C. Social media analytics—challenges in topic discovery, data collection, and data preparation. Int J Inf Manag. 2018;39:156–68.

14. Tayebi S, Manesh S, Khalili M, Sadi-Nezhad S. The role of information systems in communication through social media. Int J Data Netw Sci. 2019;3(3):245–68.
15. Stieglitz S, Meske C, Ross B, Mirbabaie M. Going back in time to predict the future-the complex role of the data collection period in social media analytics. Inf Syst Front. 2018;1–15.
16. Jansen BJ, Zhang M, Sobel K, Chowdury A. Twitter power: tweets as electronic word of mouth. J Am Soc Inf Sci Technol. 2009;60(11):2169–88.
17. Saif H, He Y, Alani H. Semantic sentiment analysis of twitter. In: International semantic web conference. Springer, Berlin, Heidelberg; 2012. pp. 508–524.
18. Jussila J, Vuori V, Okkonen J, Helander N. Reliability and perceived value of sentiment analysis for Twitter data. In: Strategic innovative marketing. Springer, Cham; 2017. pp. 43–48.
19. Radi SA, Shokouhyar S. Toward consumer perception of cellphones sustainability: a social media analytics. Sustain Prod Consum. 2021;25:217–33.
20. Chaudhary K, Kumar S. Customer satisfaction towards Flipkart and Amazon: a comparative study. Int J Acad Res Dev. 2016;35.
21. Scholz M, Schnurbus J, Haupt H, Dorner V, Landherr A, Probst F. Dynamic effects of user-and marketer-generated content on consumer purchase behavior: modeling the hierarchical structure of social media websites. Decis Support Syst. 2018;113:43–55.
22. Goldberg LR. An alternative description of personality: the big-five factor structure. J Pers Soc Ppsychol. 1990; 59(6):1216.
23. https://financesonline.com/social-media-trends/.
24. https://www.byteant.com/blog/7-ways-how-to-use-big-data-in-social-media/.
25. Bailey AA, Bonifield CM, Elhai JD. Modeling consumer engagement on social networking sites: roles of attitudinal and motivational factors. J Retail Consumer Serv. 2020;102348.
26. Hu Y, Xu A, Hong Y, Gal D, Sinha V, Akkiraju R. Generating business intelligence through social media analytics: measuring brand personality with consumer-, employee-, and firm-generated content. J Manag Inf Syst. 2019;36(3):893–930.
27. Prada A, Iglesias CA. Predicting reputation in the sharing economy with Twitter social data. Appl Sci. 2020;10(8):2881.
28. Bhattacharyya J, Dash MK. Investigation of customer churn insights and intelligence from social media: a netnographic research. Online Inf Rev. 2020. https://doi.org/10.1108/OIR-02-2020-0048.
29. https://socialmediaweek.org/blog/2017/10/not-social-platforms-created-equal-infographic/.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.