Complex & Intelligent Systems (2020) 6:621–634
https://doi.org/10.1007/s40747-020-00155-2

ORIGINAL ARTICLE

Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review

Shanshan Yi[1] · **Xiaofang Liu**[2]

Received: 11 March 2020 / Accepted: 13 May 2020 / Published online: 11 June 2020

Abstract
Big data analytics plays a major role in various industries using computing applications such as E-commerce and real-time shopping. Big data are used for promoting products and provide better connectivity between retailers and shoppers. Nowadays, people always use online promotions to know about best shops for buying better products. This shopping experience and opinion about the shopper's shop can be observed by the customer-experience shared across social media platforms. A new customer when searching a shop needs information about manufacturing date (MRD) and manufacturing price (MRP), offers, quality, and suggestions which can only be provided by the previous customer experience. The MRP and MRD are already available in the product cover or label. Several approaches have been used for predicting the product details but not providing accurate information. This paper is motivated towards applying Machine Learning algorithms for learning, analysing and classifying the product information and the shop information based on the customer experience. The product data with customer reviews is collected from benchmark Unified computing system (UCS) which is a server for data based computer product lined up for evaluating hardware, support to visualization, software management. From the results and comparison, it has been found that machine learning algorithms outperform than other approaches. The proposed HRS system has higher values of MAPE which is 96% and accuracy is nearly 98% when compared to other existing techniques. Mean absolute error of proposed HRS system is nearly 0.6 which states that the performance of the system is significantly effective.

Keywords Big data · Big data analytics · Machine learning · Feature extraction · Clustering and classification · Accuracy

Introduction

Social media and networks have probably taken over the globe by storm and have shrunk the dimensions of the world. There is a fair bit of communication happening between people across several social media platforms such as Twitter, Facebook, WhatsApp etc. These communications in most cases will be informal. These discussions convey the mood and sentiment of the persons involved in the discussion. This paves way for understanding the behavioural traits of people involved in the discussion. Sentimental analysis also called as opinion mining and natural language processing (NLP) play a vital role in analyzing and understanding the communication happening across the transcripts.

Sentiment analysis is referred as opinion mining or emotional intelligence (EI). Sentimental analysis can be defined as the art of gathering useful insight from unstructured and unorganized textual contents from several social platforms and online sources such as chats happening across social platforms such as Twitter, WhatsApp and Facebook, online blogs and comments. Opinion mining involves developing rule based automated systems that analyze the data based on a set of predefined rules. Also, automated systems that use any of the machine learning principles are developed to perform opinion mining. There are some scenarios where both rule based and automated machine learning algorithms are combined together to develop a hybrid model for sentimental analysis. Sentimental analysis can also serve as an effective means to study the positive or negative intent based on the textual contents. Sentimental analysis has served as a reliable source for providing insightful opinion about several products rolled over in the market, innovative ideas, people

✉ Shanshan Yi
shanshanyi@126.com

[1] School of Nuclear Technology and Automation Engineering, Chengdu University of Technology, Chengdu 610000, China

[2] Sichuan Aerospace Vocational College, Chengdu 610000, China

opinion about new policies framed by government etc. At the same time, sentiment analysis also plays a pivotal role in understanding the shopping experience in a particular shop based on the remarks that is shared by the customer across social media platform. Sentimental analysis thus undermines the significance in terms of understanding the customer's opinion in terms of the shopping experience gained while purchasing a product in a particular shop.

One of the key techniques that are instrumental in promoting sentimental analysis is natural language processing (NLP). It is one of the sub-domains in artificial intelligence that empowers the computer aided system to understand the different languages as communicated or spoken by human. NLP predominantly focuses on understanding the unstructured contents present in the social media and organizing the same to be made easier for sentimental analysis. NLP mainly focuses on reading and interpreting the free form text into an analysable format. The primary application area thus where NLP is a key component is opinion mining. Also, another common application where NLP plays a pivotal role is lending its help to search engines such as Google to tune their search algorithms to understand different contexts and to understand and interpret contents of different languages and produce relevant search results.

NLP predominantly deploys several machine learning algorithms for text data classification. One common and mostly preferred method deployed for classification is support vector machine (SVM). It is a typical supervised learning approach that uses machine learning algorithms to perform effective regression analysis and data classification. In the case of typical SVM based classification, the categories of classes are clearly defined and the model is well trained to fit the data into the specified classes. SVM can be used to perform linear classification as well as nonlinear classification. SVM plays a pivotal role in categorizing textual contents and hyper textual contents according to the classes defined. Also, SVMs are typically used for image classification, text and hand written character recognition. A binary SVM is focussed towards classifying data under two buckets or classes. However, human sentiment is multidimensional in nature and a simple binary SVM classifier may not suit sentimental analysis. This prompts for the use of multiclass support vector machine classifier for studying and analyzing sentiment by reading through textual data. A multiclass SVM is built by combining several binary SVMs. This approach towards sentimental analysis using twitter data uses MSVM for better classification of different types of sentiments that can be extracted out of twitter data.

The proposed approach is a machine learning based approach which uses multiclass support vector machine (MSVM) for classification of different classes of opinions and sentiments in twitter. The proposed approach comprises of importing the twitter data from the twitter source which is followed by data pre-processing which comprises of several tasks like removal of contents like punctuations, erroneous words, duplicate or redundant words and stop words. This in turn is a measure to enhance relevancy of data towards opinion mining. The data pre-processing is followed by feature extraction from the text data. Feature selection involves searching for text that conveys the mood of the person in conversation and this feature selection is carried out using Simulated Annealing. A semantic word dictionary is formed to identify the classes of opinions. This is followed by applying multiclass support vector machine as the classifier for classifying the sentiments.

The contribution of this work is to develop an effective data pre-processing method for unwanted content removal and filtering only relevant text. The feature extraction process involves extracting information pertaining to Manufacturing date (MFD), Manufacturing price (MRP), discounts and offers, quality ratings, and suggestions or reviews. Finally, multiclass support vector machine classifier is deployed to promote an efficient and effective multi class classification of different types of sentiments based on the above extracted features. Extraction of feature has wide range adoptability in automatic inspection manufacturing systems. Image feature identification in manufacturing industries fail due to poor unsuitable identification of features, thus machine learning approaches such as principal component analysis (PCA) is useful in identifying suitable features from manufacturing images. This PCA method is significant in reducing dimensionality for feature extraction approaches and features related to identification of different patterns in manufacturing.

Literature review

With the social media platforms such as Twitter, Facebook, Instagram and WhatsApp taking the communication world by storm, it has become imperative that the data residing across these social media platforms will convey insightful information about the opinion, mood and sentiment of the people over any product, idea or policies. Several works have been performed earlier to analyze the twitter contents and perform opinion mining over twitter data. The authors of [1] have proposed an approach that uses deep convolution neural network to analyze the twitter feed. The feature set was integrated into deep CNN for training and predicting sentiments by analyzing twitter data. When it comes to twitter data analysis, text pre-processing has a significant role to play. Several text pre-processing methods have been analyzed and compared as part of [2]. Since sentiment analysis involves analyzing and interpreting different types of sentiments, multi class sentiment analysis is highly essential

and is discussed in [3]. Authors of [4] have developed a tool named SENTA that uses a pattern comparison approach to perform multiclass sentiment analysis. As discussed in [5], is to understand the evolving changes of opinion mining, to the context and the dynamic events that occur during the twitter conversations.

The usage of twitter has kindled more research work towards understanding the sentiments using twitter data. One such work discussed in [6] uses a hybrid framework that uses a genetic algorithm-based approach to perform sentiment analysis. The focus of this work was to enhance the system from a scalability perspective. Authors of [7] have explored the variations of public sentiment over a given topic using a mathematical model to detect the foreground topics and promote effective ranking of candidates. An interesting work related to sentimental analysis is to classify sentiment based on the topic of discussion. Authors of [8] have provided an effective topic adaptive sentiment classification over tweets. The different challenges posed while performing multiclass sentiment analysis has been discussed in [9] and the authors have also developed a novel model that uses multiclass sentiment analysis over twitter data. Authors of [10] have proposed an effective text to speech conversion based on the sentences provided in twitter data. Authors of [11] have introduced a novel method for hierarchical topic modelling using twitter data to perform Online Analytical Processing (OLAP). To enhance the effectiveness of queries related to analytics, large OLAP database such as Vertica [12], Greenplum [13] and Teradata DB [14] has been suggested. Vertica [12] utilize projection to enhance performance of query. Instead of developing indexes which are conventional on columns, it retains the details about min/max ranges, leading to higher latency from lower pruning which are less efficient. Greenplum [13] and Teradata DB [14] approve store column wise and allow specification of indexing in column by users. Moreover, there exists, two drawbacks: initially, modification of column indexes in write path which is excessive for every column indexes; secondly, it requires more random I/Os for queries relating to point-lookup.

From the above works it can be easily understood that the contents of twitter data provide useful insights about any topic under discussion and also conveys the opinion of people involved over the particular topic. The main advantage and the unique functionality of machine learning algorithms is extracting the essence of the given problem. It provides complete information about the data and make the developer to think about choosing the appropriate algorithm which can be used for the problem. Some of the common classes of the machine learning problems are clustering, classification, regression and rule extraction.

Limitations

The data residing in twitter or any social media is mostly unstructured in nature and hence NLP has a pivotal role in structuring the unstructured data. At the same time the dynamics behind the topic under discussion and the context of discussion still remain a black box. This in turn reduces the efficiency of opinion mining and the accuracy of sentiment analysis and prediction. It is highly difficult for any sentimental analysis model to understand the external factors that may have an influence on the discussions that happen over twitter. These external factors have a direct bearing on the discussions and in turn impact the sentiment of people discussing about a particular topic. Another important aspect that is found wanting is the need for a quick search solution to spot the essential words that covers the mood and sentiment of the people discussions over twitter.

Motivation

Since human sentiment is multidimensional in nature, a simple binary support vector machine may not provide a right classification when it comes to opinion mining. Hence the proposed approach is motivated towards using a machine learning based regression model to accommodate the different types of sentiments that can be extracted from the twitter data. This in turn promotes using more classes which means that the classification accuracy is not compromised. Since there are five pivotal features namely—Manufacturing date (MFD), Manufacturing price (MRP), discounts and offers, quality ratings and suggestions or reviews that will be used to define customer shopping experience in a shop, regression based collaborative recommendation has been chosen to be the best classifier for this system. Also, it is imperative that the searching speed has a pivotal role in determining the overall efficiency of the opinion mining system. Using a high-end machine learning classifier such as regression based collaborative recommendation system ensures that the overall accuracy of the decision making is always high while classifying the customer's shopping experience from the shop perspective. Quantum computing refers to the phenomena which uses quantum mechanism such as entanglement and superposition to illustrate computing. Computers used for analysing computation is known as quantum computers. It uses quantum circuits for quantum analysis, circuits are on the basis of quantum bit. There exists two methods to implement physically which refers as quantum computer known digital and analog. Analog

methods utilise various sections such as simulation and annealing quantum. Digital quantum computes the logic gates for evaluating, whereas both methods uses bits for computing.

This paper integrates the collaborative filtering and product-product similarity method for designing the new recommendation system. Various types of recommendation systems are illustrated in Fig. 1. The collaborative filtering method applied for predicting the best shops and the product-product similarity method applied for predicting the best product. Both the products and the shops are predicted based on the highest ratings given by the customer.

Recommendation system

A recommender system is one of the information retrieval system which uses filtering method where it finds the requirement by analysing and comparing the score/rank/ rating of a product/item given by the customer. Then it predicts the high rating-based product for recommend to the new customers. This recommendation system makes the shops and shopper happy in terms of selling and buying. This recommendation system recommends not only few products or shops, but it used for wide range of products and shops added into its recommendation list. A recommendation system mainly used in various concerns like movies, videos, music, news, websites, clothes, Twitter pages, hotels, restaurants, travels, tourist places and etc. Most of the major enterprises are using recommendation system to elaborate their business, enrich customer experience.

The above Table 1 shows the relationship among the customer C_i and the product P_j as a matrix. The rows in the Table 1 represents the customer, the column represents the product, and the cells represents the rating value A_{ij} given by the customer. It is assumed that n number of products and k number of customers are involved in the paper. The term A_{ij} represents the rating value given by the customer C_i on the product P_j, where the value is from 1 to 5. If the product is not selected by the customer the value is 0, else it is 1. Hence the A_{ij} value may be binary initially. In other cases, the value of A_{ij} is from 1 to 5. Since, the customer and their option of product selection is not compulsory, the customer-product matrix is very sparse matrix. Means, all the cells have not been filled, many of the cells are not having any values.

Table 1 Customer – product matrix

	P_1	P_2	…	P_i	..	P_{n-1}	P_n
C_1			…		…		
C_2			…		…		
:	:	:	…		…	:	:
C_j			…	A_{ij}	…		
:	:	:	…		…	:	:
C_{k-1}			…		…		
C_k			…		…		

Assumptions

1. It is not compulsory for a customer to give rating value for at least one product, and the matrix is very sparse matrix.
2. One customer cannot provide rating value to all the product.
3. If the customer gives a rating value to one product, then it is assumed that the customer has experience on the product.
4. Number of customer and the products is not limited.
5. One shop can sell any number of products, and any shops can sell any product.

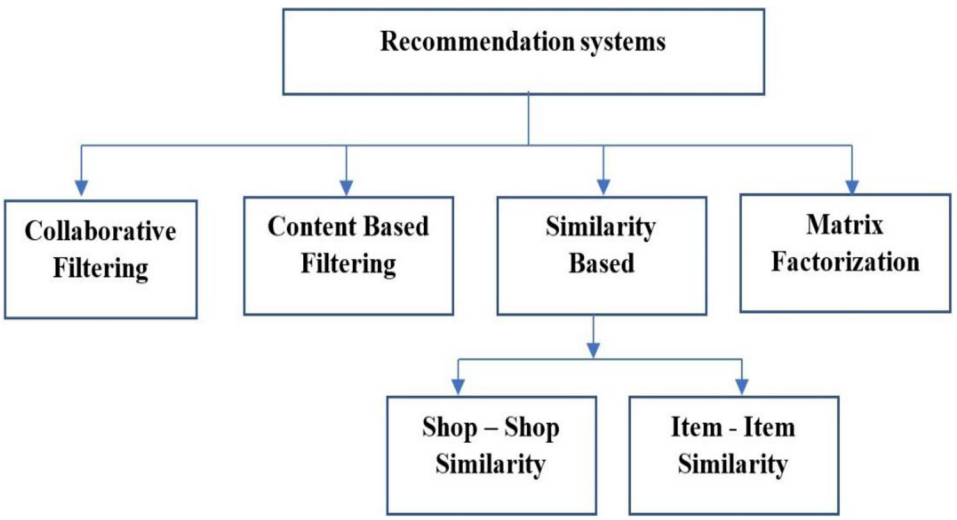


Fig. 1 Recommendation systems

Collaborative filtering is a procedure that is used to filter out components that a customer will like on reaction basis by user similarity. It is deployed on the basis of finding a larger set of people and searching a minimal group of customers with similar tastes to a particular user.

Cognitive filtering is another term illustrating the content-based filtering; it suggests the user on the basis of items content and the profile of customers. The items content is referred as a group of parameters.

Similarity based modelling is defined as a procedure wherein a normal procedure of system is performed for detecting the errors by examining their similarity to system's normal states.

Matrix factorization is a course of collaborating various filtering algorithms utilised in recommendation system. Its works by neutralizing the interaction of user item into component of two different minimal dimensionality matrices.

For example, assume that one customer cannot provide even to 1% of total products. Hence, approximately 99% of cells do not have any value. These empty cells are filled by dots "…", means, it is not a number. In case, if $n = 1$-million, and $k = 10$ k, then $n * k = 1010$, is a big number. If, an average number of customers gives rating to 5 products then, 5 * 1-million = 5 * 106 ratings. So, the customer-product matrix becomes a sparse matrix.

Due to the number of ratings filled in the matrix is comparatively less, this paper is mainly aimed to use the recommendation system, because the recommendation system can obtain the recommended product very fast than the other systems. One of the types of recommendation systems is collaborative filtering where it focused on collaboration among the customer. If a greater number of shops are linked with the products/items, they can be recommended to the customer who has not yet heard about the product. It is illustrated and can be understood easily from Fig. 2.

Among the various recommendation system, this paper focused on using collaborative filtering method. It is well-known that the collaborative recommendation system recommends a product or shop based on the rating value given by the customer. For example, there are three shops, four products, three old customers and one new customer are involved in the proposed scenario. All the three shops selling four products, but one product is selling by only one shop. Similarly, three customers buying three products, but only one product is buying by only one customer. So, based on that, product—three is recommended to the new customer. It means one product is selling by more shops and buying by more customer can be recommended to the new customer. This scenario is used for recommending a product, and it illustrates the collaborative filtering method.

The idea behind this method is, the shops and the customer who have accepted the product in past and present can agreed to the future as well. Here the first three products are accepted by the shops and the customer and is worth for buying. If the entire system is true, then making a collaborative system is easy and suitable for the application.

Product–product similarity

This product–product (**P-P**) similarity construction is also similar to customer-customer similarity [15]. It is also an imagination by inspecting Fig. 5.

P-P similarity can identify the best products which can be recommended for new customers. To understand the functionality of P-P similarity method, a similarity matrix is constructed. Calculating P-P similarity is similar to customer–customer similarity, but the customer–customer

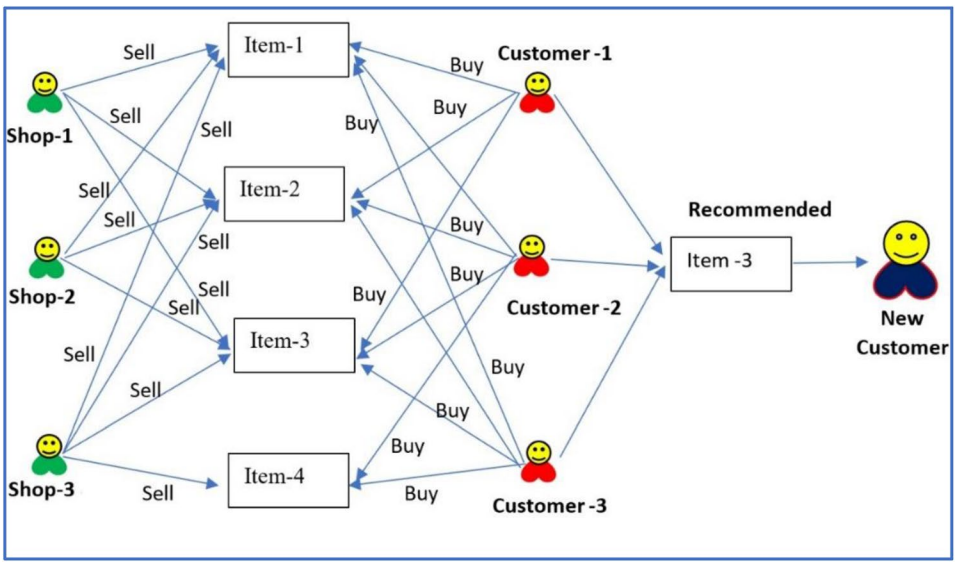


Fig. 2 Collaborative recommendation based product recommendation

Fig. 3 P-P similarity

similarity is very complicated to compare millions of customers associated with the product. The product is recommended by the customer, who has more experience. Figure 3 shows the customer who is recommending the third product, where comparing with the other two, third one is the best in number of products and has a lesser cost. To find out similar products based on the customer ratings, P-P similarity matrix is constructed and it is given in Table 2. Similarity amongst two different products is the Euclidean distance in space on which its dimensions defines the products features. If there exists the minimal distance amongst the product then the products similarity is higher, while higher distance will illustrate the minimal degree of similarity. Like product similarity, customer similarity also has the value where the lower Euclidean distance amongst two customers has higher customer similarity whereas higher the distance defines the degree of similarity between tow customers if minimal.

The rows and the columns in the P-P matrix is represented by products and customer. There are n number of products given in each column and the m number of customers given in each row. Each cell in the matrix shows the ratings Sim_{ij} given by the customer C_i for the product P_j. The symmetric matrix $n \times m$ is constructed using the rating value, is the similarity between two products. The similarity is calculated using cosine similarity value between two products. Two products are selected based on the similar and high ratings given by the customer. In the cosine similarity values such as, '1' signifies the maximum similarity

Table 2 P-P similarity matrix

	$\mathbf{P_1}$	$\mathbf{P_2}$	…	$\mathbf{P_j}$	..	$\mathbf{P_{n-1}}$	$\mathbf{P_n}$
C_1	1	Sim_{12}	…	Sim_{1j}	…		
C_2		1	…		…		
:	:	:	…		…	:	:
C_i		Sim_{i2}	…	Sim_{ij}	…		
:	:	:	…		…	:	:
C_{k-1}			…		…	1	
C_k			…		…		1

and '0' signifies the lowermost similarity. Sim_{ij} represents the ith customer ratings for jth product.

Cosine similarity is utilised as a parameter for calculating the distance where the vectors magnitude is not taken into account. It is much easier for using it for data with text demonstrated by word count. Error in testing is minimal using cosine normalization than the weight, layer and various other normalization. Cosine cantered normalization further lesser the error in testing. This normalization using cosine is very stable when compared to other non-normalization methods. Quantum computing makes sampling of matrix easy for making recommendations. Here in this paper, quantum method that samples from every row $(\widehat{U_h})j$ in a matrix mn. Normally, quantum algorithm will not yield the row $(\widehat{U_h})j$, which itself takes linear in dimension n with time, but uses row for sampling. But it is perfectly what every recommendation requires: sample a greater value component of the row, somewhat clearly output the row entirely. More correctly, the method of quantum procedure is effectively described which states that it takes input as a matrix, vector and a threshold constraint and produces the corresponding quantum state for vector projection onto the spanned space by singular row vectors of matrix whose value of singular is higher than threshold. Using this procedure, the output is clear on how to produce the sample by calculating the quantum state on the basis of computation.

Finding similar products using P-P similar matrix

For example, the customer C10 likes the product P1, P7 and P15. Then the P-P matrix provides the similar products as,

$$P_1 = \{P_4, P_5, P_6\},$$
$$P_7 = \{P_4, P_8, P_9\},$$
$$P_{15} = \{P_{10}, P_{11}, P_{12}\}.$$

Product P1 similar to P4, P5 and P6, P7 is similar to P4, P8 and P9, and P15 is similar to P10, P11 and P12. Now P4 is the common product where it is similar to P1 and P7, hence P4 is recommend to customer C10. Similar to the above, the P-P similarity works. One of the major advantages of P-P similarity is the ratings given for a product cannot be changed after predefined period. For example, consider the song "Believer". In the initial period, most of the people gave ratings of 3–4 out of 5. But after a month, people realised that the "Believer" is a great song and they are not able to change. As a rule of thumb, number of customers is more than products, if the product ratings do not modify much over time after the period, then P-P similarity-based recommendation system is highly preferred for shop-shop based recommendation.

Data collection model of shoppers shop system

Using the collaborative and P-P similarity calculation the proposed approach processes the testing data and provide the recommendation results. Social media and platform have now become the most utilized means to communicate any information among people. There is a huge impact of social platforms such as Twitter, Facebook and WhatsApp across the daily routine of a normal person. The trend has changed to such an extent that people share their experience, mood and current state of mind either in the form of a status or as a review comment. These status or review comments gets shared across a chain of linked customers and in turn acts as an indirect means of marketing the customer experience. Figure 4 depicts this scenario where the customer shares his experience by posting his experience across social media platform such as twitter or Facebook.

The proposed shoppers shop system is focussed towards gathering the insightful information pertaining to the customers shopping experience in a shop which may be a retail outlet, super market, retail distributor shop, online purchase etc. Figure 5 depicts the block diagram of a shopper's shop system. The customer has two modes of performing the shopping namely online mode where in the customer purchases the product by placing orders through online websites or a direct mode where in the customer visits the shop which may be a retail outlet, distributor stall, shopping mall etc. The customer then shares his opinion or experience by posting the same across any social media portals such as Twitter, Facebook, Instagram, WhatsApp etc. as a status or as comments or feedback or sometimes pictures with some emotions. All these experiences are recorded through the website in a social media database.

The initial step will be data extraction. Twitter or any other social media database comprises of huge chunk of data pertaining to all the customer transactions. The primary objective of data extraction is to extract the data pertaining to the customer transactions related to shopping alone, thus extracting only the essential data for further processing and discarding the remaining data from being considered for further processing. The extracted shopping related data are then subjected to pre-processing. Data pre-processing comprises of considering the extracted data as input and involves removing redundant, duplicated and unwanted noisy data to ensure that only data relevant to

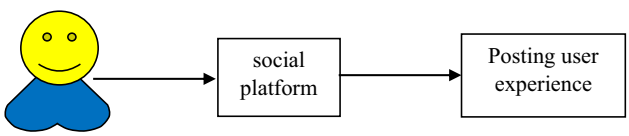


Fig. 4 Current scenario where customer shares his experience in social media

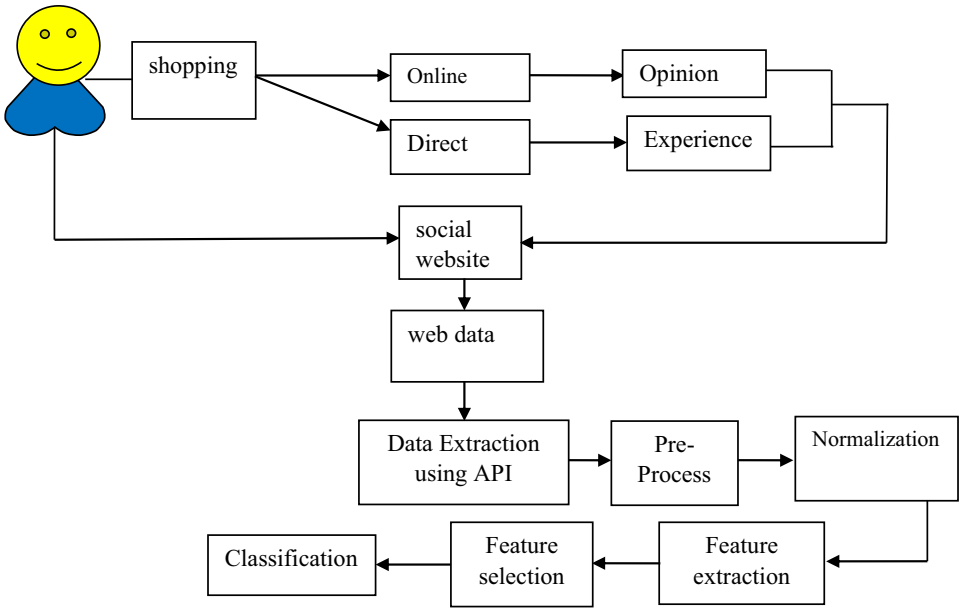


Fig. 5 Overall functionality of the proposed recommendation system

the context percolates to the next stage. Thus, all the irrelevant data are completely removed and only required data remains in the system for further processing. In the context of this work, pre-processing is highly essential in ensuring that duplicate remarks from the same customer pertaining to a same shop and for the same product is removed. Also, all the irrelevant information is removed as part of this stage.

Pre-processing is followed by normalization. Normalization is a technique where in the data is transformed and fitted into a scale that is highly suitable for the system. The primary objective of normalization is to transform the data into the best suitable form such that it enables easy data analytics and easier processing. Data analytics in shoppers shop application comprises of feature extraction and feature selection.

Feature extraction is the process of extracting essential features that are needed for data analysis. The data may comprise of n number of features. However, there may be certain salient features using which the data can be completely reconstructed. In the case of shopper's shop system, the shopping data may consist of different features and each feature may comprise of different dimensions. Feature selection comprises of selecting only the essential features from the extracted features of the data. Only five features namely Manufacturing date (MFD), Manufacturing price (MRP), discounts and offers, quality ratings and suggestions or reviews are considered as salient features and the information pertaining to the entire shopping transactions in a shop can be depicted using these features and the customer experience can be portrayed using these features.

Feature extraction

Data extraction demonstrates some original transformation of features to produces certain other features that has more momentous. Feature extraction is normally utilised to construct mean of linear combination $\propto$ Ry of unceasing features which consists of discriminant. It is mainly utilised in the context of reducing complexity and forecast some useful information from the gathered data. Here the shopping data gathered are enforced for data extraction as it may include some redundancy in it. Extraction techniques such as principle component analysis may be used to analyse the drawbacks in the database and produce a solid data for further investigation used the proposed model.

Once, the feature extraction and feature selection are completed, the refined data is then subjected to classification. The classification model in this shopper's shop system uses machine learning based approach namely regression-based recommendation system to perform multi class classification. The classifier analyses the five selected features and based on the values, determines whether the shop where the product was purchased provided good, average or bad customer shopping experience.

Experimental results and discussion

The proposed machine learning approach is implemented, experimented, and the results are verified. Programming and executing the proposed approach, the system configuration used is given in Table 3. The dataset is taken from various data publicly available in "https://jmcauley.ucsd.edu/data/amazon/links.html" [16, 17]. The complete

Table 3 System requirement for experiment

Element	Performance
OS	Windows
CPU	Intel Core i7, 7th Gen Processor @3.2 GHz
Memory	8 GB RAM
HDD	1 TB
Software	MATLAB
Data set	*Amazon*—product recommendation data sets

explanation about the dataset is given in [18]. The dataset comprises of various product information, reviews and meta information taken from Amazon, Shop clues and Flicker websites. The total number of reviews available in the dataset is 142.8 million has been given up to 2014. Including the data collected from [16–18], various product dataset given in [19] is also taken for experiment. From the experiment the performance of the proposed ML algorithm is verified. Since, ML based predicting the ratings is true, the error values are calculated for finding the incorrect recommendation output. It takes very less time and fast, because the number of incorrect recommendations is very less than the correct one. The error values are calculated like supervised learning methods discussed in [20, 21]. The performance comparison is more accurate since the dataset has predefined recommendation result. It is assumed that there is a portion of recommendation results is not correct. Using the difference between the correct and in-correct answers is calculated and it is defined as error value.

The main objective of this paper is to create a well suitable and highly applicable recommendation system for shoppers' shop. To improve the efficiency and provide an optimal solution for recommendation system, this proposed HRS use the collaborative and P-P similarity algorithms for predicting customer ratings-based product and shops. The problem is considered as the regression problem and it is implemented in MATLAB software and the results are verified. The volume of data used for training process is 100,400,000 ratings where 400,000 customers given for 25,000 products. The products include clothes, kitchen products, cosmetics, electrical and electronic items used in our daily life. The format of the data (see Table 3) used in the training process is

< customer-ID(CID), product-ID(PID), rating, date, shop >

where, CID, PID and rating are integers. The rating range is between 1 and 5. In general, the ratings are given as stars. In this paper, to reduce the computational complexity, the number of stars is counted and an equal integer value is given. For each product and the customer, it is essential to predict the ratings given by the customer. The recommendation system is considered as **regression problem**. The machine learning algorithm is used for minimising the RMSE.

Modelling the recommendation system with the help of a regression model using mapping function (f) with approximation from the input variable (Y) to an output variable (X) which is continuous. Real value is the variable in the output, for instance it may be integer or floating value. These are represented often by quantities such as sizes and amounts. Various models can be adopted for recommendation system but linear regression is utilised for analysing the best hyper plane which makes the prediction of data effective.

Objectives

It can be obtained only by predict the rating values given by the customer to a product which is not already rated. One of the main objectives of this experiment is to calculate the real rating value by finding the difference between actual rating (MAPE) and predicted rating (RMSE).

Constraints

The application and the data have a defined interoperability.

There is no recommended product with less latency have not been precomputed in the existing researches.

Type of data

The volume of data collected is more than 50,000 and we obtained only good, complete data for the experiment to avoid manual and computational errors with time complexity.

1. There are 35,500 unique product ID.
2. There are 85,000 unique customer ID.
3. Ratings with 1 to 5 values (it may be indicated in stars).
4. Each product with the shop name, MRD, MRP, SP, quality and ratings are collected.

Sample data (one row) from the data file are given in the following Table 4. The data selected based on the best-selling price, quality and ratings, the product and shop name are given in Table 4, which is used for experiment. It says that MRD is the latest date, MRP is the original given by the production company, SP is given by the shop after offer, the quality of the product and the highest ratings.

Table 4 Data size: category wise

Product category	Number of products	Total number of ratings
Electronics	6,498,196	27,824,482
Beauty	9,259,204	12,023,070
Food	1,671,760	19,297,156
Clothes	19,503,384	5,748,920
Total	36,932,544	64,893,628

Table 5 Sample data

Sr. no	CID	PID	Rating	Date	Shop
1	1209954	093456	5	2005–05–09	0942
2	2263586	085632	4	2004–08–20	0452
3	1009622	082843	4	2005–01–19	0942
4	1910569	075634	3	2004–04–12	0942
5	401047	081263	4	2005–06–03	0983

Table 6 Time complexity

Process	Time (min)
Reading and converting into known format	1.45
Preprocessing	3.87
Clustering	7.34
Product recommendation	2.9
Shop recommendation	1.9

The data are clustered based on the categories, product dataset has a greater number of products and sub items.

Process on the data

Initially read the datafile, eliminate the unwanted spaces, trailing semicolon, and separate each field by comma. The time taken for compiling and executing each process of the total proposed approach is given in the Tables 5 and 6 for understanding the time complexity.

Next, the basic statistics of the data are considered for verifying the efficiency of the approach. The size of the total data, data used in training process and testing process are computed and given in Table 7. The data size is verified from actual data collected, divided for training and testing process. It can determine the efficiency of the proposed approach. From Table 7, it is noticed that, the data loss before and after processing is not high, and ML algorithm saves the data.

Then the result obtained from the experiment is clustered data. After clustering, the data are arranged (it is given for the sample data taken from the actual data). The product is clustered based on the product-categories such as cosmetics, groceries, Medical, ready-food and etc. Some of the products are clustered under same cluster with high ratings are recommended. Then those products selling by similar companies are considered for recommendation. In this paper Table 8a and b shows the products and the related shops are recommended for the new customer (highlighted in the Table).

Table 7 Data size verification

Total Data

Concern	Values
Number of Product Ratings	64,893,628
Number of Customer	508,000
Number of unique products	10,000

Trained Data

Concern	Values
Total Number of Product Ratings in trained data	40,000,000
Number of unique Customers in train data	400,000
Number of unique products in train data	10,000
Highest value of customer ID	2,345,121
Highest value of a product ID	18,000

Testing Data

Concern	Values
Total Number of Product Ratings in Test data	10,000,000
Number of unique Customers in Test data	300,000
Number of unique products in Test data	9,000
Highest value of customer ID	2,650,500
Highest value of a product ID	18,760

Table 8 **a** Product based clustering, **b** Product – shop based clustering

a

Sr. No.	PID	CLID	Rating	Date	Shop
1	093456	13	5	2005-05-09	0942
2	085632	18	4	2004-08-20	0452
3	082843	13	4	2005-01-19	0942
4	075634	19	3	2004-04-12	0942
5	081263	11	4	2005-06-03	0983

b

Sr. No.	PID	CLID	Rating	Shop
1	093456	13	5	Amazon
2	085632	18	4	Flicks
3	082843	13	4	Amazon
4	075634	19	3	Shop clues
5	081263	11	4	Amazon

For example, the product id 093456, 082843 have obtained rating value 5 and both the products is selling by the shop amazon. Based on the rating value and the similar shop the shopper's shop is amazon. Since the product has high rating value and both the products are selling by Amazon, it is recommended by the hybrid recommendation system. Similarly, the entire dataset is processed. To improve the quality of the recommendation 40% of the dataset is used for training process and the recommended products are updated for future testing process. Whenever a new product or customer looks for a product it is compared with the trained process and apply recommendation. If the product is not available in the list then it will be applied for testing process. Testing process is applying the same process applied in the training process and obtain the recommendation result as "1" or "0".

The size of the total data and the ratings are high, it is not able to count it by simple addition or multiplication method. To obtain the accurate ratings initially the rating distributed from 1 to 5 is calculated for all the products. It makes easy to classify the ratings for a specific product and shop. Figure 6 shows the rating distribution of the training data set. The ratings have been categorized into five values with 1 being the least rating and 5 being the maximum rating value across different products (Table 9).

Fig. 6 Rating distribution calculation

Table 9 Product versus recommendation

Product category	Total number of products	Number of products recommended
Food	10,000	9899
Clothes	10,000	9945
Electronics	10,000	9965
Kitchen	10,000	9389

The set of all performance measures used to determine the accuracy of the proposed approach is MAE (Mean absolute error), MSE (Mean squared error) and MAPE (mean absolute percentage error), and are calculated using the following equation.

Mean absolute error is used to estimate the inaccuracy amongst combined findings expressing the similar phenomenon. MAE is similar metric used for estimating the error caused in forecasting time series analysis, sometimes utilised in confusion having huge standard definition.

$$\mathrm{MAE} = \frac{1}{n}\sum_{1}^{n}\left|d_i - \hat{d}_i\right|$$

$$\mathrm{MPAE} = \frac{100}{n}\sum_{1}^{n}\frac{d_i - \hat{d}_i}{d_i}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{1}^{n}\left(d_i - \hat{d}_i\right)^2$$

where d_i represents the actual rating value, $\hat{d}_i$ represents the predicted rating value,n represents the total number of products. Mean Absolute Error (MAE) denotes the measure of differences persisting between two continuous variables while performing regression analysis. MAE provides the average vertical distance between a variable and identity line. Figure 7 strikes a comparison of MAE values for the proposed Hybrid Recommendation System (HRS) and its contemporary methods. It can be observed that the MAE values are significantly lower when compared to other approach which is a clear indicator that the proposed HRS outperforms other existing approaches.

Mean absolute percentage error (MAPE) is used to denote the prediction accuracy of any forecasting approach. Greater the MAPE value, the accuracy is said to be precise. Figure 8 strikes a comparison of MAPE values with existing methods and it can be observed that the proposed HRS system has higher values of MAPE and the accuracy is nearly 98%. Thus it can be observed that the proposed Hybrid Recommendation System is highly accurate and provides accurate customer recommendations when compared to other contemporary approaches such as collaborative filtering with machine learning (CBF-ML), smart recommender hybrid system of learning (SRHL) and Collaborative filtering with machine learning (CF-ML).

Mean Square Error (MSE) is the deviation of the square of the estimated value from that of the actual value. This is another key metric that determines the deviation and provides a clear indicator in terms of determining the accuracy of the estimation technique. Lesser MSE value indicates very minimal deviation between the estimated and observed values. Figure 9 provides a comparison of MSE values of the proposed HRS with other existing approaches. It can be observed that the MSE value for HRS is 6 which is significantly lesser when compared to SRHL which has 7, CBF-ML has 13 and CF-ML has 17 as MSE value. Lower the MSE value higher the performance of the recommendation system.

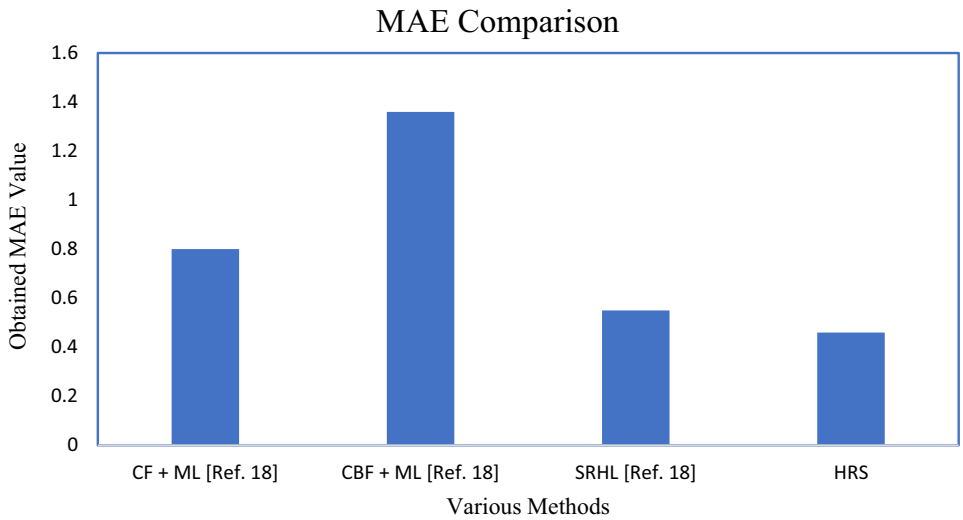


Fig. 7 MAE comparison

Fig. 8 MAPE comparison

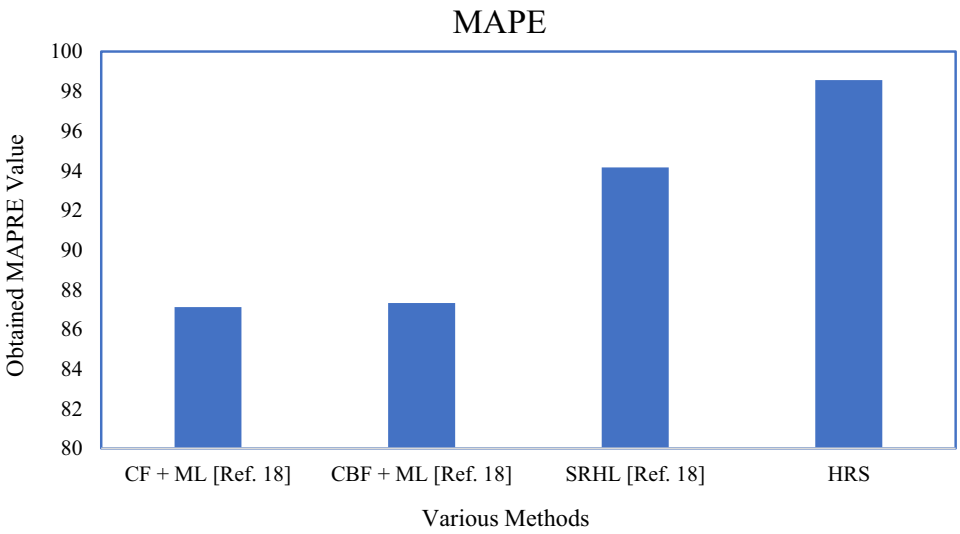
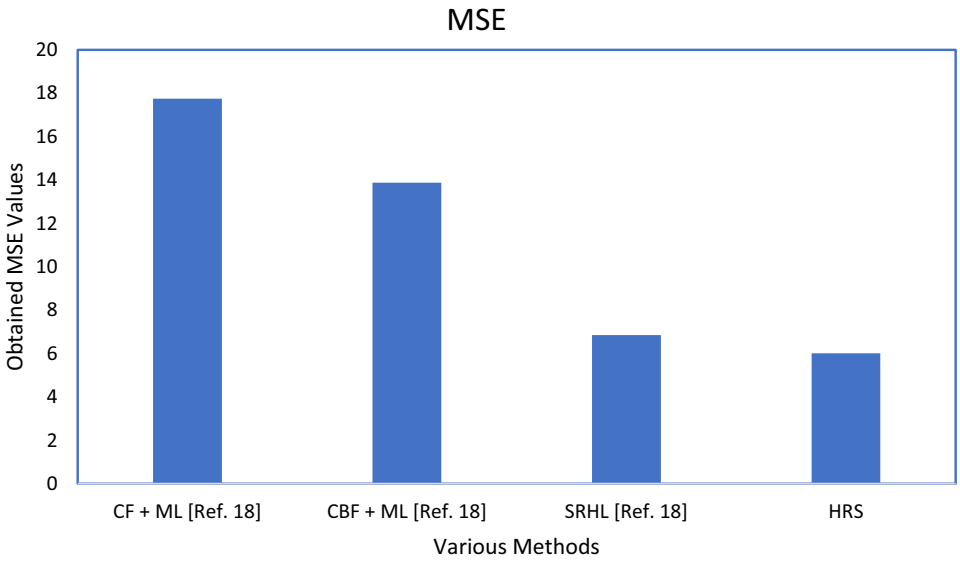


Fig. 9 MSE comparison



Conclusion

This work was directed towards developing a suitable hybrid recommendation system that can self-study the customer shopping data that is available in the form of reviews, understand the pattern and predict the interest of customers towards buying a particular product in selected shop. Five essential features were extracted from the customer data and using this, the Hybrid Recommendation System (HRS) was designed using machine learning based regression model. This system has been found to be instrumental in classifying the preferred choice of shops based on the products purchased by the customer. The salient feature of this HRS approach is that there is zero intervention of human element involved when it comes to predicting the customer choice of shops.

The performance of this Hybrid Recommendation System was evaluated using three metrics namely MAE (Mean absolute error), MSE (Mean squared error) and MAPE (mean absolute percentage error) and the results were compared and analyzed with other contemporary approaches. These results that were discussed in the previous section show that the MAE value of HRS is significantly lower than the existing approaches that were compared. Also, the MAPE value for HRS was nearly 98% which is a clear indicator of a high degree of accuracy. Similarly, the MSE value for HRS showed minimal deviation which is another strong indicator of high degree of precision and accuracy. Thus, it

can be concluded that the proposed Hybrid Recommendation System (HRS) clearly outperforms other contemporary approaches in terms of accurate prediction of customer sentiment with respect to shopping a product in a particular shop. Future work can be extending this approach towards gathering the customer interest for multiple products across different geographical locations.

References

1. Jianqiang Z, Xiaolin G, Xuejun Z (2018) Deep convolution neural networks for twitter sentiment analysis. IEEE Access 6:23253–23260. https://doi.org/10.1109/ACCESS.2017.2776930
2. Jianqiang Z, Xiaolin G (2017) Comparison research on text preprocessing methods on twitter sentiment analysis. IEEE Access 5:2870–2879. https://doi.org/10.1109/ACCESS.2017.2672677
3. Bouazizi M, Ohtsuki T (2018) Multi-Class sentiment analysis in twitter: What if classification is not the answer. IEEE Access 6:64486–64502. https://doi.org/10.1109/ACCESS.2018.2876674
4. Bouazizi M, Ohtsuki T (2017) A pattern-based approach for multi-class sentiment analysis in twitter. IEEE Access 5:20617–20639. https://doi.org/10.1109/ACCESS.2017.2740982
5. Ebrahimi M, Yazdavar AH, Sheth A (2017) Challenges of sentiment analysis for dynamic events. IEEE Intell Syst 32(5):70–75. https://doi.org/10.1109/MIS.2017.3711649
6. Iqbal F et al (2019) A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. IEEE Access 7:14637–14652. https://doi.org/10.1109/ACCESS.2019.2892852
7. Tan S et al (2014) Interpreting the public sentiment variations on twitter. IEEE Trans Knowl Data Eng 26(5):1158–1170. https://doi.org/10.1109/TKDE.2013.116
8. Liu S, Cheng X, Li F, Li F (2015) TASC:topic-adaptive sentiment classification on dynamic tweets. IEEE Trans Knowl Data Eng 27(6):1696–1709. https://doi.org/10.1109/TKDE.2014.2382600
9. Bouazizi M, Ohtsuki T (2019) Multi-class sentiment analysis on twitter: classification performance and challenges. Big Data Min Anal 2(3):181–194. https://doi.org/10.26599/BDMA.2019.9020002
10. Trilla A, Alias F (2013) Sentence-based sentiment analysis for expressive text-to-speech. IEEE Trans Audio Speech Lang Process 21(2):223–233. https://doi.org/10.1109/TASL.2012.2217129
11. Yu D, Xu D, Wang D, Ni Z (2019) Hierarchical topic modeling of twitter data for online analytical processing. IEEE Access 7:12373–12385. https://doi.org/10.1109/ACCESS.2019.2891902
12. Lamb AF, Varadarajan M, Tran R, Vandier N, Doshi BL, Bear C (2012) The vertica analytic database: C-store 7 years later. arXiv:1208.4173
13. Greenplum Database. greenplum.org/
14. Solutions TW (2002) Teradata Database technical overview, pp 1–7. http://www.teradata.com/brochures/Teradata-Solution-Technical-Overview-eb3025
15. Data Driven Investor – Medium (2020) https://medium.com/datadriveninvestor. Retrieved 6 June 2020
16. Ni J, Muhlstein L, McAuley J (2019) Modeling heart rate and activity data for personalized fitness recommendation. In: WWW'19: proceedings of the 2019 World Wide Web conference, San Francisco, CA, USA, May 2019
17. He R, Kang W-C, McAuley J (2017) Translation-based recommendation. In: Proceedings of the eleventh ACM conference on recommender systems, 2017
18. McAuley J (2020) Recommender systems datasets. https://cseweb.ucsd.edu/~jmcauley/datasets.html. Retrieved 6 June 2020
19. Kretser A, Murphy D, Starke-Reed P (2017) A partnership for public health: USDA branded food products database. J Food Compos Anal 64:10–12
20. Elahi M, Ricci F, Rubens N (2016) A survey of active learning methods in collaborative filtering recommender systems. Comput Sci Rev 20:29–50
21. Nouh RM, Lee H-H, Lee W-J, Lee J-D (2019) A smart recommender based on hybrid learning methods for personal well-being services. Sensors 19(431):1790–1801