Journal of Big Data

RESEARCH Open Access

Improving prediction with enhanced Distributed Memory-based Resilient Dataset Filter

Sandhya Narayanan[1*], Philip Samuel[2] and Mariamma Chacko[3]

*Correspondence:
nairsands@gmail.com
[1] Information Technology, School of Engineering, Cochin University of Science & Technology, Kochi 682022, India
Full list of author information is available at the end of the article

Abstract

Launching new products in the consumer electronics market is challenging. Developing and marketing the same in limited time affect the sustainability of such companies. This research work introduces a model that can predict the success of a product. A Feature Information Gain (FIG) measure is used for significant feature identification and Distributed Memory-based Resilient Dataset Filter (DMRDF) is used to eliminate duplicate reviews, which in turn improves the reliability of the product reviews. The pre-processed dataset is used for prediction of product pre-launch in the market using classifiers such as Logistic regression and Support vector machine. DMRDF method is fault-tolerant because of its resilience property and also reduces the dataset redundancy; hence, it increases the prediction accuracy of the model. The proposed model works in a distributed environment to handle a massive volume of the dataset and therefore, it is scalable. The output of this feature modelling and prediction allows the manufacturer to optimize the design of his new product.

Keywords: Distributed Memory-based, Resilient Distribution Dataset, Redundancy

Introduction

Analyzing and processing massive volumes of data in different applications like sensor data, health care and e-Commerce require big data processing technologies. Extracting useful information from the enormous size of unstructured data is a crucial thing. As the amount of data becomes more extensive, sophisticated pre-processing techniques are required to analyze the data. In social networking sites and other online shopping sites, a massive volume of online product reviews from a large size of customers are available [1]. The impact of online product reviews affects 90% of the current e-Commerce market [2]. Customer reviews contribute the product sale to an extent and product life in the market depends on online product recommendations.

Online feedback is one of the communication methods which gives direct suggestions from the customers [3, 4]. Online reviews and ratings from customers are another information source about product quality [5, 6]. Customer reviews can help to decide on a new successful product launch. Online shopping has several advantages over retail shopping. In retail shopping, the customers visit the shop and receive price information but less product

Springer Open

information from shop owners. On the other hand, online shopping sites give product reviews and previous customer feedbacks without extra cost and effort for the customers [7–10].

Investing in poor quality products potentially affects an industry's brand loyalty and this strategy should be changed by the eCommerce firms [5, 11]. Consumer product success depends on different criteria, such as the quality of the product and marketing strategies. The users should provide their valuable and accurate reviews about the products [12]. Customers bother to give reviews about products, whether they liked it or not. If the users provide reviews, then other retailers can create some duplicated reviews [13, 14]. In online marketing, the volume and value of product reviews are examined [15, 16]. The number of the product reviews on the shopping sites, blogs and forums has increased awareness among the users. This large volume of the reviews leads to the need for significant data processing methods [17, 18]. The value is the rating on the products. The ratio of positive to negative reviews about the product leads to the quality of the product [19, 20].

Feature selection is a crucial phase in data pre-processing [21]. Selecting features from an un-structured massive volume of data reduce the model complexity and improves the prediction accuracy. Different feature selection methods existing are the filter, wrapper and embedded. The wrapper feature selection method evaluates the usefulness of the feature and it depends on the performance of the classifier [22]. The filter method calculates the relevance of the features and analyzes data in a univariate manner. The embedded process is similar to the wrapper method. Embedded and wrapper methods are more expensive compared to the filter method. The state-of-art methods in customer review analysis generally discuss on categorizing positive and negative reviews using different natural language processing techniques and spam reviews recognition [23]. Feature selection of customer reviews increases prediction accuracy, thereby improves the model performance.

An enhanced method, which is a combination of filter and wrapper method is proposed in this work, which focuses on product pre-launch prediction with enhanced distributive feature selection method. Since many redundant reviews are available on the web in large volumes, a big data processing model has been implemented to filter out duplicated and unreliable data from customer reviews in-order to increase prediction accuracy. A scalable big data processing model has been applied to predict the success or failure of a new product. The realization of the model has been done by Distributed Memory-based Resilient Dataset Filter with prediction classifiers.

This paper is organized as follows. "Related work" section discusses related work. "Methodology" section contains the proposed methodology with System design, Resilient Distributed Dataset and Prediction using classifiers. "Results and discussions" section summarizes results and discussion. The conclusion of the paper is shown in "Conclusion and future work" section.

Related work

Makridakis et al. [24] illustrate that machine learning methods are alternative methods for statistical analysis of multiple forecasting field. Author claims that statistical methods are more accurate than machine learning [25] methods. The reason for less accuracy is the unknown values of data i.e., improper knowledge and pre-processing of data.

Different works have been implemented using the Matrix factorization (MF) [14] method with collaborative filtering [26]. Hao et al. [15] focused on a work based on the factorization of the user rating matrix into two vectors, i.e., user latent and item latent with low dimensionality. The sum of squared distance can be minimized by training a model that can find a solution using Stochastic Gradient Decent [27] or by least squares [28]. Salakhutdinov et al. [29] proposed a method that can be scaled linearly by probability related matrix factorization on a big volume of datasets and then comparing it with the single value decomposition method. This matrix factorization outperforms other probability factorization methods like Bayesian-based probabilistic analysis [29] and standard probability-based matrix factorization methods. A conventional approach, like traditional collaborative Filtering [13, 30] method depends on customers and items. The user item matrix factorization technique has been used for implementation purpose. In the recommender system, there is a limitation in the sparsity problem and cold start problem. In addition to the user item matrix factorization method, various analyses and approaches have been implemented to solve these recommendation issues.

Wietsma et al. [31] proposed a recommender system that gives information about the mobile decision aid and filtering function. This has been implemented with a study of 29 features of student user behavior. The result shows the correlation among the user reviews and product reviews from different websites. Jianguo Chen et al. [32] proposed a recommendation system for the treatment and diagnosis of the diseases. For cluster analysis of disease symptoms, a density-peaked method is adopted. A rule-based apriori algorithm is used for the diagnosis of disease and treatment. Asha et al. [33] proposed the Gini-index feature method using movie review dataset. The sentimental analysis of the reviews are performed and opinion extraction of the sentences are done. Gini-index impurity measure improves the accuracy of the polarity prediction by sentimental analysis using Support vector machine [34, 35]. Depending on the frequency of occurrence of a word in the document, the term frequency is calculated and opinion words are extracted using the Gini-index method. In this method, high term frequency words are not included, as it decreases the precision. The disadvantage of this method is that for the huge volume of data, the prediction accuracy decreases.

Luo et al. [36] proposed a method based on historical data to analyze the quality of service for automatic service selection. Liu et al. [37] proposed a system in a mobile environment for movie rating and review summarization. The authors used Latent Semantic Analysis (LSA-based) method for product feature identification and feature-based summarization. Statistical methods [38] have been used for identifying opinion words. The disadvantage of this method is that LSA-based method cannot be represented efficiently; hence, it is difficult to index based on individual dimensions. This reduces the prediction accuracy in large datasets.

Lack of appropriate computing models for handling huge volume and redundancy in customer review datasets is a major challenge. Another major challenge handled in the proposed work is the existence of a pre-launch product in the industry based on the product features, which can be predicted based on the customer feedback in the form of reviews and ratings of the existing products. This prediction helps to optimize the design of the product to improve its quality with the required product features. Many of the relational database management systems are handling structured data, which is not scalable for big data that handles a large volume of unstructured data. This proposed model solves the problem of redundancy in a huge volume of the dataset for better prediction accuracy.

Methodology

A pre-launch product prediction using different classifiers has been analysed by huge customer review and rating dataset. The product prediction is done through the phases consisting of data collection phase, feature selection and duplicate data removal, building prediction classifier, training as well as testing.

Figure 1 describes the various stages in system design of the model. The input dataset consists of multivariate data which includes categorical, real and text data. Input dataset is fed for data pre-processing. Data pre-processing consists of feature selection, redundancy elimination and data integration which is done using Feature Information Gain and Distributed Memory-based Resilient Dataset Filter approach. The cleaned dataset is trained using classification algorithms. The classifiers considered for training are Support Vector Machine (SVM) and Logistic Regression (LR). Further the dataset is tested for pre-launch prediction using LR and SVM.

Data collection phase

This methodology can be applied for different products. Several datasets like Amazon and flip cart customer reviews are available as public datasets [39–41]. The dataset of customer reviews and ratings of seven brands of mobile phones for a period of 24 months are considered in this work. The mobile phones product reviews are chosen because of two reasons. New mobile phones are launched into the market industry day by day which is one of the unavoidable items in everyone's life. Market sustainability for the mobile phones is very low.

Table 1 shows a sample set of product reviews in which input dataset consists of user features and product features. User features consists of *Author*, *ReviewID* and *Title* depending on the user. Product feature consists of *Product categories, Overall ratings and Review Content*. Since mobile phone is taken as the product, the categorization is done according to the features such as *Battery life*, *price*, *camera*, *RAM*,

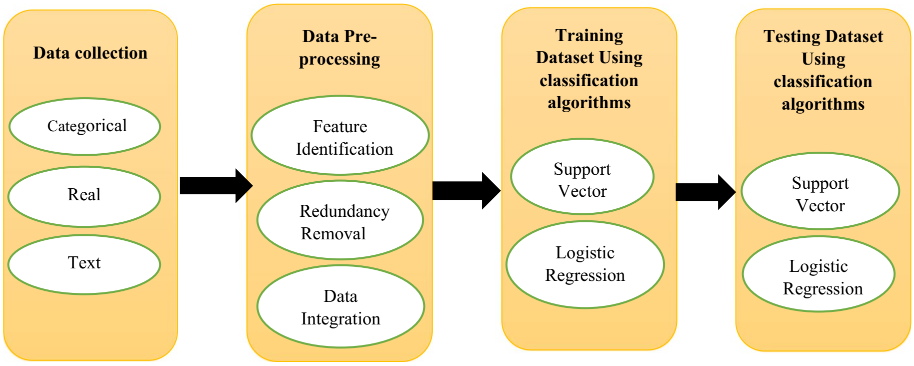
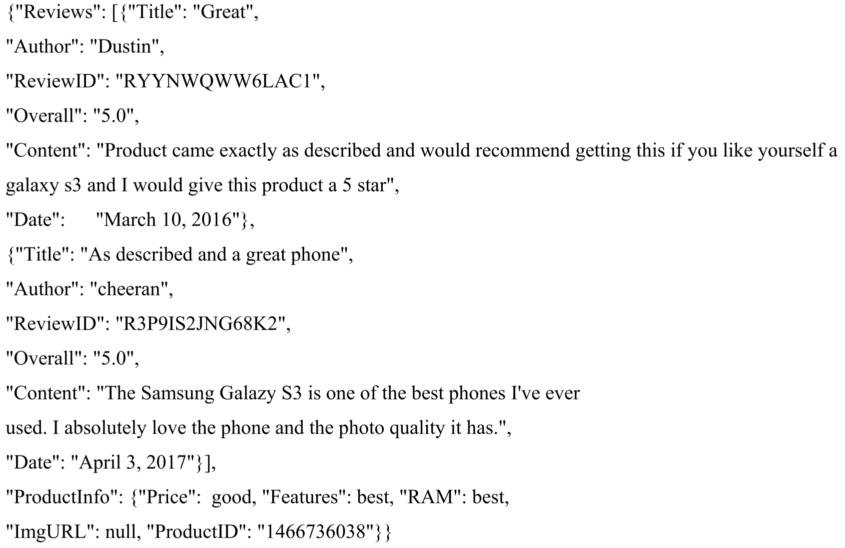


Fig. 1 Product prelaunch prediction System Design

Table 1 Sample set of Product Reviews



```
{"Reviews": [{"Title": "Great",
"Author": "Dustin",
"ReviewID": "RYYNWQWW6LAC1",
"Overall": "5.0",
"Content": "Product came exactly as described and would recommend getting this if you like yourself a
galaxy s3 and I would give this product a 5 star",
"Date":      "March 10, 2016"},
{"Title": "As described and a great phone",
"Author": "cheeran",
"ReviewID": "R3P9IS2JNG68K2",
"Overall": "5.0",
"Content": "The Samsung Galazy S3 is one of the best phones I've ever
used. I absolutely love the phone and the photo quality it has.",
"Date": "April 3, 2017"}],
"ProductInfo": {"Price":  good, "Features": best, "RAM": best,
"ImgURL": null, "ProductID": "1466736038"}}
```

processor, *weight* etc. Some features are given a priority weightage depending on the product and user requirements. Input dataset with JSON file format is taken.

Dataset pre-processing

In data pre-processing, feature selection plays a major role. In the product review dataset of a mobile phone, a large number of features exist. Identifying a feature from customer reviews is important for this model to improve the prediction accuracy. Enhanced Feature Information Gain measure has been implemented to identify significant feature.

Features are identified based on the content of the product reviews, ratings of the product reviews and opinion identification of the reviews. Ratings of the product reviews can be further categorized based on a rating scale of 5 (1—Bad, 2—Average, 3—Good, 4—very good, 5—Excellent). For opinion identification of the product, the polarity of extracted opinions for each review is classified using Senti-WordNet [42].

Feature Information Gain measures the amount of information of a feature retrieved from a particular review. Impurity which is the measure of reliability of features in the input dataset should be reduced to get significant features. To measure feature impurity, the best information of a feature obtained from each review is calculated as follows

- Let P_i be the probability of any feature instance (f) of k feature set $F = \{f_1, f_2, \ldots f_k\}$ belonging to ith customer review R_i, where i varies from 1 to N.
- Let N denotes the total number of customer reviews.
- Let O_R denotes the polarity of extracted opinions of the Review.
- Let S_R denotes product rating scale of review (R).

The information of a feature with respect to review rating and opinion is denoted by I_f

$$I_f = log_2\left(\frac{1}{P(R = F)}\right) * O_R * S_R. \tag{1}$$

Expected information gain of the feature denoted as E_f

$$E_f = \sum_{i=1}^{N} -P_i(R = F).\left\|I_f\right\|_1. \tag{2}$$

Review Feature Impurity R(I) is calculated as

$$R(I) = -\sum_{i=1}^{N} P_i.log_2 E_f. \tag{3}$$

Then Feature Information Gain (Δ_G) to find out significant features are calculated as

$$\Delta_G = R(I) - \sum_{i=1}^{N}\left[\left(\frac{O_R}{N} * E_f\right) - \left(\frac{S_R}{N} * E_f\right)\right]. \tag{4}$$

Features are selected based on the Δ_G value and those with an Information gain greater than 0.5 is selected as a significant feature. Table 2 shows the significant feature from customer reviews and ratings.

Next step is to eliminate the redundant reviews and to replace null values of an active customer from the customer review dataset using an enhanced big data processing approach. Reviews with significant features obtained from feature identification are considered for further processing.

Table 2 Significant Features from Customer Reviews and Ratings

No	Customer reviewed features	No	Customer reviewed features
1	Author	17	RAM
2	Title	18	Sim type
3	ReviewID	19	Product category
4	Content	20	Thickness
5	Product brand	21	Weight of mobile phone
6	Ratings	22	Height
7	Battery life	23	Product type
8	Price	24	Product rating
9	Feature information gain	25	Front camera
10	Review type	26	Back camera
11	Product display	27	Opinion of review
12	Processor	28	Multi-band
13	Operating system	29	Network support
14	Water proof	30	Quick charging
15	Rear camera	31	Finger sensor
16	Applications inbuilt	32	Internal storage

Resilient Distributed Dataset

Resilient Distributed Dataset (RDD) [43] is a big data processing approach, which allows to store cache chunks of data on memory and persevere it as per the requirements. The in-memory data caching is supported by RDD. Variety of jobs at a point of time is another challenge which is handled by RDD. This method deals with chunks of data during processing and analysis. RDD can also be used for machine learning supported systems as well as in big data processing and analysis, which happens to be an almost pervasive requirement in the industry.

In the proposed method the main actions of RDD are:

- Reduce (β): Combine all the elements of the dataset using the function β.
- First (): This function will return the first element
- takeOrdered(n): RDD is returned with first 'n' elements.
- saveAsSequenceFile(*path*): the elements in the dataset to be written to the local file system with given *path*.

The main Transformations of RDD are:

- map(β): Elements from the input file is mapped and new dataset is returned through function β.
- filter(β): New dataset is returned if the function β returns true.
- groupBykey(): When called a dataset of (key, value) pairs, this function returns a dataset of (key, value) pairs.
- ReduceBykey(β): A (key, value) pair dataset is returned, where the values of each key are combined using the given reduce function β.

In the proposed work an enhanced Distributed Memory-based Resilience Dataset Filter (DMRDF) is applied. DMRDF method have long Lineage and it is recomputed themselves using prior information, thus it achieves fault-tolerance. DMRDF has been implemented to remove the redundancy in the dataset for product pre-launch prediction. This enhanced method is simple and fast.

- Let the list of n customers represented as $C = \{c_1, c_2, c_3 \ldots, c_n\}$
- Let the list of N reviews be represented as $R = \{r_1, r_2, r_3 \ldots, r_N\}$
- Let x significant features are identified from feature set (F) represented as $F_x \subset F$
- An active customer consists of significant feature having information Gain value denoted by Δ_G

In the DMRDF method, a product is chosen and its customer reviews are found out. Eliminate customers with similar reviews on the selected product and also reviews with insignificant features. Calculate the memory-based Resilient Dataset Filter score between each of the customer reviews with significant features.

Let us consider a set C of 'n' number of customers, the set R of 'N' number of reviews and a set of significant features $'F_x'$ are considered. The corresponding vectors are represented as K_C, K_R and K_{F_x}. Then K_{R_i} is represented using a row vector and K_{F_j} is represented using the column vector. Each entry K_{C_m} denote the number of times the m$^{\text{th}}$ review arrives in customers. The similarities between ith review of mth customer is found out using L_1 norm of K_{R_i} and K_{C_m}. The Distributed Memory-based resilient filter score δ is calculated using the Eq. (5).

$$\delta = \sum_{\substack{i=1\\m=1}}^{\substack{N\\n}} \left(\frac{\left[K_{R_i} * \left(\sum_{j=1}^{x} K_{F_j}\right)\right] * K_{C_m}}{K_{R_i} \cdot K_{C_m}} \right) * |\Delta_G| \tag{5}$$

The δ score is calculated for each customer review whereas the score lies between [0,1]. The significant features are found out using Eq. 4. For customer reviews without significant features, Δ_G value will be zero. The reviews with δ score value 0 are found to be insignificant without any significant feature or opinion and hence those reviews are eliminated and not considered for further processing in the work. More than one Distributed Memory-based resilient filter score value is identified then the second occurrence of the review is considered as duplicate.

Prediction classifiers

Logistic regression and Support Vector Machine classifiers are the supervised machine learning approaches used in the proposed work for product pre-launch prediction.

Logistic regression (LR)

We have implemented proposed model using logistic regression analysis for prediction. This model predicts the failure or success of a new product in the market by analysing selected product features from customer reviews. A case study has been conducted using the dataset of customer reviews of mobile phones. Success or failure is the predictor variable used for training and testing the dataset. For training the model 75% of the dataset is used and for testing the model, remaining 25% is used.

- Let p be the prediction variable value, assigning 0 for failure and 1 for success.
- p_0 is the constant value.
- b is the logarithmic base value.

Then the logit function is,

$$L_0 = b^{p_0 + p \sum\limits_{i=1}^{x} f_i} \tag{6}$$

Then the Logistic regression value γ is shown in Eq. (7),

$$\gamma = \frac{L_0}{\left(b^{p_0 + p \sum_{i=1}^{x} f_i}\right) + 1} \tag{7.1}$$

$$= \frac{1}{1 + b^{-\left(b^{p_0 + p \sum_{i=1}^{x} f_i}\right)}} \tag{7.2}$$

The probability value of γ lies between [0,1]. In this work, if this value is greater than 0.5 the pre-launch prediction of the product is considered as success and for values less than 0.5, it is considered as failure.

Support Vector Machine (SVM)

SVM is the supervised machine learning method, used to learn from set of data to get new skills and knowledge. This classification method can learn from data features relationships (z_i) and its class $\left(y_i\right)$ that can be applied to predict the success or failure class the product belongs to.

- For a set T of t training feature vectors, $z_i \in R^D$, where i = 1 to t.
- Let $y_i \in \{+1, -1\}$, where +1 belongs to product success class and -1 belongs to product failure class.
- The data separation occurs in the real numbers denoted as X in the D dimensional input space.
- Let w be the hyper plane normal vector element, where $w \in X^D$.

The hyper plane is placed in such a way that distance between the nearest vectors of the two classes to the hyperplane should be maximum. Thus, the decision hyper plane is calculated as,

$$\alpha(w) = \frac{2}{\|w\|} \tag{8}$$

The conditions for training dataset $d \in X$, is calculated as

$$w^t z_i + d \geq 1, \quad \text{where} \quad y_i = +1. \tag{9}$$

$$w^t z_i + d \leq -1, \ \text{where} y_i = y_i - 1. \tag{10}$$

To maximize the margin the value of w should be minimized.

The products in the positive one class (+1) are considered as successful products, [from Eq. (9)] and those in the negative one class (−1) [from Eq. (10)] are in failure class.

Experimental setup

The proposed system was implemented using Apache Spark 2.2.1 framework. Spark programming for python using PySpark version 2.1.2, which is the Spark python API has been used for the application development. An Ubuntu running Apache web server using Web Server Gateway Interface is used. Amazon Web Services is used to run some components of the software system large servers (nodes), having two Intel Xeon E5-2699V4 2.2 G Hz processors (VCPUs) with 4 cores and 16 GB of RAM on different Spark cluster configurations. According to the scalability requirements the software components can be configured and can run on separate servers.

Results and discussions

To evaluate our prediction system several case studies have been conducted. Support Vector Machine and Logistic regression classifiers are employed to perform the prediction. Most significant customer review features are used to analyse the system performance. The prediction accuracy evaluation is taken as one of the system design factors. The system response time is another major concern for big data processing system. In the customer review feature identification, we propose feature information gain and DMRDF approach to identify significant features and to eliminate redundant customer reviews from the input dataset.

Figure 2 illustrates significant features required for the mobile phone sustainability. Customer reviews and ratings of 7 brands of mobile phones are identified and evaluated with DMRDF using SVM and LR. The graph shows the significant features identified by the model against the percentage of customers whose reviews are analysed. 88% of the customers identified internal storage as a significant feature. Product price has been identified by 79% of customers as significant feature. With this evaluation customer requirements for a product can be analysed in a better manner, thus can optimize the design of the product for better product quality and for product sustainability in the industry.

Figure 3 shows the comparison of the processing time taken by the proposed model with different dataset size against that of the state of art techniques. DMRDF method takes less time for completion of the application compared to other gini-index and latent semantic analysis methods. Hence the proposed model is fast and scalable. It provides a high-speed processing performance with large datasets. This shows the DMRDF applicability in big data analytics, whereas gini-index and LSA-based methods processing time is larger for large volume of dataset. From the Fig. 3 it can be seen that with 9 GB dataset time taken for prediction using LSA-based model, Gini-index model and DMRDF model is 342 s, 495 s and 156 s respectively. With 18 GB dataset time taken for prediction using LSA-based model, Gini-index model and DMRDF model 740 s, 910 s and 256 s respectively. Gini-index and LSA-based methods time taken for 18 GB dataset is twice that of 9 GB dataset. But for DMRDF model time taken for 18 GB dataset is 1.6 times that of

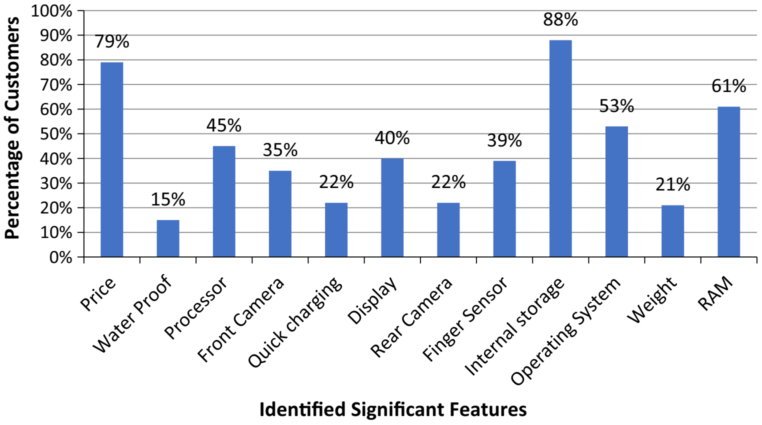


Fig. 2 Identified Significant Features from Customer reviews and Ratings

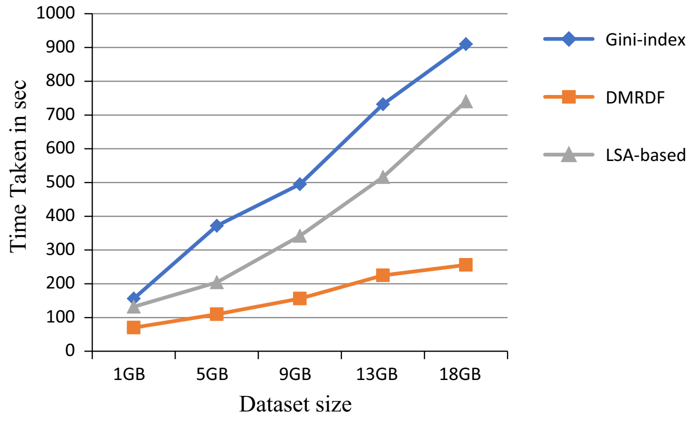


Fig. 3 Dataset Size versus Processing Time Graph

Table 3 Performance comparison of the proposed model with state of art techniques

Classifier	**Support vector machine**		
Method used	**P@R (precision)**		**PA % (prediction accuracy)**
DMRDF	0.941	0.92	95.4
LSA-based	0.894	0.79	87.5
Gini-index	0.66	0.567	83.2
Classifier	**Logistic regression**		
Method used	**P@R**	**R@R %**	**PA %**
DMRDF	0.915	0.849	93.5
LSA-based	0.839	0.753	83
Gini-index	0.62	0.52	79.8

9 GB dataset and also it is 3 times lesser than Gini-index method. DMRDF model has more advantage compared to the other state of art techniques in the case of application execution and performance.

The reliability of the methods considered for the pre-launch prediction depends on precision [44], recall and prediction accuracy measurement. Table 5 shows a comparison of precision, recall and accuracy measures of DMRDF, Gini-index and LSA-based methods with Support Vector Machine and Logistic Regression classifiers using customer reviews dataset over a period of 24 months. The results shown in Table 3 are best proved using DMRDF with Support Vector Machine classification with prediction accuracy of 95.4%. The DMRDF outperforms LSA-based and Gini-index methods in P@R, R@R and PA measures. Using proposed method, true positive (TP), false positive (FP), true negative (TN) and false negative (FN) are found out. The prediction accuracy (PA), precision (P@R) and recall (R@R) are computed using Eqs. (10), (11), and (12) respectively.

$$\mathrm{PA} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$\text{P@R} = \frac{TP}{TP + FP} \tag{11}$$

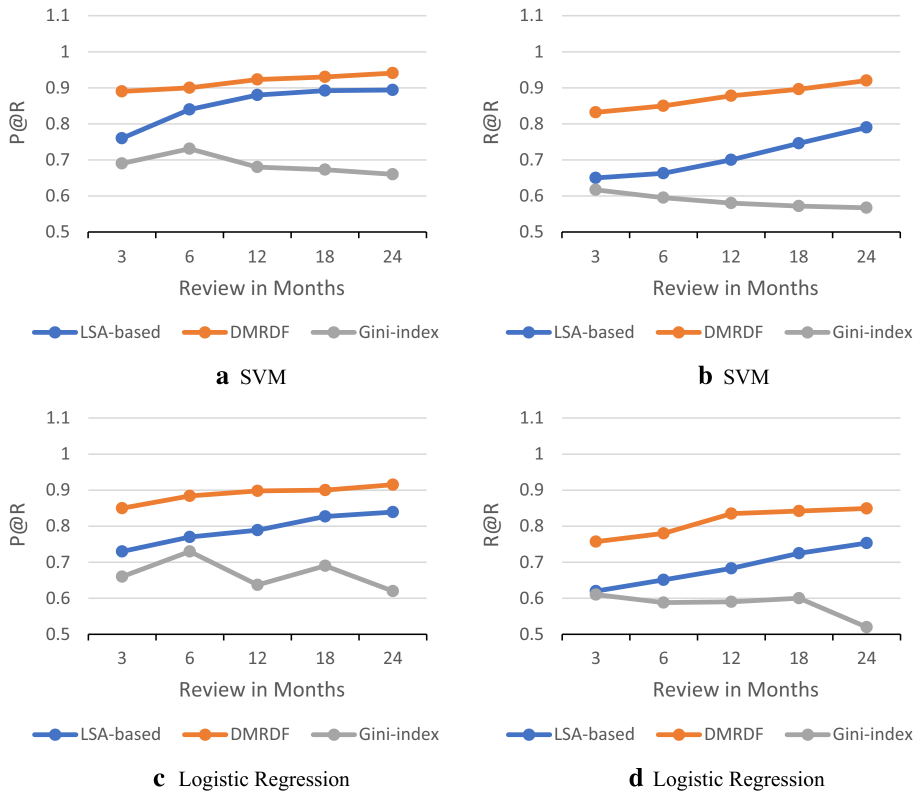
$$\text{R@R} = \frac{TP}{TP + FN} \tag{12}$$

Using DMRDF with SVM classifier and LR classifier, the prediction accuracy variations are less compared to LSA-based and Gini-index methods. Hence DMRDF outperforms the other two methods for customer review feature prediction.

Furthermore Fig. 4, shows the DMRDF, LSA-based and Gini-index approaches as applied to the customer reviews and ratings datasets for 3, 6, 12, 18 and 24 months. In DMRDF many features may appear in different customer review aspects, hence performance evaluation will not consider duplicate customer reviews. In Gini- index, features are extracted based on the polarity of the reviews and for large dataset P@R and R@R are less. The results show that DMRDF method outperforms the other two methods in big data analysis. Gini-index approach does not perform well in customer review feature prediction.

Conclusion and future work

Technological development in this era brings new challenges in artificial intelligence like prediction, which is the next frontier for innovation and productivity. This work proposes the implementation of a scalable and reliable big data processing model



a SVM

b SVM

c Logistic Regression

d Logistic Regression

Fig. 4 Precision and Recall of DMRDF, LSA-based and Gini-index methods using SVM and LR classifiers

which identify significant features and eliminates redundant data using Feature Information Gain and Distributed Memory-based Resilient Dataset Filter method with Logistic Regression and Support Vector Machine prediction classifiers. A comparison of the analysis has been conducted with state of art techniques like Gini-index and LSA-based approaches. The prediction accuracy, precision and recall of DMRDF method outperforms the other methods. Results show that the prediction accuracy of the proposed method increases by 10% using significant feature identification and elimination of redundancy from dataset compared to state of art techniques. Large feature dimensionality reduces the prediction accuracy of the LSA-based method where as number of significant features plays an important role in prediction modelling. Results show that proposed DMRDF model is scalable and with huge volume of dataset model performance is good as well as time taken for processing the application is less compared to state of art techniques.

Resilience property of DMRDF method have long lineage, hence this can achieve fault-tolerance. DMRDF model is fast because of the in-memory computation method. Proposed design can be extended to other product feature identification big data processing domains. As a future work, the model may be developed to make real time streaming predictions through a unified API that searches customer comments, ratings and surveys from different reliable online websites concurrently to obtain synthesis of sentiments with an information fusion approach. Since the statistical properties of customer reviews and ratings vary over time, the performance of machine learning algorithms can also come down. To cope with the limitations of deep learning matrix factorization integrated with DMRDF can be adapted.

Abbreviations

DMRDF: Distributed Memory-based Resilient Dataset Filter; FIG: Feature information gain; RDD: Resilient distributed dataset; SVM: Support vector machine; LR: Logistic regression; LSA: Latent semantic analysis; PA: Prediction accuracy; P@R: Precision; R@R: Recall; MF: Matrix factorization.

Acknowledgements

Not applicable.

Authors' contributions

SN designed and implemented the model for Pre-launch product prediction. SN analysed and interpreted the customer reviews and ratings dataset regarding the pre-launch product prediction. PS supervised the design, implementation and analysis of the model for pre-launch product prediction. MC was a major contributor in writing the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the Kaggle repository. [snap.stanford.edu/data/web-Amazon.html] [40] and [http://www.kaggle.com/PromptCloudHQ/flipkart-products] [39].

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Information Technology, School of Engineering, Cochin University of Science & Technology, Kochi 682022, India. [2] Department of Computer Science, Cochin University of Science & Technology, Kochi 682022, India. [3] Department of Ship Technology, Cochin University of Science & Technology, Kochi 682022, India.

Received: 25 October 2019 Accepted: 17 February 2020
Published online: 28 February 2020

References

1. Lau RY, Liao SY, Kwok RC, Xu K, Xia Y, Li Y. Text mining and probabilistic modeling for online review spam detection. ACM Trans Manag Inform Syst. 2011;2(4):25.
2. Lin X, Li Y, Wang X. Social commerce research: definition, research themes and the trends. Int J Inform Manag. 2017;37:190–201.
3. Matos CAD, Rossi CAV. Word-of-mouth communications in marketing: a meta-analytic review of the antecedents and moderators. J Acad Market Sci. 2008;36(4):578–96.
4. Jeon S, et al. Redundant data removal technique for efficient big data search processing. Int J Softw Eng Appl. 2013;7.4:427–36.
5. Dave K, Lawrence S, and Pennock D. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. WWW'2003.
6. Zhou Y, Wilkinson D, Schreiber R, Pan R. Large-scale parallel collaborative filtering for the netflix prize. 2008. p. 337–48. https://doi.org/10.1007/978-3-540-68880-8_32.
7. Zhang KZK, Benyoucef M. Consumer behavior in social commerce: a literature review. Dec Support Syst. 2016;86:95–108.
8. Cui Geng, Lui Hon-Kwong, Guo Xiaoning. The effect of online consumer reviews on new product sales. Int J Electron Comm. 2012;17(1):39–58.
9. Manek AS, Shenoy PD, Mohan MC, et al. Detection of fraudulent and malicious websites by analysing user reviews for online shopping websites. Int J Knowl Web Intell. 2016;5(3):171–89. https://doi.org/10.1007/s11280-015-0381-x.
10. Singh S, and Singh N. Big data analytics. In: Proceedings of the 2012 international conference on communication, information & computing technology (ICCICT), institute of electrical and electronics engineers (IEEE). 2012. p. 1–4. http://dx.doi.org/10.1109/iccict.2012.6398180.
11. Demchenko Yuri et al. Addressing big data challenges for scientific data infrastructure. In: IEEE 4th Int. conference cloud computing technology and science (CloudCom). 2012.
12. Sihong Xie, Guan Wang, Shuyang Lin and Yu Philip S. Review spam detection via time-series pattern discovery. In: ACM Proceedings of the 21st international conference companion on World Wide Web. 2012. p. 635–6.
13. Koren Y, Bell R, Volinsky C. matrix factorization technique for recommender systems. Computer. 2009;8:30–7.
14. Salakhutdinov R, Mnih A, & Hinton G. Restricted boltzmann machines for collaborative filtering. In: Proc. of the 24th Int. conference on machine learning. 2007. p. 791–8.
15. Hao MA, King I, Lyu MR. Learning to recommend with explicit and implicit social relations. ACM Trans Intell Syst Technol. 2011;2(3):29.
16. Bandakkanavar V, Ramesh M, Geeta V. A survey on detection of reviews using sentiment classification of methods. IJRITCC. 2014;2(2):310–4.
17. Gu V, and Li H. Memory or time—performance evaluation for iterative operation on hadoop and spark. In: Proc. of the 2013 IEEE 10th Int. Con. on high-performance computing and communications. 2013. https://doi.org/10.1109/hpcc.and.euc.2013.106.
18. Zhang Hanpeng, Wang Zhaohua, Chen Shengjun, Guo Chengqi. Product recommendation in online social networking communities—an empirical study of antecedents and a mediator. J Inform Manag. 2019;56(2):185–95.
19. Ghose A, Ipeirotis PG. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In: Int Conference Electron Comm ACM. 2007. p. 303–10.
20. Chong AY, Ch'ng E, Liu MJ, Li B. Predicting consumer product demands via Big Data: the roles of online promotional marketing and online reviews. Int J Prod Res. 2015;55:1–15. https://doi.org/10.1080/00207543.2015.1066519.
21. Yang H, Fujimaki R, Kusumura Y, & Liu J. Online Feature Selection. In: Proceedings of the 22nd ACM SIGKDD Int. Conference on KDD '16, 2016. https://doi.org/10.1145/2939672.2939881.
22. Breese JS, Heckerman D, and Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: Proc. of the 14th Conf. on Uncertainty in Artifical Intelligence, 1998.
23. Mukherjee A, Kumar A, Liu B, Wang J, Hsu M, Castellanos M, Ghosh R. Spotting opinion spammers using behavioral footprints. In: Proc. of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining Chicago, ACM. 2013. p. 632–40.
24. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: concerns and ways forward. PLoS ONE. 2018;13(3):e0194889. https://doi.org/10.1371/journal.pone.0194889.
25. Imon A, Roy C, Manos C, Bhattacharjee S. Prediction of rainfall using logistic regression. Pak J Stat Oper Res. 2012. https://doi.org/10.18187/pjsor.v8i3.535.
26. Chen T, Zhang W, Lu Q, Chen K, Zheng Z, Yu Y. SVD Feature: a toolkit for feature-based collaborative filtering. J Mach Learn Res. 2012;13(1):3619–22.
27. Shi Y, Larson M, Hanjalic A. Collaborative filtering beyond the user-item matrix—a survey of the state of art and future challenges. ACM Comput Surv. 2014;47(1):3.
28. Shan H, & Banerjee A. Generalized probabilistic matrix factorizations for collaborative filtering, In Data mining (ICDM), IEEE 10th international conference. 2010. p. 1025–30.
29. Salakhutdinov R, & Mnih A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In: Proc. of the 25th int. conference on machine learning. 2008. p. 880–7.
30. Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Al Najada H. Survey of review spam detection using machine learning techniques. J Big Data. 2015;2(1):23.
31. Wietsma TA, Ricci F. Product reviews in mobile decision aid systems. Francesco: PERMID; 2005. p. 15–8.
32. Jianguo C, et al. A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. Inform Sci. 2018;435:124–49.
33. Manek AS, Shenoy PD, Mohan MC, Venugopal KR. Aspect term extraction for sentiment analysis in large movie reviews using Gini-index feature selection method and SVM classifier. World Wide Web. 2017;20:135–54. https://doi.org/10.1007/s11280-015-0381-x.
34. Fan RE, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: A library for large linear classification. J Mach Learn Res. 2008;9:1871–4.

35. Ribeiro MT, Singh S, and Guestrin C. Why should I trust you?: Explaining the predictions of any classifier. In: Proc. ACMSIGKDD Int. Conf. Knowl. Discov. Data Mining. 2016. p. 1135–44.
36. Luo X, et al. An effective scheme for QoS estimation via alternating direction method-based matrix factorization. IEEE Trans Serv Comput. 2019;12(4):503–18.
37. Liu CL, Hsaio WH, Lee CH, Lu GC and Jou E. Movie rating and review summarization in mobile environment. In: IEEE trans. systems, man and cybernetics, Part C: applications and reviews. 2012. p. 397–407.
38. Vapnik, VN. The nature of statistical learning theory, Springer, 2nd ed, 1999. Translated by Xu Jianghua, Zhang Xuegong. Beijing: China Machine Press; 2000.
39. [Dataset] Flipkart-products. http://www.kaggle.com/PromptCloudHQ/flipkart-products.
40. [Dataset] https://snap.stanford.edu/data/web-Amazon.html.
41. [Dataset] He R, McAuley J. Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW; 2016.
42. Popescu AM, Etzioni O. Extracting product features and opinions from reviews. 2005; EMNLP.
43. Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, Franklin M, Shenker S, Stoica I. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing Technical Report UCB/EECS-2011-82. UC Berkeley: EECS Department; 2011.
44. Davis J, Goadrich M. The relationship between precision-recall and ROC curves, In ICML. 2006. p. 233–40.
45. Lee JS, Lee ES. Exploring the usefulness of predicting people's locations. Procedia Soc Beh Sci. 2014. https://doi.org/10.1016/j.sbspro.2014.04.451.

Publisher's Note

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