Data Mining and Machine Learning Approach for Online Product Recommendation System Using Sentiment Analysis

Hettikankanama HKSK[1], Vasanthapriyan S[2], and Rathnayake RMKT[3]

[1, 2] Department of Computing and Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka
[3]Department of Physical Sciences and Technology, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka
[#] Hettikankanama HKSK; <sanduni@silverstripers.com>

Abstract— With the evolution of Internet and e-commerce physical stores and businesses were moved to web breaking geographical barriers. Most online businesses use recommender systems to find right product for right customer at right time to increase customer satisfaction. Product recommendation systems are filtering tools which use data mining and machine learning algorithms to suggest the most relevant items to a particular user. This study illustrate how recommender systems increase the quality of the decisions that customers make while selecting a product by reducing the information overload and complexity. The goal of this study is to propose a novel product recommendation algorithm considering user reviews which provide multiuser recommendation. In this research a data set was taken through some different supervised and unsupervised learning methods, available recommendation systems and finally through proposed recommendation system. New recommendation model and its workflow is illustrated here which analyse review text and provide rating value for reviews through sentiment analysis and polarity estimation. This paper presents the methodology and techniques used in novel recommendation algorithm and its evaluation.

*Keywords— **Machine Learning, Recommendation system, Review rating, Sentiment analysis.***

I. INTRODUCTION

With the evolution of Internet and people's enthusiasm to perform tasks with few clicks, a thing that got affected is how vendors and customers perform their buying and selling tasks. The physical stores and businesses were moved to web by extending their services and capabilities worldwide by allocating a part of cyber space rather than being limited to one single physical location creating e-market and e-commerce. These systems can promote the business and organizations online and make it able for customers to perform their tasks with minimal human intervention [1]. This allowed vendors to expand the business beyond geographical limitation and acquire huge customer base worldwide more efficiently.

As many of the physical stores began to move to the web, only being a part of the web didn't fulfilled the expectations of businesses. There emerged a huge competition between online businesses. Therefore they strive to achieve competitive advantage to get more customer attraction and increase their sales. By increasing customer satisfaction the businesses can gain more customer attraction. Which shows that e-retailers should provide e-satisfaction to create a good customer base [2].

Number one goal of many leading businesses worldwide is to make sure their customers are satisfied [3]. To make customers satisfy it is vital to provide the service they expect by providing value for their money and facilitate them. It should be made easier for customers to find what they want with good quality and in less time and the process should be less exhaustive. For that most online businesses use recommender systems to find right product for right customer at right time [4]. . These systems have the ability to increase the quality of the decisions that customers make while searching and selecting a product by reducing the information overload and complexity caused by them [5].

Product recommendation systems can be identified as filtering tools which use data, pattern recognition and filtering to suggest the most relevant items to a particular user [6]. Those systems work on a specific parameters like ratings or product properties. These systems are having the challenge to give high quality and relevant recommendations without errors in reduced time. This will increase the traffic to the website [7]. Recommendation systems use various data mining and machine learning algorithms [7]. As there is a huge competition between online businesses the quality and accuracy of product recommendation system chooses the winner in online sales nowadays.

It is found that people read reviews given by other customers before purchasing. The survey done by Myles

Anderson shows that 88% consumers trust online reviews like personal recommendations [8].

In this paper it talks proposes a new model for product recommendation by rating the products according to the reviews given by users. According to this method the textual reviews given by previous customers of a particular product are considered and the products are ranked and sorted according to the cumulative score given to the product and the recommendation is done as customer search by keywords. With this method it will suggest good quality products as reviews comes from many users and they provide a practical opinion with a wide view which provide multiuser recommendation the upcoming trend of product recommendation. Architecture of the proposed model and method took to conduct the research and the results are discussed in this paper.

II. METHODOLOGY

This research was conducted according to a specific method as shown in Figure 1. In this method first of all an questionnaire based online survey to collect requirements is done and the results given by the questionnaire was analyzed to understand drawbacks of current system, how much people trust the base of current system and how much will people trust the base of proposed system and suggestions to improve ranking for recommendation are take. After that the relevant data set which has enough reviews on product was taken and then as the next step the data pre-processing is done to take the data into an applicable condition. Then sentiment values for the review text is calculated and sentiment polarity is estimated. Then some insights of data set and behaviors of some variables in relation to others are measured. The supervised learning methods are applied on data set and learning algorithm with highest accuracy is identified. After this a product ranking method which consider sentiment value of user reviews. Next step is model validation for high accuracy as it is a vital characteristic for a recommender system. To conduct the research as the programming language Python programming language is used and Jupiter Microsoft Azure Notebook is used as development environment.

Here the customer product reviews given by users are extracted in native format. In the extracted data set there are more than 400,000 review data from Amazon on some different brands of mobile phones. It contains reviews from Amazon.com about unlocked mobile phones. Data set carry basic product information, price, ratings and review vote ratings. Here the data set was extracted using web scrapper PromptCloud.

Python programming language is used in analysing and implementing the model. A data set which contain 400 thousand user reviews on 4410 products from 385 brands are taken.

III. EXPERIMENTAL DESIGN

A. Requirement Analysis

Here the requirement analysis was done in two ways as from literature review where the capabilities, pros and cons of current systems are evaluated and the next way is through a questionnaire where it demonstrates the acceptability and needfulness of this kind of a System. Here a questionnaire was done in the form of online questionnaire which was answered by 164 IT/CS and IS students of foreign and local universities. Questions included in questionnaire and their context are as shown in table 1 in appendixes. It shows surprisingly many people do not trust rating based product ranking systems solely and they tend to read reviews. Thus the results show a huge needfulness of considering reviews to rank products.

B. Proposing Best Accurate Algorithm to Predict Polarity Value for Textual Reviews

Here a data mining approach to predict values for the products considering positivity and negativity of its reviews is proposed. Five algorithms were used on data prediction and their performance measurement results were recorded and compared to estimate and propose the best performing algorithm in sentiment polarity prediction. Here Decision tree classification algorithm, random forest classification algorithm, K-Neighbours classification algorithm, Ada-boost algorithm and finally the ensemble approach was used to predict review polarity.

C. System Workflow Design

Here in this research the final outcome is a system which can recommend products considering the positivity and negativity of review that a particular product has taken. Workflow of the proposed System is shown in Figure 2. Here this system will be able to predict polarity of a review as negative or positive and then calculate average value for each product considering review negativity and positivity.
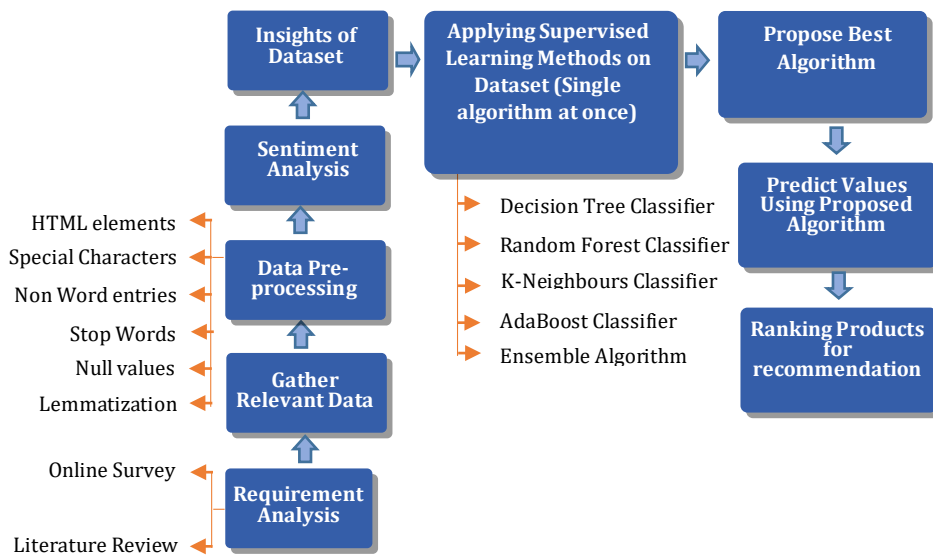
Figure 1 (Methodology flowchart):

Insights of Dataset

Applying Supervised Learning Methods on Dataset (Single algorithm at once)

Propose Best Algorithm

Sentiment Analysis

Predict Values Using Proposed Algorithm

HTML elements
Special Characters
Non Word entries
Stop Words
Null values
Lemmatization

Data Pre-processing

Ranking Products for recommendation

- Decision Tree Classifier
- Random Forest Classifier
- K-Neighbours Classifier
- AdaBoost Classifier
- Ensemble Algorithm

Gather Relevant Data

Online Survey

Literature Review

Requirement Analysis

Figure 1. Methodology

Figure 2 (Workflow flowchart):

User Query

Keyword extracting

Customer product review data

Recommend products according

Data pre processing

Null value treatment
HTML elements remover
Special character remover
Non word remover
Stop word remover
Lemmatize

Ranking products accordingly for recommendation

Average estimation score

Keyword extraction

Novel review-based score estimation mechanism

Sentiment Intensity analyzer

Product sentiment score calculation

Polarity estimation

Negative   Positive
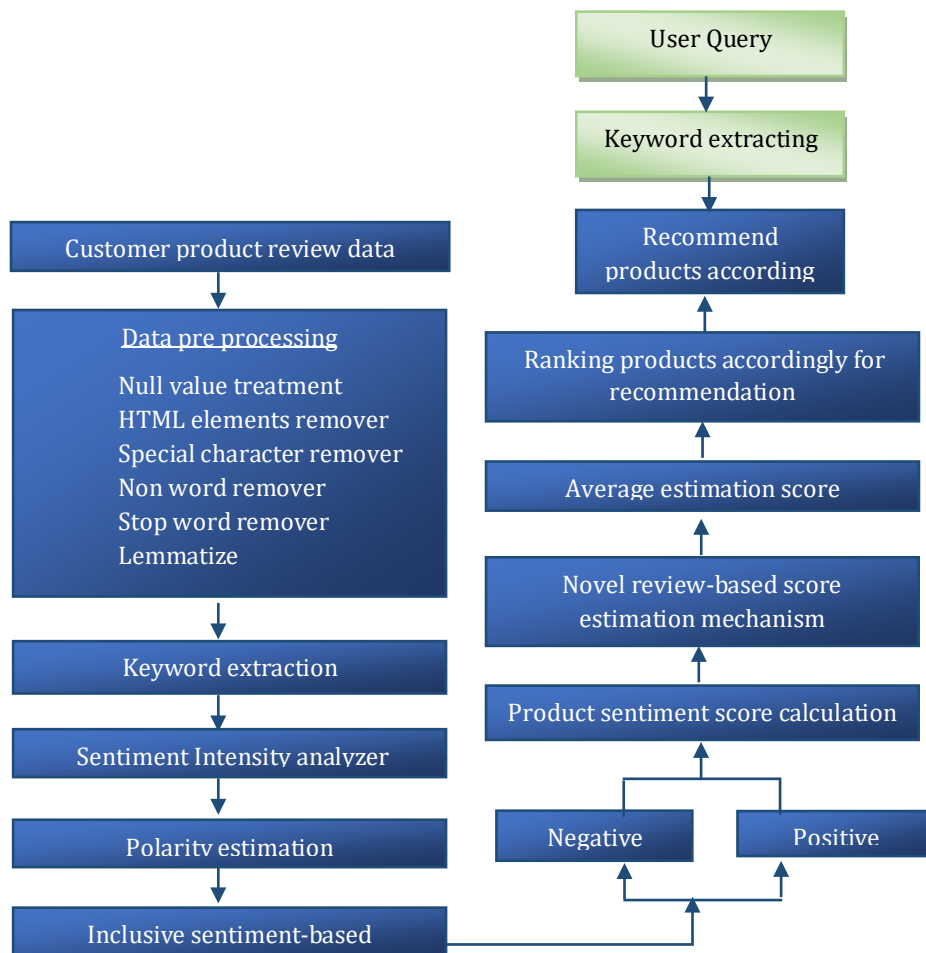
Inclusive sentiment-based

Figure 2. Workflow of the proposed Recommendation system

IV. Results

In this study as described earlier random forest classifier, decision tree classifier, K neighbours classifier and AdaBoost classifier are used. Their results were recorded as shown in Table 2. According to results K-Neighbours classifier is proposed as most appropriate algorithm as it has highest precision, recall, cross validation precision, F1 measure and also least Mean error. Results comparison is graphically illustrated in Figure 3.
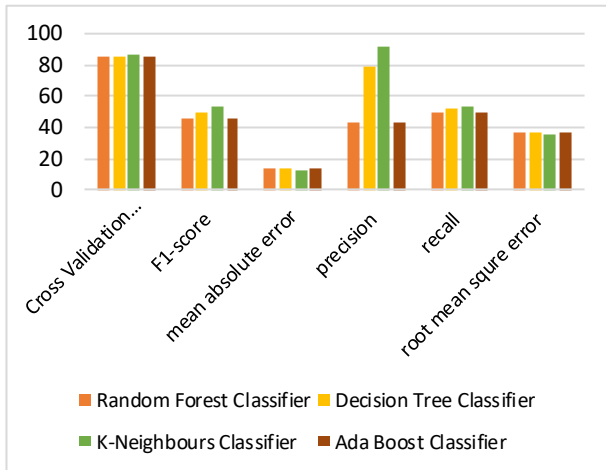

Figure 3. Algorithm results comparison

Ensemble approach which cause to increase the accuracy is also used to predict the sentiment values. The comparison of results along with results of K-Neighbours as shown in Figure 4 ensures that K-Neighbours is still having highest accuracy and that algorithm was proposed as best algorithm to predict sentiment polarity of reviews.
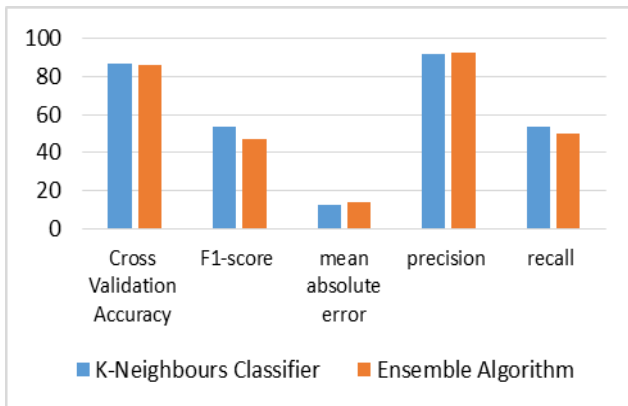

Figure 4. Comparison of ensemble and K-neighbours algorithms

For the evaluation of the model some evaluating mechanisms which consider different types of measurements as accuracy or coverage are used. Accuracy measures number of correct recommendation divided by all possible recommendations. Coverage is objects considered divided by objects in the search space. Accuracy measurements are categorized in to two parts as statistical and decision support accuracy metrics. Statistical Accuracy Metrics do the evaluation by comparing predicted recommendations with actual user recommendations. Correlation, MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) are statistical accuracy measuring metrics here.

$$MAE = \frac{1}{N} \sum_{u,i} |P_{u,i} - R_{u,i}| \qquad (1)$$

Here $P_{u,i}$ means predicted rating for user u on item i. $R_{u,i}$ is actual rating and N is the total number of rating on item set. Lower the MAE is better and same for the RMSE.

$$RMSE = \sqrt{\frac{1}{n} \sum_{u,i} (p_{u,i} - r_{u,i})^2} \qquad (2)$$

Precision recall curve (PRC), Reversal rate, Receiver Operating Characteristics (ROC), weighted errors are some mostly used Decision support accuracy metrics. Precision, recall and F-measure are calculated as follows.

$$Precision = \frac{Correctly\ recommended\ items}{Total\ recommended\ problems} \qquad (3)$$

$$Recall = \frac{Correctly\ recommended\ items}{Total\ useful\ recommended\ problems} \qquad (4)$$

F-measure used to take precision and recall in to single metric.

$$F\_measure = \frac{2PR}{P+R} \qquad (5)$$

Coverage is measured considering number of users that recommender system help when consider the number of users of the product.

According to results K-Neighbours algorithm is used to predict sentiment value and products were ranked according to those scores. Product ranking of the final system results are as Shown in Figure 5.

		1.000000
Doogee X5 Pro 16GB Black, Dual Sim, 2GB Ram, 5.0 inch, Unlocked International Model, No Warranty		1.000000
Smartwatch + Unlocked Watch Cell Phone All in 1 Bluetooth Watch for iPhone Android Samsung Galaxy Note,Nexus,htc,Sony White		0.971429
Smartwatch + Unlocked Watch Cell Phone All in 1 Bluetooth Watch for iPhone Android Samsung Galaxy Note,Nexus,htc,Sony Silver		0.961538
CNPGD All-in-1 Watch Cell Phone & Smart Watch Sync to Android IOS Smart Phone (Gold)		0.950000
CNPGD [U.S. Warranty] All-in-1 Smartwatch and Watch Cell Phone Gold		0.947368
BLU Advance 4.0L Unlocked Smartphone -US GSM - White		0.937500

Figure 5. Product ranking according to review polarity score

Table 2. Performance measurement results of algorithms

Algorithm name	Cross Validation Accuracy	elapsed time	F1-score	mean absolute error	precision	recall	root mean square error
Random Forest Classifier	86	15.86	46	13.95	43	50	37.35
Decision Tree Classifier	86	0.2	50	13.63	79	52	36.91
K-Neighbours Classifier	87	75.33	54	12.83	92	54	35.81
Ada Boost Classifier	86	10.11	46	13.95	43	50	37.35

V. Discussion and Conclusion

Here in this research it demonstrated following objectives at the initializing phase and here at the conclusion emphasize that all three objectives of the research are fulfilled.

- Obj1: Identify different types of methods used for product recommendation (through literature review)
- Obj2: Propose more effective algorithm for estimating sentiment polarity of a review text. (K neighbours – Proposed as the algorithm with best accuracy)
- Obj3: Facilitating customers by providing recommendations based on textual reviews given by product users. (Ranking is done and products are presented to user considering highest mean sentiment score)

As some limitations of this research followings can be highlighted.

- Achieving 100% accuracy by automating text analyzing is not possible as human also feel conflicts when trying to understand natural language statements.
- Limited to English language
- Deep learning was barred due to limited resources

As for future research direction here the followings are recommended.

- Train the model to identify Sinhala and Singlish languages to be used in local online businesses (any other language).
- Combining review sentiment based recommendation system with an already using recommendation algorithms.

- Consider some more variables which affect product credibility along with review sentiment intensity to create a better recommendation algorithm.

The originality of this paper or the contribution of this study towards the enhancement of industry and knowledge are as follows.

Most research are based only on analyzing co relation and sentiment using lexicon data bases. Here algorithm training was done to predict sentiment polarity and rank products and compared the lexicon data base. Most sentiment based researches are proposal researches while this is an implementation research. Proposed novel method can be used to create or extend lexical database. Highest accuracy was achieved up to 92% here, other researches which were based only on using lexical data bases like Vader limited to highest 80% accuracy.

And considering system's perspective this system overcome some drawbacks of currently using star rating based system. The drawbacks that were overcame are;

-Fake rating generation (As it is hard to generate textual reviews than generating fake star ratings as it can easily detected)

-Systems that show positive reviews at first (When systems were designed to show only positive reviews first the user will mislead as they make choice based on those reviews only)

-Inability to present real experience in number (It is hard for anyone to present their real experience with the product through star ratings which are in a range of 1 to 5)

REFERENCES

[1] Mahmood, S M Feroj. (2016). E-commerce, online *shopping and customer satisfaction: an empirical study on e-commerce system in dhaka*.

[2] Szymanski, D. and Hise, R. (2000*). E-satisfaction: an initial examination. Journal of Retailing,* 76(3), pp.309-322.

[3] Fieldboom.com. (n.d.). Customer Satisfaction Survey. [online] Available at: https://www.fieldboom.com/customer-satisfaction-surveys.

[4] Wang, J. and Zhang, Y. (2013). Opportunity Model For E-Commerce Recommendation. *Proceedings Of The 36th International ACM SIGIR Conference On Research And Development In Information Retrieval* - SIGIR '13, Pp.303-312.

[5] Spiekermann, S. *Online Information Search with ElectronicAgents: Drivers, Impediments, and Privacy Issues, unpublished doctoral dissertation,* Humboldt University Berlin, Berlin, 2001

[6] MacKenzie, I., Meyer, C. and Noble, S. (2013). *How retailers can keep up with consumers. [online] McKinsey & Company.* Available at:
 https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers

[7] N. Vaidya and A. R. Khachane, "*Recommender systems-the need of the ecommerce ERA,*" 2017 International Conference on Computing Methodologies and Communication (ICCMC), Erode, 2017, pp. 100-104. doi: 10.1109/ICCMC.2017.8282616.

[8] Anderson, M. (2014). *88% Of Consumers Trust Online Reviews As Much As Personal Recommendations.* [online] Search Engine Land. Available at:
https://searchengineland.com/88-consumers-trust-online-reviews-much-personal-recommendations-195803

[9] Wang, H. (2019). Market Basket Analysis in R | Henry Wang. [online] Henrywang.nl. Available at: https://henrywang.nl/market-basket-analysis-in-r/.

[10] Chapman, C., Fiet, M. and Christoper, N. (2015). R for Marketing Research and Analytics. Cham: Springer International Publishing.

APPENDIXES

Table 1. Question and their context in questionnaire

Objective	Question
Learn the awareness of online product recommendation systems among online shopping sites.	1. Have you heard about Product recommendation systems? 2. What do you think a product recommendation system should do?
Identify whether the person is a user of online shopping.	3. Have you done online shopping?
Identify whether the person have been subjected to receiving poor quality products than expected in online shopping and they care about that.	4. Have you ever been victim of receiving poor products than expected when purchased online? 5. How it feels for you?
Get idea of the person about current mechanism used to recommend products.	6. How much do you trust ratings given for products? (Star Ratings)
Get idea of how much people consider user given textual reviews before making an online purchase. Identify the needfulness of considering user reviews in product recommendation.	7. Do you prefer asking from real users of product before purchasing online, even after seeing the product has acceptable ratings? 8. Do you read reviews given by real users of the product before purchasing online? 9. How many reviews normally you read before make a purchase? 10. How much time would you prefer to allocate normally for reading reviews? 11. What do you think? Analysis of reviews by reading by yourself is ,
Get the user's acceptance and agreement on a review based ranking system for recommendation	12. Do you think it will helpful if a product recommendation system analyse real user reviews and recommend products for you? 13. How much will you trust a system which analyse textual real user reviews and recommend products?
Get suggestions from people to improve the system	14. Do you have any suggestion to improve current product recommendation systems?