
DATA MINING APPROACH FOR CUSTOMER REVIEW BASED PRODUCT RANKING IN ONLINE PRODUCT RECOMMENDATION

Hettikankanama H.K.S.K.^{1*}, Vasanthapriyan S.¹, Rathnayake R.M.K.T.²

¹Department of Computing and Information Systems,

²Department of Physical Sciences and Technology,

Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya,
Sri Lanka

*keshalasaniduni@gmail.com**

ABSTRACT

Product recommendation systems use ratings from users to rank products. There's a gap between analyzing user ratings and reviews in purchase decision making process. 97% people read reviews to confirm product quality and trust reviews than ratings. This research proposes a data mining and machine learning model to rank products based on textual reviews. When considering methodology and design of this study, a survey was conducted and outcomes show a needfulness of using reviews in ranking. A model to rank products considering review sentiment polarity is proposed and implemented using Python programming language. Well-structured unique workflow of data pre-processing, sentiment-polarity estimation, algorithm training for high accuracy, best algorithm selection, value prediction and calculation for ranking products for recommendation is used. As the results, Survey indicates that 98.8% of people read reviews though the star rating is presented. 85.8% say they trust this kind of system more. Among four algorithms, K-Neighbors algorithm was proposed as best performing algorithm for value prediction for this type of research. Products were successfully ranked based on sentiment score. Most of existing researches are proposal researches while this is an implementation research. Proposed algorithm and model with high accuracy can use as base for future researches. Illustrated Python implementation method also can be used for future work. There are some practical implications as Fake review generation can mislead the outcomes and reviews from other languages, rather than English, will not be considered for calculations and have to train and create lexical database for other languages.

Keywords: *Data Mining, Product Recommendation, Sentiment Analysis, User Reviews*

1 INTRODUCTION

Internet has evolved as a powerful tool through many years and nowadays each and every thing around human is getting digitalized and moving to web. As Internet and digital devices bring things which are nearby and things which are much far away to the finger tips people use internet facilities in there day to day life. In this era where people live busy lives, they try to make their day to day life events much easier and faster. People seek ways to perform life tasks efficiently. Internet, in this context has become inevitably beneficial thing.

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 (Szymanski & Hise, 2000).

Product recommendation systems use different filtering tools with algorithms and data to suggest most appropriate product for customer (Techlabs, 2017). A recommendation system has two major parts involves, as user and item. Recommender systems use some parameters for recommendation. Providing accurate recommendations in high quality, ability to serve many users in less time and ability to increase sales by user satisfaction are the challenge of this kind of system (Vaidya & Khachane, 2017). The technology opened the door for people to the world of information where it brought its own obstacle of finding needed information in less time. This is called information overload (AIMultiple, n.d.). Most of the time customers get similar looking information according to their inquiry but not relevant for actual needs. This makes it difficult to find products. Recommendation systems which filter all irrelevant information and retrieve appropriate information prevent this problem mainly. This improves the accessibility to information relieving the burden of product discovery.

These recommendation systems can be used in many types of businesses as e-commerce, online retail, Media, Banking, Telecom, Utilities shopping. When the way a recommender system works is taken to consideration, the base of them are information filtering and matching algorithms which can combine the user and the content considering two types of information as explicit data, which is created by user by performing a direct action to show a preference for example ranking a product, and the other one is implicit behavioural data by learning behaviour of user like what products does user view most. Basically the working process of a recommendation system can be listed under 6 stages: gathering relevant data, finding patterns in actions or behaviour, extracting valuable insights, calculations, comparisons, recommending most suitable

products. For the first three steps mentioned above the data mining techniques are used.

2 LITERATURE REVIEW

Several literatures have been studied to understand the definition of product ranking in online purchases and from which some of the concepts related to the product are explained.

In the article (Adomavicius & Tuzhilin, 2005) They provide a study on how to recommend a product to others based on its usability, content and method of advance recommendation. They also discussed expanding the recommendation system to provide consumers with more thorough knowledge, providing consumers with a more flexible and less protruding form of recommendation. In study the recommendation system based on social networks, (He & Chu, 2010) debated product recommendations via internet or social networks that would benefit internet customers. They addressed how information can be extracted from social media through semantic filtering that analyses the information and is provided to a collection of individuals by giving their views or recommendations and sharing it with each other in order to get a cooperative concept, so that each user can benefit from this advice for the product they choose with the assistance of other views on the product they choose.

In study sentence dependency parsing for opinion mining (Felix & Sasipraba, 2016), they provided comprehensive data on how to extract views from countless reviews in online shopping. The writers have implemented a phrase-dependence parsing that helps extract the product and user opinion specification and creates a relationship between them. Text stream clustering algorithm based on the selection of adaptive feature (Gong, Zeng, & Zhang, 2011), in which the authors discussed the analysis of text streams that highlights a particular topic in a survey. A new text stream cluster algorithm has been introduced by the authors to focus on the specific information that needs to be extracted. This data enables us to understand better how to extract all the reviews from the internet portal. Product Aspect Ranking Technique (Wu, Zhang, Huang, & Wu, 2009) the writers offer a comprehensive overview of how products are evaluated in this study based on the significant elements that benefit both the company and customers. They provide a description of the identification and classification of a product. Consumers may vary in the significant elements of a product. Based on the advantages and disadvantages of the significant element, despite the good value of other elements, the general evaluation of the product may influence.

3 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 are 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 considers 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.

4 DATA ANALYSIS & RESULTS

In this research reviews are analysed using NLTK and the reviews were given a value according to the sentiment intensity of the words in review text. These sentiment value score was calculated at first which is ranging from -1 to 1 where it is considered that the degree of minus intensity a value is more intended to be a negative sentiment. Here as described earlier in the chapter some classification algorithms were used. Here the best algorithm to perform the task of product review sentiment analysis has to be identified. In this research some classifier algorithms are used and their results were collected. By comparing results of accuracy of each algorithm one most appropriate algorithm was proposed. In this study random forest classifier, decision tree classifier, K neighbours classifier and AdaBoost classifier are used. Their results were recorded as presented in Table 1 (a and b).

Ensemble algorithm which causes to increase the accuracy is used to predict the sentiment values in this part of the implementation. Here it uses best combination of single algorithms to create a high accurate algorithm.

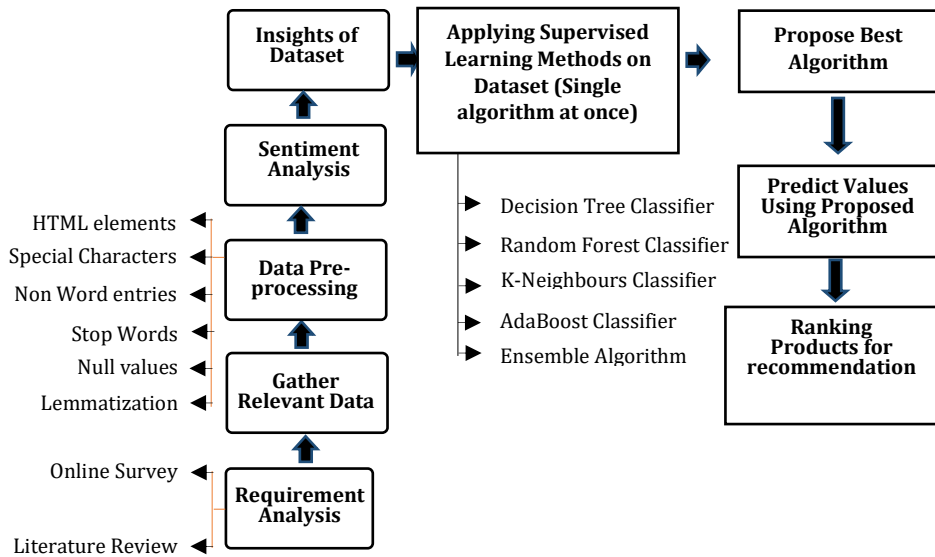


Figure 1: Methodology of Research

In this phase ensemble algorithm is used to predict sentiment value and products were ranked according to those scores. Proposed ranking system workflow is as in Figure 2 Ranked Results are as Shown in Table 2 .

Table 1(a): Results for performance measurements of algorithms

Classifier	Cross Validation Precision [Test] %	Cross validation Precision [Training] %	Elapsed Time (Sec.)	F1-Score [Test] %	F1-Score [Training] %	Mean Absolute Error [Test] %	Mean Absolute Error [Training] %
Random Forest Classifier	85.0	86.0	25.47	46.0	46.0	14.96	14.0
Decision Tree Classifier	84.0	86.0	0.4	52.0	50.0	14.05	13.6
K Neighbors Classifier	85.0	87.0	86.1	49.0	54.0	14.50	12.8
Ada Boost Classifier	86.0	86.0	14.51	46.0	46.0	14.95	14.0

Table 1 (b): Results for performance measurements of algorithms

Precision [Test] (%)	Precision [Training] (%)	Precision Difference (%)	Recall [Test] (%)	Recall [Train] (%)	Root Mean Square Error[Test] (%)	Root Mean Square Error [Training] (%)
43.0	43.0	0.0	50.0	50.0	38.67	37.35
93.0	79.0	-14.0	53.0	52.0	37.48	36.91
88.0	92.0	4.0	52.0	54.0	38.08	35.81
43.0	43.0	0.0	50.0	50.0	38.67	37.35

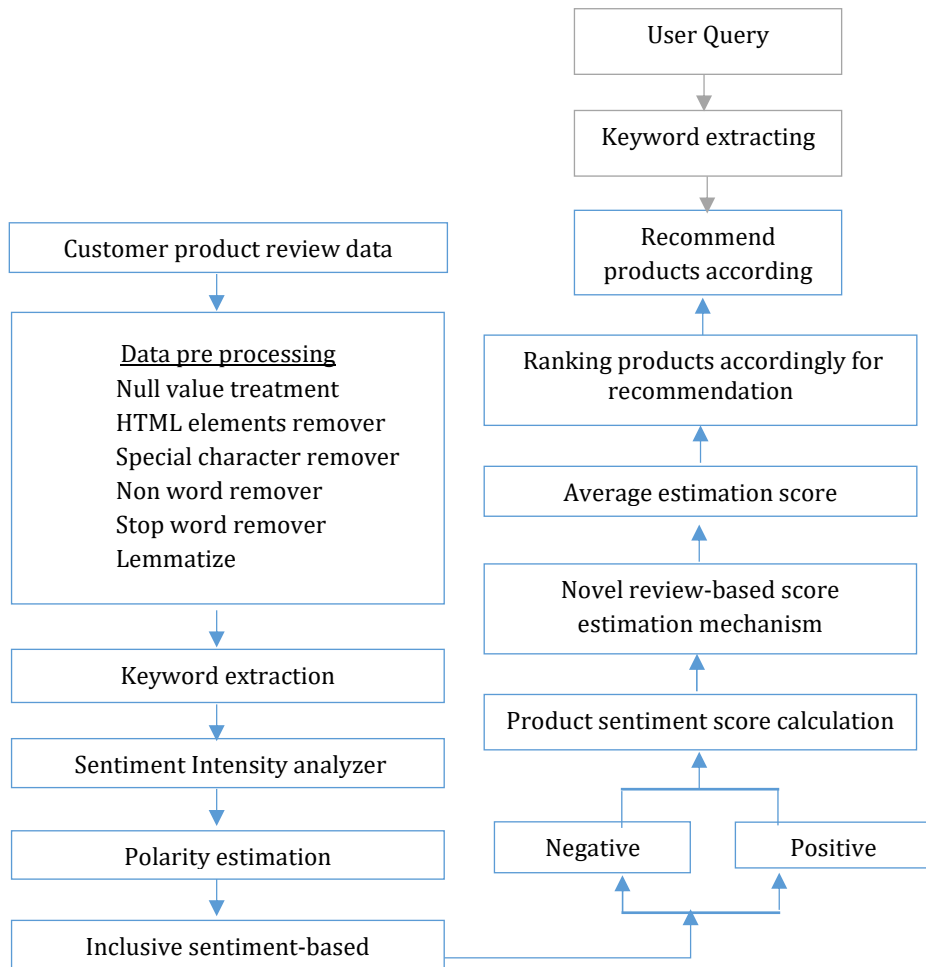


Figure 2: Workflow of the proposed system

Table 2: Sample of ranked products according to given score

Doogee X5 pro 16GB Black, Dual Sim, 2GB Ram, 5.0 inch, Unlocked International Model, No Earrenty	1.0000000
Smart watch +Unlocked Watch Cell Phone All in 1 Bluetooth Watch for iPhone Android Samsung Galaxy Note, Nexus, htc, Siny White	0.971429
Smart watch+ Unlocked Watch Cell Phone All in 1 Bluetooth Watch for iPhone Android Samsung Galaxy Note, Nexus, htc, Siny Silver	0.961538
CNPGD All-in-1 Watch Cell Phone & Smart Watch Sync ti Android IOS Smart Phone(Gold)	0.950000
Smart watch and Watch Cell Phone Gold	0.947368
L Unlocked Smartphone US GSM White	0.937500

5 DISCUSSION OF THE FINDINGS

Here the K-Neighbors algorithm is resulted as the best algorithm for classification of this dataset. This was compared with Ensemble algorithm results and still K-Neighbors shows best performance. Comparison results are graphically presented in following Figure5 and Figure 6. This algorithm can be taken to predict sentiment polarity and resulted system to rank products can be used to rank products in online retail sites to facilitate searching and purchase decision making of customers and this will turn shoppers into buyers as this decrees information overloading and consider real user given textual reviews for product ranking.

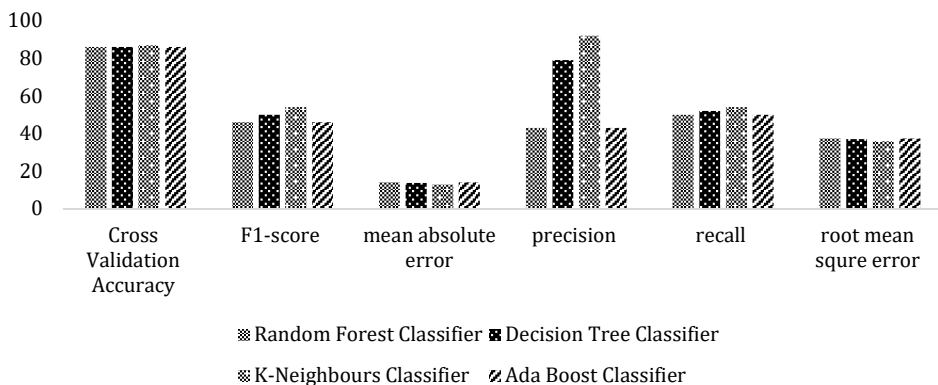


Figure 5: Graphical representation of performance measurement results of algorithms

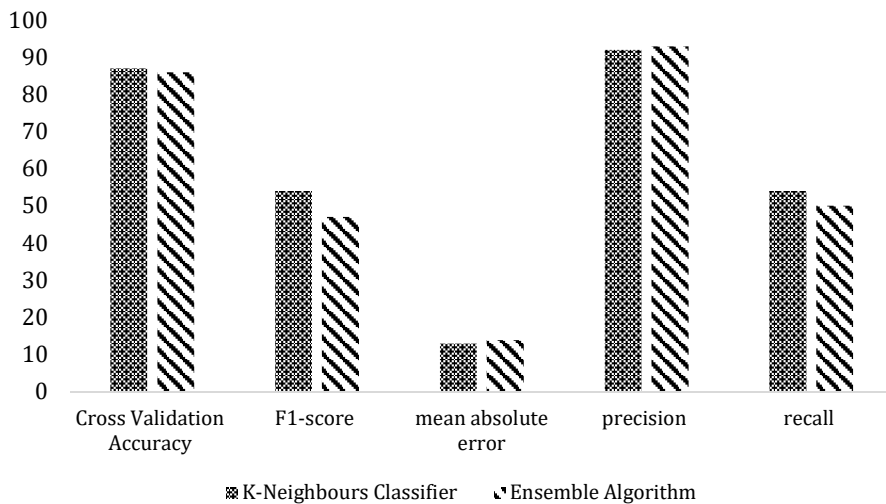


Figure 6: Performance measurement results of K-Neighbors and Ensemble algorithms

6 CONCLUSION & CONTRIBUTIONS

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.

- Objective 1: To identify different types of methods used for product recommendation (through literature review)
- Objective 2: To propose more effective algorithm for estimating sentiment polarity of a review text. (K neighbours – Proposed as the algorithm with best accuracy)
- Objective 3: To facilitate 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 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 researches are based only on analyzing co relation and sentiment using lexicon data bases. Here algorithm training was done to 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 overcomes some drawbacks of currently using star rating-based system. The drawbacks that were overcome 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

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734-749. <https://doi.org/10.1109/TKDE.2005.99>

-
- AIMultiple. (n.d.). Recommendation systems increase sales [2019 update]. Retrieved from Applied AI website: <https://blog.aimultiple.com/recommendation-system>
- Felix, A. Y., & Sasipraba, T. (2016). *INCIDENT MAPPING AND EAS USING DECISION SUPPORT SYSTEM*. 11(15), 4.
- Gong, L., Zeng, J., & Zhang, S. (2011). Text stream clustering algorithm based on adaptive feature selection. *Expert Systems with Applications*, 38(3), 1393–1399. <https://doi.org/10.1016/j.eswa.2010.07.041>
- He, J., & Chu, W. W. (2010). A Social Network-Based Recommender System (SNRS). In N. Memon, J. J. Xu, D. L. Hicks, & H. Chen (Eds.), *Data Mining for Social Network Data* (pp. 47–74). https://doi.org/10.1007/978-1-4419-6287-4_4
- Szymanski, D.M, & T. Hise, R. (2000). E-Satisfaction: An Initial Examination. *Journal of Retailing*, 76, 309–322. [https://doi.org/10.1016/S0022-4359\(00\)00035-X](https://doi.org/10.1016/S0022-4359(00)00035-X)
- Techlabs, M. (2017, September 28). What are Product Recommendation Engines? And the various versions of them? Retrieved from Towards Data Science website: <https://towardsdatascience.com/what-are-product-recommendation-engines-and-the-various-versions-of-them-9dcab4ee26d5>
- Vaidya, N., & Khachane, A. R. (2017). Recommender systems-the need of the ecommerce ERA. *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, 100–104. <https://doi.org/10.1109/ICCMC.2017.8282616>
- Wu, Y., Zhang, Q., Huang, X., & Wu, L. (2009). Phrase Dependency Parsing for Opinion Mining. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1533–1541. Retrieved from <https://www.aclweb.org/anthology/D09-1159>