

An Innovative Recommender System for E-Commerce Websites using Natural Language Processing



Sushmita Bose, Chaitali Choudhary, Ashok Behra, Sumit Kumar Sar

Abstract: A Recommender System has become the go-to application for the internet generation these days. Mono-variate, bi-variate and multi-variate Recommender Systems are available to consumers of various products and services for the last 10 years or so only. In this paper, opinion mining dependent sentiment analysis using NLP tools will be used to recommend products to their purchasers on e-commerce websites. The application can be developed on the Python platform can be commercially used and will be precisely used to people who have to spend money without traditionally touching or feeling the item.

Keywords: Recommender System, mono-variate, bi-variate, multi-variate, opinion mining, sentiment analysis, NLP tools, Python.

I. INTRODUCTION

Before we start with Recommender Systems (RS), we must understand the super-class of such platforms: the Information Filtering systems. These are systems, software or hardware, that remove iterative, redundant, duplicate, unwanted and noisy data from the stream of data available for analysis [1]. Thus, here the data is being cleansed, with the objective of improving the signal to noise ratio in the information. A filter of such kind can be developed using techniques such as Bayesian Networks, support vector machines, neural networks, logistic regressions, etc. RS are a sub-class of the above-mentioned category [2]. RS 'Recommend', which simply means it puts one article before others in terms of 'preference', 'learning' from the history of the user's use. E-commerce as well as retail sale companies are giving a sharp edge to the power of data and boosting sales by implementing RS on their web sites [3].

The successful cases of such set-ups have been steadily progressing in the last years. It is now the right time to delve deeper into this utilitarian machine learning (ML) technique. In this paper, we have studied the major types of popular RS, their working, and their use by companies in the industry, requirements for implementing RS, and their evaluation [4].

II. RECOMMENDER SYSTEM

The objective of RS is to try to calculate and predict user interests plus suggest consumable services and items which may be interesting to them. These are amongst the most potent ML systems that online retailer sellers execute to encourage sales.

Information needed for a RS are sourced from specific ratings given by users for a movie, a song or a consumable product, from inferable search-engine queries and purchase histories, or other learnings regarding the users or the items themselves. websites such as Spotify, YouTube, Netflix, etc utilize that data to suggest playlists, named theme mixes, or make related video recommendations, apiece.

A. The need for Recommender Systems

Businesses using RS concentrate on boosting sales as a consequence of highly personalized offers and enhanced buyer experience.

Recommendations commonly accelerate searches and thus, users find it easier to access the content they're interested in. Sites then surprise them with suggestions they would have never actively sought. Businesses can gain and retain loyal customers by sending emails containing hyperlinks to new offers that match the recipients' interests, or recommendations of films and TV shows that befit their profiles.

The user begins feeling appreciated and is now likely to buy additional commodities or consume added content. By comprehending what a user desires, the company gains a sharp competitive edge. Now, the peril of losing a customer to a competitor drops.

This supplementary value given to users, by incorporating recommendations in e-commerce websites and products results in customer delight. Moreover, it empowers companies to stay ahead of their rivals and ultimately boost their profits.

With some understanding of RS, it has to be evaluated, when it's worthwhile to implement one[5]. An existing successful business could probably survive without a RS for a further long period.

Manuscript received on February 15, 2020.

Revised Manuscript received on February 20, 2020.

Manuscript published on March 30, 2020.

* Correspondence Author

Sushmita Bose, CSE, BIT, Durg, C.G., India, Email: czarinademarquis@gmail.com

Chaitali Choudhary, CSE, BIT, Durg, C.G., India, Email: chaitali.choudhary@gmail.com

Ashok Behra, CSE, BIT, Durg, C.G., India, Email: ashokctc15@gmail.com

Sumit Kumar Sar, CSE, BIT, Durg, C.G., India, Email: sumitsar@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Yet, the power of data to create a better user experience and to increase earnings is much leveraged [6].

B. Working of Recommender Systems

RS employ two classes of data and extract information in the following pattern :

1. Characteristic data/ information – Keywords used by users, categories of media searched by them, their specific profile details and preferences, are known as characteristic data/ information.

2. Interactions or transactions between users and their preferred items - This is information such as ratings, number of purchases, likes, etc.

Taking the above into consideration, the three algorithms that are being used in RS can be easily differentiated between:

1. The algorithms that use characteristic data/ information are known as Content – Based systems.

2. The algorithms that use interactions or transactions between users and their preferred items are known as Collaborative filtering systems.

3. The algorithms that combine both the above types of algorithms to avoid problems that are generated when working with just one of the above classes of information are called Hybrid systems.

When we will drill deeper into content-based and collaborative filtering systems and we can observe in what all ways they are different from each other.

C. Content-based systems

This kind of system makes recommendations which are based on the purchase and preference history of the user. Such algorithms reason that it is very probable that if a user has bought an article earlier, he/ she may purchase it again. Such articles are customarily clustered on the basis of their characteristics. Websites create user profiles by asking the users about their interests, preferences as well as their past purchase history. Certain systems, which are not considered totally content-based, employ user’s online data for creating user profiles, such as personal or social data [7].

One issue that exists is that the system will make repetitive recommendations due to extreme fitting. Any user may have purchased only a very few items, the system will not recommend items outside of that type of items, even if they may be appealing to the user.

The next frequent concern is, that newly registered users are short of a detailed and concise profile particularly if they have not been specifically asked for preferential data. Although, it is moderately easy to append new articles to the system, it must be ensured that the system assigns them a specific group as per their characteristics.

D. Collaborative filtering systems

Collaborative filtering is, now, one of the most commonly employed methods which normally renders better outcomes than the above mentioned content-based RS. Instances of this are found in the RS of popular multimedia websites such as Spotify, YouTube, Netflix, voot, Amazon prime, etc.

Collaborative filtering systems employ user transactions to filter for articles which may intrigue them. It may safely

be imagined as a generalization system of classification algorithms and regression algorithms. As opposed to other algorithms, where it is intended to predict a parameter that directly depends on other parameters (feature), in collaborative filtering, feature variables and class variables are not discriminated between. [8].

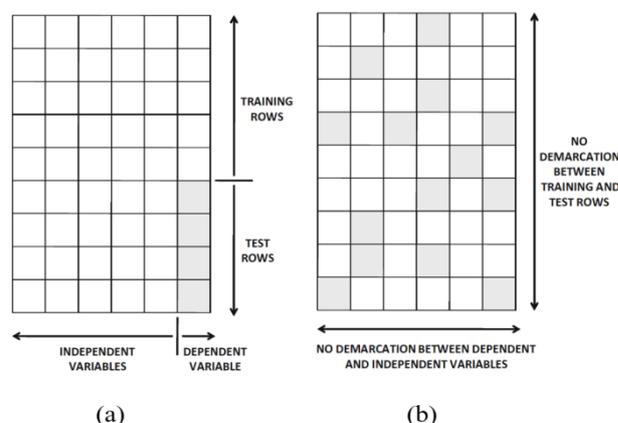


Fig. 1 (a) Classification (b) Collaborative Filtering

Precisely, the basis of collaborative filtering systems is that suppose a user likes article X, another user prefers the same article X plus another article, article Y, it can be safely assumed that the first user may as well be interested in the second article. Consequently, the systems intend to predict new transactions based on past ones [9]. Following are the two types of algorithms that have been designed for achieving the above: memory-based and model-based.

- **Memory-based algorithms :** Two memory-based approaches are being used: the first recognizes groups of users and employs transactions generated by a particular user for forecasting transactions from comparable users of the system. However, the second algorithm recognizes rated or viewed groups of articles by user X and uses the same to forecast the transaction of user X with a similar yet different article N.[10]

- **Model-based algorithms:** The basis of the Model-based algorithms are the in-vogue ML as well as data mining algorithms. Training of models to make them capable of making forecasts being the aim, we can take the example as following: the model would put to use the available user-article transactions for training the designed model to predict which are the five most likely articles which our user will prefer[11]. The capability of recommending a huge variety and number of articles to an even greater number of users, as opposed to memory-based algorithms and others.

E. Collaborative filtering systems and related concerns

There are the following two major problems with the above systems:

1. **Cold start:** The system developed must be provided with adequate data (user-article transactions) so that it can work satisfactorily. It is very difficult for a newly set-up e-commerce site to give recommendations. Only after several users have made significant number of transactions will the site be capable of doing so.



2. Adding new users/products or services to the RS: Whether the entire system has been newly set-up or it is an already functioning one, if a new product or user has to be added, the recommendations become somewhat less accurate.

In order to mitigate the above issues, the users have to be requested for various types of information when they sign-up, such as their location, age, preferences, gender, etc.. This meta-data about the products can then be used to relate them to the other existing products available in the existing database.

III. SENTIMENT ANALYSIS

Opinion mining and emotion AI are the two other names of Sentiment Analysis. It is defined as “the automated process of analysing text data, and classifying consumer opinions as positive, neutral or negative” [15]. This technology empowers users of the RS by helping them to identify and evaluate sentiments in online conversations, and also provides valuable insights to understand the feelings of the customers about their brands, products or services [16].

A. Sentiment based Recommender System

Sentiment analysis has become a key for businesses to learn how their clients feel about their presence, services, product line, brand, etc and how they express their feelings online.[17].

These days, people express their experience with a product or brand on social media. This has become a resource of information for the RS being implemented. Mentions of the item on twitter as #tags, on facebook as well as blogs or forums are accounted for in terms of not only number of mentions, but their quality as well. Positive or negative sentiments of users expressed online have a major influence on the result of the RS. [18] [19] [20] [21].

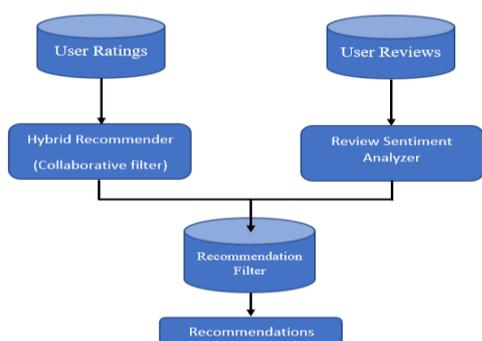


Fig. 2: Proposed RS.

B. Evaluating the RS

RS can be evaluated by various methods and how to select which method would be best is based on the objectives of the e-commerce website. If the sole interest resides in advising the top 5 articles the user shall look for, the predictions regarding the rest of the items need not be considered when conducting the evaluation [22].

Yet, the management of the e-commerce website may mostly be interested in particularly the above 5 recommendations. The preferred method of evaluating has a

significant effect on the way the system needs to be designed. The two types of RS evaluations methods, which are usually used are as follows: online methods and offline methods.

- **Online methods** : Also known as A/B testing), the online methods take the following steps:
 1. user reactions are measured,
 2. recommendations - direct as well as indirect are made
 3. recommendations are taken into consideration.
 This is the ideal method of evaluation but at the same time, hard to implement. This because interactions with the real time systems which are already in operation are mandatory.[23][24].
- **Offline methods** : Real time users are not directly involved in the offline methods, which are thus ideal for experimental stages. The available information is divided into separate training and validation sets of data. As per definition, this means that one part of the data will be employed to set the system up and the other part to evaluate it.

IV. PROPOSED METHODOLOGY

There are hardly any commercial RS which consider reviews given by users of the products as data set. Most use user-ratings, which may or may not be very appropriate. But, reviews are given deliberately by users after using the product, where they put in conscious effort. Thus it can be considered more authentic than ratings, although fake reviews have also been observed.

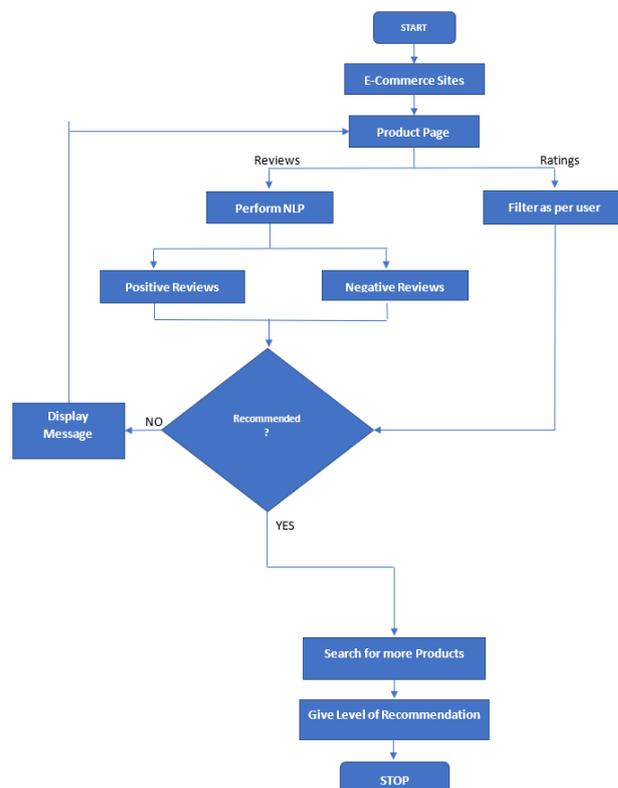


Fig. 3: Reviews based RS using NLP.

In our proposed methodology, we are using reviews given by users of the product to categorize it as either recommended or not-recommended, along with ratings.



For this purpose, Natural Language Processing tools will be used to analyse the data set obtained from the e-commerce website.

The description of the proposed methodology, as given in fig. 4, is as follows:

Step 1: The selected product is located on the e-commerce site. The landing page is scanned for the link to the page where reviews and ratings given by users of the product are available. The address of this page is used by web scraping tools to extract the reviews and ratings data and populate to a data base, such as Microsoft Excel Spreadsheet.

Step 2: Review data is fed to the NLP tools to segregate positive and negative describing words connected to the entities in the collected reviews. If the ratio of positive words is found to be more than 50% of the total count of describing words, will render the review as positive and vice versa. In a similar fashion, if the number of instances of 3 and above stars or marks is more than 80% of the total number of ratings, then, the rating of the product will be considered positive and vice versa.

Step 3: We then come to the step where we have to decide whether the identified product is to be considered recommendable or not. We propose to consider the product recommendable only and only if it is branded so in both reviews and ratings categories. In either case, the same product is searched for on 02 more similar e-commerce website and the above process is to be repeated. We propose to the most popular 03 e-com websites, Amazon.in, Flipkart.in and shopclues.in.

Step 4: Whether or not the product is found recommended on 2 out of 3 above mentioned websites, the result is displayed on our web page.

The proposed methodology is innovative, as it considers reviews instead of ratings.

V. EXPERIMENTS & RESULTS

We performed this experiment on python 3.0 using scrapy API used for web scraping and wordcloudnet for graphical representation of extracted reviews.

We took a sample space of Pilot-Metropolitan-Collection-Fountain reviews from Amazon sales data and extracted all kinds of reviews and then using NLP we segregated user comments into positive, negative and neutral words. Reviews were extracted as csv file using python code which was filtered only to extract ratings and comments. Fig 4 was plotted using those ratings.

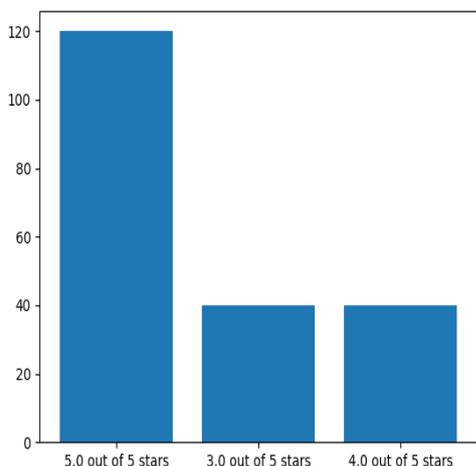


Fig 4: Graph showing various ratings given to product

An image of wordcloud was plotted in fig 5 using wordcloudnet by filtering out various specific keywords.



Fig 5 :wordcloud of pilot metropolitan fountain pen

The product Pilot-Metropolitan-Collection-Fountain was with more positive rating and reviews so the keywords highlighted in wordcloud was right, smooth, useful, well balanced and fineness as their frequency was high compared to others and we can conclude that negative words was much less. On the contrary another product we considered was Tatero-ultrasonic-repellent-non-toxic-cockroaches with more confusing ratings as 5-star rating count was nearly equal to 1-star rating but reviews showed that it falls more into negative side and thus results varied accordingly as shown in fig 6 & 7. The words with higher frequency for this product was waste, useless, will return etc.

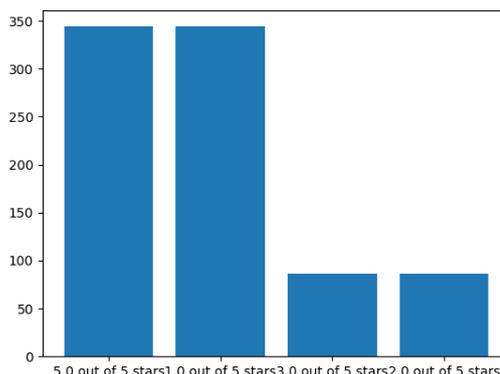


Fig 6: Graph rating for product TATERO-Ultrasonic-Repellent-Non-Toxic-Cockroaches



Fig 7: wordcloud for product TATERO-Ultrasonic-Repellent-Non-Toxic-Cockroaches

VI. CONCLUSION & FUTURE SCOPE

Much future scope is available in this area. Cross-platform user ratings can be used to further enhance the relevance of the recommendation. This system may extract reviews about a product from several websites and also develop a system to identify and remove fake reviews.

RS are being touted as the elixir for businesses in today's competitive markets. They give a feeling of delight to the users, creates and increases engagement and loyalty to the website. the business benefits from leveraged revenue and a very loyal set of buyers as well as sellers to them. As has been found, it is more beneficial to begin with implementing a simple RS for a limited group of test users, and then spend on more complex algorithms after the user, both buyer and seller base grows.

Business objectives will command the type of RS should be invested into. Will it provide more involvement to the existing users and will it encourage those occasional customers to become more involved and active, such questions have to be kept in mind.

REFERENCES

1. K. Shah, A. Salunke, S. Dongare, and K. Antala, "Recommender Systems: An overview of different approaches to recommendations," 2017.
2. A. V. Singh and M. Sharma, "Building an Effective Recommender Systems Using ML Based Framework."
3. X. Zhao, "A Study on E-commerce Recommender Systems Based on Big Data," 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), no. 1, pp. 222–226, 2019.
4. M. Lerato and O. A. Esan, "A Survey of Recommender Systems Feedback Techniques , Comparison and Evaluation Metrics," 2015.
5. A. Al-doulat, "Surprise and Curiosity in A Recommender Systems," 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), pp. 1–2, 2018.
6. G. H. Bossenbroek H, "Recommendation in E - commerce," 2015.
7. F. Sun, Y. Shi, and W. Wang, "Content-based Recommendation System Based on Vague Sets," pp. 3–6, 2013, doi: 10.1109/IHMSC.2013.218.
8. S. Y. Chen Y, Xia X, "A Collaborative Filtering Recommendation Algorithm Based on Contents' Genome," vol. 9, pp. 5–8.
9. F. S. Gohari and M. J. Tarokh, "Classification and Comparison of the Hybrid Collaborative Filtering Systems," vol. 6, no. 2, pp. 129–148, 2017.
10. X. Zhao and H. Wang, "A Collaborative Filtering Algorithm based on Memory."
11. S. Gong and H. Tan, "Combining Memory-Based and Model-Based Collaborative Filtering in RS," pp. 14–17, 2009, doi: 10.1109/PACCS.2009.66.
12. R. Katarya, "An Improved Recommender Systems Using Two-Level Matrix Factorization for Product Ontologies," 2017 International Conference on Intelligent Sustainable Systems (ICISS), no. Icss, pp. 223–226, 2017.
13. O. Barkan and N. Koenigstein, "ITEM2VEC: NEURAL ITEM EMBEDDING FOR COLLABORATIVE FILTERING," pp. 1–6, 2016.
14. A.- Amin, "Sentiment Analysis of Bengali Comments With Word2Vec and Sentiment Information of Words," pp. 186–190, 2017.
15. J. Desai and J. Majumdar, "Opinion Mining and Sentiment Analysis on Online Customer Review," 2016.
16. S. Vanaja, "Aspect-Level Sentiment Analysis on E-Commerce Data," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), no. Icirca, pp. 1275–1279, 2018.
17. A. Salinca, "Business reviews classification using sentiment analysis," 2016, doi: 10.1109/SYNASC.2015.46.
18. Z. Wang, C. S. Chong, L. Lan, Y. Yang, S. Ho, and J. C. Tong, "Fine-Grained Sentiment Analysis of Social Media with Emotion Sensing," no. December, pp. 1361–1364, 2016.

19. C. Cai and L. Li, "New Words Enlightened Sentiment Analysis in Social Media," pp. 202–204, 2016.
20. H. P. Patil, "Sentiment Analysis for Social Media : A Survey," 2015.
21. M. H. A. El-jawad and R. Hodhod, "Sentiment Analysis of Social Media Networks Using ML," 2018 14th International Computer Engineering Conference (ICENCO), pp. 174–176, 2017.
22. T. Hornung, M. Przyjacieli-zablocki, A. Sch, S. Franz, and G. Lausen, "Evaluating Hybrid Music Recommender Systems," pp. 57–64, 2013, doi: 10.1109/WI-IAT.2013.9.
23. B. R. Cami, H. Hassanpour, and H. Mashayekhi, "A Content-based Movie Recommender Systems based on Temporal User Preferences," pp. 121–125, 2017.
24. P. Venil, G. Vinodhini, and R. Suban, "Performance Evaluation of Ensemble based Collaborative Filtering Recommender Systems," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 1–5.
25. online <https://www.tryolabs.com>

AUTHORS PROFILE



Sciences.

Sushmita Bose received the B.Tech Degree in Information Technology from Pt. Ravi Shankar Shukla University Raipur, She is currently pursuing M.Tech in Computer Science and Engineering from CSVTU Bhilai, India. Her current research interest is Data



working as Assistant Professor in Bhilai Institute of Technology, Durg.

Chaitali Choudhary received her B.Tech. in CSE from GEC, Bilaspur and M.Tech. from Chhattisgarh Swami Vivekanand Technical University, Bhilai. She is currently pursuing Ph.D. from University of Petroleum and Energy Services, Dehradun. She is presently



Ashok Kumar Behra completed his M. Tech Degree in Computer Technology and Applications in 2011. He has over 20 years of experience. His research area is in the field of Software engineering. He is also the member of ISTE.



Sumit Kumar Sar received the B.Tech Degree in CSE from Pt. RSSU Raipur, and M. Tech. degree in C. T. & A. from CSVTU Bhilai, India. He is currently pursuing Ph.D. degree in CSE from CSVTU Bhilai, India. He is presently employed as Associate Professor with the Dept of CSE in Bhilai Institute of Technology, Durg.